conference

proceedings

SOUPS 2016: Twelfth Symposium on Usable Privacy and Security

Denver, CO, USA
June 22–24, 2016

Sponsored by

usenix
THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

ISBN 978-1-931971-31-7

Proceedings of the Twelfth Symposium on Usable Privacy and Security

Denver, CO, USA June 22–24, 2016

usenix ASSOCIATION

Thanks to Our SOUPS 2016 Sponsors

Gold Sponsor

Silver Sponsors

facebook Google

Bronze Sponsor

Media Sponsors and Industry Partners

CRC Press

Distributed Management Task Force (DMTF)

LXer

ISBN 978-1-931971-31-7

Thanks to Our USENIX Supporters

USENIX Patrons

Facebook Google Microsoft Research NetApp VMware

USENIX Benefactors

ADMIN *Linux Pro Magazine*

USENIX Partners

Booking.com Can Stock Photo

Open Access Publishing Partner

PeerJ

USENIX Association

Proceedings of SOUPS 2016: Twelfth Symposium on Usable Privacy and Security

June 22–24, 2016
Denver, CO, USA

SOUPS 2016 Symposium Organizers

General Chair/
Steering Committee Chair
Mary Ellen Zurko, *Cisco Systems*

Technical Papers Co-Chairs
Sunny Consolvo, *Google*
Matthew Smith, *University of Bonn*

Technical Papers Committee
Lujo Bauer, *Carnegie Mellon University*
Richard Beckwith, *Intel*
Konstantin Beznosov, *University of British Columbia*
Cristian Bravo-Lillo, *Universidad de Santiago de Chile*
Sonia Chiasson, *Carleton University*
Alexander De Luca, *Google*
Serge Egelman, *University of California, Berkeley, and International Computer Science Institute*
Sascha Fahl, *CISPA, Saarland University*
Alain Forget, *Google*
Simson Garfinkel, *National Institute of Standards and Technology (NIST)*
Marian Harbach, *International Computer Science Institute*
Cormac Herley, *Microsoft Research*
Iulia Ion, *Google*
Jaeyeon Jung, *Microsoft Research*
Mike Just, *Heriot-Watt University*
Apu Kapadia, *Indiana University Bloomington*
Janne Lindqvist, *Rutgers University*
Heather Lipford, *University of North Carolina at Charlotte*
Michelle Mazurek, *University of Maryland, College Park*
Heather Patterson, *Intel Labs and NYU Information Law Institute*
Emilee Rader, *Michigan State University*
Rob Reeder, *Google*
Jessica Staddon, *North Carolina State University*
Frank Stajano, *University of Cambridge*
Janice Tsai, *Microsoft*
Emanuel von Zezschwitz, *University of Munich (LMU)*
Rick Wash, *Michigan State University*
Tara Whalen, *Google*
Allison Woodruff, *Google*
Mary Ellen Zurko, *Cisco Systems*

Invited Talks Chair
Yang Wang, *Syracuse University*

Lightning Talks and Demos Chair
Elizabeth Stobert, *ETH Zürich*

Panels Chair
Tim McKay, *Kaiser Permanente*

Posters Co-Chairs
Michelle Mazurek, *University of Maryland, College Park*
Florian Schaub, *Carnegie Mellon University*

Tutorials and Workshops Co-Chairs
Adam Aviv, *US Naval Academy*
Mohammad Khan, *University of Connecticut*

Publicity Chair
Patrick Gage Kelley, *University of New Mexico*

Steering Committee
Lujo Bauer, *Carnegie Mellon University*
Konstantin Beznosov, *University of British Columbia*
Robert Biddle, *Carleton University*
Sunny Consolvo, *Google*
Lorrie Cranor, *Carnegie Mellon University*
Simson Garfinkel, *National Institute of Standards and Technology (NIST)*
Jason Hong, *Carnegie Mellon University*
Heather Richter Lipford, *University of North Carolina at Charlotte*
Andrew Patrick, *Office of the Privacy Commissioner of Canada*
Stuart Schechter, *Microsoft Research*
Matthew Smith, *University of Bonn*
Mary Ellen Zurko, *Cisco Systems*

External Reviewers

Yasemin Acar
Muhammad Adnan
Tousif Ahmed
Nirav Ajmeri
David Crandall
Julie Haney
Qatrunnada Ismail
David Llewellyn-Jones
Billy Melicher
Nicholas Micallef
Kristopher Micinski
Pradeep Murukannaiah
Sameer Patil
Tobais Seitz
Karthik Sheshadri
Blase Ur
Akash Verma

SOUPS 2016: Twelfth Symposium on Usable Privacy and Security
June 22–24, 2016
Denver, CO

Thursday, June 23, 2016

Security Interfaces

Behavior 1

Encryption and Surveillance

(Thursday, June 23 continues on the next page)

Friday, June 24, 2016

Authentication

Behavior 2

Privacy

SOUPS 2016
Twelfth Symposium on Usable Privacy and Security
Message from the Chairs

Welcome to SOUPS 2016!

The Twelfth Symposium on Usable Privacy and Security has been a year of transition for us. For the first time, we are not sponsored and held by Carnegie Mellon's CyLab. We are a USENIX conference this year (and we look forward to being one again in 2017). In addition, SOUPS founder Lorrie Faith Cranor stepped down as General Chair in January, since her new position as FTC Chief Technologist precluded her from dealing with industry sponsorships. We look forward to hearing much more about her FTC experience to date as our keynote speaker! Mary Ellen Zurko took over as General Chair on an accelerated timeline ("Some are born great, some achieve greatness, and some have greatness thrust upon 'em.") Every member of the organizing committee thus had additional responsibilities with this transition, and they all stepped up splendidly. USENIX learned about SOUPS, SOUPS learned about USENIX, and SOUPS 2016 is the result.

We retained the traditional SOUPS structure, with workshops (5), a tutorial, technical papers (22), posters (with a happy hour), lightning talks, demos, a panel, an invited talk, a social event, and an ice cream social. Please visit our Web site to learn the results of the SOUPS 2016 awards—Distinguished Paper, IAPP SOUPS Privacy Award, Distinguished Poster, and the John Karat Usable Privacy and Security Student Research Award (new this year).

This year we received 79 technical paper submissions. The 30-person program committee provided two rounds of reviews. In the first round, each submission received at least three reviews, and in the second round, some submissions received additional reviews. In the end, all submissions received at least three and as many as six reviews. After the second round of reviews, authors had an opportunity to provide a short rebuttal to respond to the reviews. Following the rebuttal period and an online discussion, the program committee held an in-person one-day meeting, which resulted in 22 papers being selected for presentation and publication. The acceptance rate was ~28%.

We would like to thank all of the authors and the members of the technical papers committee and every member of the SOUPS organizing committee for helping to produce this program. We would also like to thank the International Computer Science Institute in Berkeley, CA for hosting the in-person PC meeting. We especially thank the US National Science Foundation, Google, Facebook, Cisco, CRC Press, DMTF, and LXer for their sponsorship of this event. And finally, a big shout-out to each member of the USENIX staff for all the additional work they took on with producing SOUPS this year, including our liaison, Casey Henderson.

See you next year, July 12–14, at the Hyatt Regency Santa Clara!

Mary Ellen Zurko, *Cisco Systems*
General Chair

Sunny Consolvo, *Google*
Technical Papers Co-Chair

Matthew Smith, *University of Bonn*
Technical Papers Co-Chair

Rethinking Connection Security Indicators

Adrienne Porter Felt[1], Robert W. Reeder[1], Alex Ainslie[1], Helen Harris[1], Max Walker[1], Christopher Thompson[2], Mustafa Emre Acer[1], Elisabeth Morant[1], Sunny Consolvo[1]
Google[1], UC Berkeley[2]
security-enamel@chromium.org[1], cthompson@cs.berkeley.edu[2]

ABSTRACT

We propose a new set of browser security indicators, based on user research and an understanding of the design challenges faced by browsers. To motivate the need for new security indicators, we critique existing browser security indicators and survey 1,329 people about Google Chrome's indicators. We then evaluate forty icons and seven complementary strings by surveying thousands of respondents about their perceptions of the candidates. Ultimately, we select and propose three indicators. Our proposed indicators have been adopted by Google Chrome, and we hope to motivate others to update their security indicators as well.

1. INTRODUCTION

Security indicators are the most commonly seen browser security UI. Every major browser displays security indicators — a lock, a shield, or some other symbol — to summarize the security states of websites. (Figure 1 shows an example of a green lock in Google Chrome.) Yet, despite this ubiquity, people often find browser security indicators confusing.

Researchers have cautioned since 2002 that people don't always understand security indicators [7, 8, 16]. Two anecdotal experiences convinced us that the problem remained. While doing field work in India, we met many tech-savvy people who didn't associate Google Chrome's security indicators with security. Later, we discovered that one author's American sibling was similarly confused. This spurred us to formally revisit the problem of security indicators.

Our goal is to create new security indicators that non-expert browser users can understand. Ideally, security indicators should at least communicate whether a given website connection is currently secure or dangerous. We focus specifically on comprehension, leaving the question of how to draw attention to the indicators for future work [5].

In order to improve security indicators, we first needed to learn more about the shortcomings of existing indicators. We surveyed 1,329 people about Google Chrome's connection security indicators in the course of their normal web browsing. Although most of our tech-savvy (but non-expert) respondents had at least a basic understanding of the HTTPS indicator, many were unfamiliar with the HTTP indicator.

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22–24, 2016, Denver, Colorado, USA.

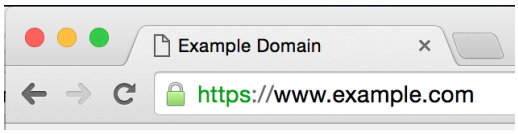


Figure 1: The green lock is a security indicator.

We then began the task of creating and testing new security indicators, working within the additional constraints posed by modern browser needs. Browsers are used by diverse audiences on diverse devices. Security indicators therefore face several design constraints:

- The indicators need to scale down for small devices. Icons should not rely on small decorations that become illegible when small. We can optionally use text, but there will not always be space to display it.
- The icon shape alone — without color — needs to communicate the level of risk to meet accessibility needs. 8% of men are colorblind [17], and many others have vision impairments.
- The indicator's meaning needs to be taught with words when possible. Millions of new Internet users have recently come online via smartphones without learning "standard" iconography from desktop browsers.

We identified forty candidate icons and seven accompanying strings that meet these constraints. Through a series of surveys, we narrowed the set down to the most promising icons and strings. Ultimately, we selected three sets of browser security indicators based on survey results, prior research, and our design constraints. Our proposed indicators will be deployed with Google Chrome 53.

Contributions. We contribute the following:

- Most security indicator research was performed in 2002 – 2008, but requirements have changed over time. We evaluate browsers' security indicators and determine whether they meet modern browser users' needs.
- We identify specific shortcomings of Chrome's connection security indicators with an in-the-moment survey of 1,329 respondents.
- We propose three new security indicators, based on multiple rounds of user testing and our constraints.

2. THE ROLE OF SECURITY INDICATORS

Browsers use security indicators to communicate connection security states, website trustworthiness, or a combination of the two. Security indicators are trusted browser UI, and they appear in or near the URL bar. They are distinct from website-controlled UI (such as favicons), although websites sometimes do use icons that appear similar to security icons. For example, the favicon for the website shown in Figure 1 looks extremely similar to Chrome's HTTP indicator.

2.1 Connection security

Connection security describes *how* a website was fetched over the network. Ideally, the HTTP connection should use well-authenticated TLS to protect end users' web traffic from eavesdroppers and attackers on the network.

Valid HTTPS. This is the best case scenario. The browser can establish a valid TLS connection to the server. The connection is private and tamper-free, even in the presence of malicious parties on the network. However, the website itself could be malicious or compromised; HTTPS only provides security guarantees about the connection.

HTTPS with minor errors. Although the browser was able to establish a valid TLS connection, there are minor problems (e.g., including an image over plain HTTP).

HTTPS with major errors. This is the worst case scenario. The website was supposed to load over HTTPS, but the certificate chain fails to validate. Most browsers show a warning that might (or might not) be overridable.

HTTP. The connection does not use HTTPS, so anyone on the network can see or modify the contents of the website. Although HTTP used to be the default for web browsing, more than half of page loads are now over HTTPS [9].

2.2 Website trustworthiness

In addition to connection security, browsers may also want to check whether the website itself is trustworthy.

EV HTTPS. A website can pay a certificate authority to confirm the website's identity, and the certificate authority will issue an Extended Validation (EV) certificate with the organization's name. EV was originally envisioned as a strong phishing defense.

Malware and phishing. Browsers may perform phishing and malware checks on websites. Services like Microsoft SmartScreen [2] and Google Safe Browsing [6] provide phishing and malware verdicts for browsers. Many browsers show full-page malware warnings.

3. RELATED WORK

Security indicators were well-studied in the mid-2000s, and this literature motivated a shift in how browsers treated security indicators. Security indicators used to be displayed in several areas (e.g., the bottom right corner of the browser), but browsers moved the indicators into the URL bar.

Warnings are complementary to indicators for communicating security issues to users, and have also received considerable research attention. While full coverage of the warnings literature is out of scope for this paper, readers may consult Sunshine et al. [21] and Sotirakopoulos et al. [20] as works specifically on connection security (i.e., SSL/TLS) warnings.

3.1 Connection security

Connection security indicators have received mixed results over the last fifteen years of research.

People look at indicators. Using eye tracking, Whalen and Inkpen found that most of their lab study participants looked at the lock icon while performing common online tasks [23]. Although some participants were confused about the significance of the icons, Whalen and Inkpen advised browser vendors against changing the lock. *"Making major modifications to this [lock] symbol, such as using a different object, may be disorienting: users now expect to find a lock in a browser window."*

Some people understand indicators. Friedman et al. interviewed people from a rural community in Maine, a suburb in New Jersey, and a Silicon Valley community [8]. Across these three communities, roughly half of participants could identify a secure connection from browser screenshots. While not terrible, we hope that someday more than half of users will understand how to differentiate secure and insecure connections. Lin et al. found similar results; some (but not all) of their participants knew about connection security indicators and checked them during study tasks [12].

No one heeds indicators. In contrast, Schechter et al. found that security indicators fail to change user behavior [16]. *None* of their participants withheld their passwords when asked to log in to their bank over HTTP. Similarly, several people incorrectly told Dhamija et al. that a lock icon is *"more important when it is displayed within the page than if presented by the browser"* [7].

Some mobile indicators are lacking. Amrutkar et al. studied SSL indicators in mobile browsers, where screen space is limited [3]. They found high rates of non-compliance with web security user interface standards on connection security indicators. In some cases, mobile browsers lacked any indicator at all of potential attacks, such that even experts would not have enough information to detect these attacks.

We expand on prior literature by evaluating Google Chrome's existing security indicators at much larger scale. With more than a thousand respondents, we were able to collect a broad, nuanced set of qualitative data. Furthermore, all of the cited studies took place in laboratories as either semi-structured interviews or researcher-directed tasks. Our survey respondents naturally encountered security indicators in the course of browsing on their own computers.

3.2 Website trustworthiness

In the past, HTTPS was viewed as a sign of website trustworthiness; getting a valid HTTPS certificate was too difficult for typical phishing websites. Dhamija et al. challenged 22 people to identify phishing websites, and 17 of them failed to check the connection security indicator during the study [7]. This demonstrated that connection security indicators were ineffective at preventing phishing attacks. Subsequently, HTTPS has ceased to be a useful signal for identifying phishing websites because it is no longer unusual

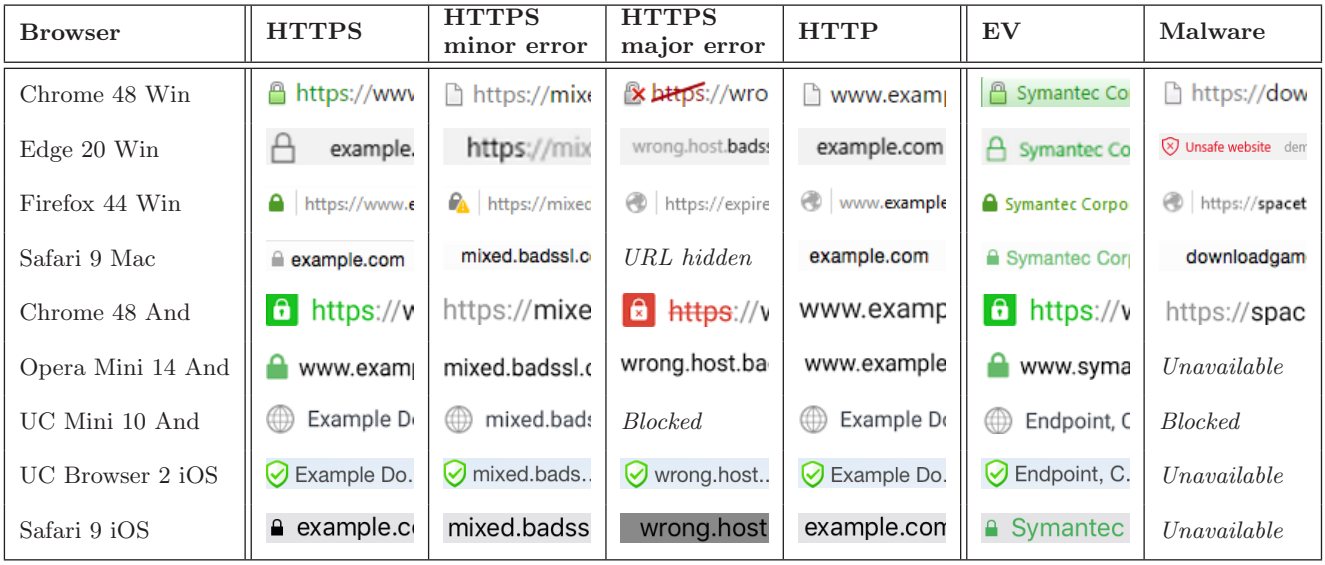


Figure 2: Security indicators for major browsers on Windows (Win), Mac, Android (And), and iOS. For categories that trigger warnings (e.g., malware), we include the security indicator state during the warning.

to find malicious websites that support HTTPS. We therefore do not aim to use HTTPS as an anti-phishing defense.

EV is an anti-phishing defense, although its use is limited by lack of support from popular websites and some major mobile browsers. All major desktop browsers display EV information, but some mobile browsers (including Chrome and Opera for Android) do not display EV information. Older literature suggests that EV indicators may need improvement. Jackson et al. asked study participants to identify phishing attacks and found that "extended validation did not help users defend against either attack" [10]. When testing new security indicators, Sobey et al. concluded that Firefox 3's EV indicators did not influence decision making for online purchases [19]. Improving EV indicators are out of scope for our current work.

3.3 Security indicator proposals

We propose changes to browser security indicators, and our proposal draws from prior research.

Sobey et al. suggested expanding security indicators into a "chip" that provides both an icon and explanatory text [19]. We like this format because it teaches and contextualizes the icon. However, Sobey et al. found that half of study participants did not notice the chip [19]. We have restricted our focus to comprehension, but their results suggest that we will need to do additional future work to draw attention to security indicators.

Maurer et al. proposed changing the entire toolbar to reflect the connection security state [14]. They surveyed participants about their proposal (and Firefox's existing security indicators) using a Firefox extension. With their proposal, study participants found valid HTTPS websites more trustworthy. In practice, however, we find it unlikely that a browser vendor would adopt a proposal that consumes the entire toolbar area as a security indicator. We took a similar methodological approach (using an extension) to survey people about Chrome's security indicators, but our surveys focused on comprehension instead of trustworthiness.

Although we specifically study security indicators, closely related UI also influences users' perceptions of security. For example, domain highlighting emphasizes the hostname in the URL bar (and de-emphasizes the potentially confusing path). Lin et al. found *"that domain highlighting works [to identify phishing], but nowhere near as well as we would like"* [12]. And what UI should be displayed when the user clicks on the security indicator? Biddle et al. proposed a way to display the identity information associated with the HTTPS connection [4]. Their proposal helped study participants find web site ownership and data safety information.

4. CURRENT BROWSER INDICATORS

Figure 2 illustrates how different security states are represented in major desktop and mobile browsers, according to our testing in February 2016. We describe and critique them to motivate the need for improved security indicators.

Similar shapes. Chrome and Firefox overload the meanings of shapes. Firefox's two lock icons have different meanings: a green lock for HTTPS, and a gray lock with a small yellow triangle for HTTPS-with-minor-errors. Chrome similarly has two locks: a green lock for HTTPS, and a red lock with a slash for HTTPS-with-major-errors. In both cases, the states look similar — particularly at small scale — unless the viewer is already familiar with the meaning. Chrome further compounds the problem by using colors that colorblind people commonly cannot distinguish.

Secure but untrustworthy. Most browsers use security indicators primarily to convey connection security information. If a browser's security indicator reflects only connection security, the browser can end up in a confusing state. When a user clicks through a warning to a malicious website, the browser will show a neutral or positive indicator in the URL bar. This might cause a user to believe the website is safe despite having seen the warning. Edge notably mitigates this by updating the security indicator to reflect malware or phishing verdicts.

Evergreen indicators. UCWeb's browsers (UC Mini and UC Browser)[1] stand out from other browsers by not displaying connection security information. Neither distinguishes between HTTP and HTTPS. UC Browser for iOS always displays a green shield regardless of the connection security state, which provides a sense of unmerited security.

Missing HTTP indicators. Many major browsers lack any indicator at all for HTTPS with minor errors or HTTP. As a result, the user does not have a click target to learn more about the connection security state. This is arguably reasonable on mobile, where small screens might necessitate removing or hiding indicators by default. However, desktop URL bars have sufficient space for an indicator.

HTTPS with minor errors. With the exception of Firefox, most browsers treat HTTPS with minor errors as if it were HTTP. We agree with this decision. This state is less risky than HTTP, but the website does not deserve to be displayed as fully secure. This state often occurs when websites are transitioning from HTTP to HTTPS. If we were to make the minor error state look *worse* than HTTP, it would discourage transitioning.

5. PERCEPTIONS OF CHROME'S SECURITY INDICATORS

We surveyed 1,329 people to understand user perceptions of Chrome's security indicators. We hoped to learn what people think Chrome's HTTPS and HTTP indicators mean, with an emphasis on identifying common misconceptions.

5.1 Method

We built a Chrome extension to deliver in-context surveys about Chrome's connection security indicators. The extension enabled us to survey respondents about indicators immediately after the respondents had an opportunity to see an indicator during normal browsing. Supplementary screenshots of the extension are in Appendix A, and the extension code is available on GitHub.[2]

5.1.1 How the extension worked

Setup. Immediately after installation, the extension displayed a consent form. If a respondent consented, s/he was then shown a short demographic survey, after which the extension shut down for a fifteen minute quiet period. Since the extension was intended for use with additional surveys later, we wanted respondents to learn that they would see surveys during regular browsing and not just upon installation of the extension.

Notification. After the quiet period ended, the extension waited until the respondent visited an HTTP or valid HTTPS website. (We avoided websites with major or minor errors by using a whitelist of popular websites without HTTPS errors.) When a qualifying website loaded, the extension prompted the respondent with a system notification to take a survey. If the respondent clicked on the notification, a survey would appear in a new window. Respondents were only notified once, and the extension stopped offering the survey after six hours from installation. So, not all people who installed the extension provided a survey response.

Survey. We created two versions of the survey, one for HTTP and one for HTTPS. The appropriate survey was selected based on the first website the respondent visited that triggered a survey notification.

5.1.2 Deployment

Our extension was publicly available for download in the Chrome Web Store, which is Google's official central repository for Chrome apps and extensions. We encouraged downloads via a press release, which was picked up by several popular tech news sources (e.g., [11, 18]) and a post in Chrome's help forum [13]. The promotional materials offered an opportunity to provide feedback on Chrome.

We collected surveys from May 11, 2015 to September 10, 2015 (122 days). We received 5,041 completed demographic surveys, and 1,329 completed HTTP(S) surveys, including 733 HTTPS surveys and 596 HTTP surveys.

To preserve respondent privacy, we chose not to monetarily compensate respondents. This decision allows us to collect data pseudonymously.

5.1.3 Questions

We asked respondents to describe the meaning of the indicators. To contextualize our question, the survey prominently included a screenshot of Chrome's URL bar with a red circle around the security indicator. The HTTPS and HTTP versions had screenshots of the appropriate indicators. Beneath the screenshot were the instructions:

> *You just now saw a URL bar like the one shown above. The following questions are about the URL bar.*

Each survey included three questions. In this paper, we focus on responses to the second question. (The other two are available in Appendix B, along with a screenshot of the survey in Appendix A.) We asked two versions, one for HTTPS and one for HTTP:

> HTTPS: *What does the green symbol to the left of the URL mean to you?*

> HTTP: *What does the white symbol to the left of the URL mean to you?*

5.1.4 Data coding

Seven security experts coded the qualitative responses. One team member (the *codemaster*) used open coding to create an initial codebook, in consultation with another expert. The remaining six coders did two partial coding rounds, each time giving feedback to the codemaster about shortcomings in the codebook. In the second round, all coders coded the same 40 responses to measure consistency. Fleiss's κ, a measure of inter-rater reliability, was 0.81, which we considered sufficiently consistent to proceed.

For the final round of coding, the codemaster divided the 1,329 responses between three pairs of coders. The coders

[1] `http://www.ucweb.com/company/about/`
[2] `https://github.com/GoogleChrome/experience-sampling`

worked in pairs so that two people independently coded each response. Each coder was responsible for approximately 400 responses, split between HTTP and HTTPS responses. Fleiss's κ was 0.89 before the codemaster reconciled remaining conflicts. Coders agreed on codes for 91% of responses, whereas 9% required resolution. The codemaster resolved the conflicting responses.

5.1.5 Demographics

While we hoped to reach a representative sample of Chrome users, our recruiting method may have provided a biased sample. In particular, we could not control which publishers ran our press release. Based on our demographic survey, respondents were most likely to learn about our survey from the Chrome Web Store, TechCrunch, omgchrome.com, and Reddit. These websites cater to technology enthusiasts, so our sample may be biased toward tech-savvy users. Furthermore, our decision to preserve respondent privacy by using non-compensated volunteers may have attracted a sample of people excited about improving Chrome.

Table 1 summarizes the demographics of our sample. Compared with Wash and Rader [22], a recent usable security paper that emphasized a representative sample of US Internet users, our sample skews young and heavily male. Educational level is closer to Wash and Rader's sample, though ours is skewed somewhat toward higher educational levels. Our sample was international (65% from outside the US), so cannot be expected to mimic Wash and Rader exactly, but we still note that the sample skews young and male.

Nevertheless, the sample is moderately large, at 5,041 installs and 1,329 survey responses. It is also diverse across age, educational level, and geography. Our survey was in English, and we filtered out non-English responses, but our respondents nonetheless were heavily international. The size and diversity of the sample suggest that the responses we received represent the understanding of a significant portion of the Chrome user population. And, since the bias is likely toward the tech-savvy, our results are likely an upper bound on the true understanding of security indicators amongst the general Chrome user population. That is, since our results show a lack of understanding of the indicators even amongst our sample, the understanding amongst all Chrome users is likely even lower. Our ultimate conclusion that users at large could benefit from a redesign of the indicators still holds.

5.1.6 Ethics

Consent. Respondents were shown a consent form that explained how the survey platform worked and how their answers would be used. If they did not consent, the extension would automatically uninstall itself. If they did consent, they proceeded to the demographic questionnaire. Respondents could view the consent form again later by clicking on "What is this?" in the extension notification.

Minors. Respondents needed to be age 18 or older. If a respondent claimed to be below the age of 18 in the demographic survey, the extension automatically uninstalled itself without sending any data to our server.

PII. We did not ask respondents to provide any personally identifiable information. The questions focus on the respondents' opinions of and beliefs about Chrome's security UI, as well as general demographic information. Each installation was assigned a random pseudonymous identifier to link demographic surveys with HTTP(S) surveys, but we cannot link the pseudonyms to individual people.

	Respondents	Installers
Male	90.4%	81.0%
Female	7.0%	14.2%
Other or not specified	2.6%	4.8%
Age 18-24	30.1%	25.8%
Age 25-34	40.7%	33.9%
Age 35-44	18.3%	20.0%
Age 45-54	6.9%	10.1%
Age 55-64	2.7%	6.3%
Age 65 or over	1.3%	3.8%
Some High School	2.6%	7.0%
HS or equiv	40.6%	48.9%
College degree	33.3%	28.2%
Graduate degree	20.2%	16.6%
Prefer not to answer	3.3%	6.4%
US	35.4%	27.8%
France	10.0%	6.8%
UK	5.9%	4.0%
Russian Federation	5.8%	4.4%
Germany	5.7%	3.7%
Canada	2.6%	3.6%
Other	34.6%	49.7%

Table 1: Demographics of the 1,329 respondents who provided completed surveys and of all 5,041 people who installed our extension.

Approval. Our organization does not have an IRB, but our study was internally reviewed before launch.

5.2 Results

We analyze responses to *What does the (white|green) symbol to the left of the URL mean to you?* by examining how many responses fall into each of our categories. We find that most respondents understand the HTTPS indicator, but are less sure about the meaning of the HTTP indicator. Table 2 summarizes the responses, including representative quotes.

5.2.1 HTTPS survey

Almost all of the 733 respondents mentioned security-related concepts when describing the green lock indicator. We categorized survey responses into seven high-level categories — CONNECTION, IDENTITY, PROTOCOL, SECURITY, ICON APPEARANCE, DON'T KNOW, and INCORRECT THEORIES — and ordered them by technical correctness and completeness, with CONNECTION demonstrating the most knowledge and INCORRECT THEORIES demonstrating the least. As shown in Table 2, most responses were at least partially correct; a majority fell in the first four categories, although the responses contain varying levels of technical depth and sophistication. We explain the categories, codes, and corresponding results in more detail below.

Connection and Identity. Responses in these categories are the most technically sophisticated and nuanced. CON-

HTTPS Category	Responses	Representative Quotes
CONNECTION	40.1%	
Encrypted connection	18.8%	"a secure encripted page"; "Connection is encrypted by HTTPS/SSL."
Secure connection	17.0%	"Secure connection"; "Secure connection I associate with the https vs http"
Safe to enter data	2.2%	"this site is safe to proceed to send data"
Private connection	1.5%	"The connection is private"; "It's a secure and private session"
Connection in general	0.6%	"https connection"
IDENTITY	13.4%	
Valid certificate	8.6%	"Secured connection, valid certificate"
Verified or authenticated	2.6%	"it's a verified domain – it's safe"
Trusted site	1.0%	"that is's a trustworthy page with a known identity."
Authority/Root CA/Chain of trust	1.0%	"...the certificate is in my database of trusted CA."
Identity applies only to name	0.3%	"does not guarantee the identity of the recipient (other than the hostname that is)"
PROTOCOL	34.4%	
HTTPS	18.7%	"HTTPS-using website."; "Secured via HTTPS"
SSL	12.1%	"SSL"; "SSL is enabled on the current site"
TLS	2.5%	"The page was served over TLS"; "That the site is SSL/TLS"
Secure form of HTTP	1.0%	"secure http"; "Site using encripted http"
SECURITY	35.7%	
Security or safety in general	23.7%	"Security."; "Security, safe, protection"
Secure site or page	12.0%	"The website is secure"; "Is a secure page"
ICON APPEARANCE	0.4%	
Lock	0.4%	"locked"; "closed lock = locked..."
DON'T KNOW	0.6%	
Don't know	0.6%	"I do not Know."
INCORRECT THEORIES	0.4%	
Miscellaneous	0.4%	"it is password?"; "website has user secured information on it"

HTTP Category	Responses	Representative Quotes
NOT SECURE	21.2%	
Not secure in general	10.9%	"This web page is purely a web page with no security"; "The page is unsecure"
Not encrypted	6.8%	"An unencrypted connection to the site."; "Unencrypted transmission of the page."
Insecure connection	2.0%	"white symbol to me means unsecure connection and page info."; "Unsecure connection"
PROTOCOL	17.4%	
Not HTTPS	6.6%	"Means that it is not https"; "Unencrypted connection (non-HTTPS)"
HTTP	4.3%	"unencrypted page transmitted over http protocol"; "http"
HTTP and not HTTPS	1.9%	"HTTP, not HTTPS"; "The site is being served via HTTP rather than HTTPS"
Protocol in general	1.5%	"Web protocol + Certificate"; "It represents either the favicon or the security protocol..."
HTTPS	0.6%	"security something (https?)"; "https I think?"
Not TLS	0.4%	"It's not TLS/SSL secures. so no https"
ABOUT SECURITY	7.1%	
Security in general	6.2%	"Security"; "Safety!"
Connection in general	0.6%	"The type of connection that was made with the server."
Site identity in general	0.4%	"Whether or not the identity of the site is verified"
REGULAR WEBPAGE	8.4%	
Regular webpage	8.4%	"regular web page"; "I am looking at a regular web page with no known issues"
CONTEXT MENU ITEMS	23.8%	
Site information	11.8%	"Provides Site Information"; "Click - see details for website"
Cookies	4.7%	"cookies"; "It gives a quick glance at permissions and cookies."
Permissions	2.4%	"information about privacy permissions"
SSL certificate status	1.9%	"Information on current page (cookies, ssl certificat)"
Connection	1.7%	"Access to the details of the connection to the site."
Security status	1.3%	"It offers information about the security of the webpage you are visiting."
ICON APPEARANCE	5.3%	
Document	2.6%	"document"; "Something to do with paper or a document..."
Page	1.3%	"page icon"
Piece of paper	0.9%	"Something to do with paper or a document..."
File icon	0.4%	"For me this symbol is the 'computer' file symbol..."
FAVICON	9.4%	
Website with no favicon	6.8%	"no favicon"; "No favicon for the current website."
Is the favicon for the site	2.4%	"Favicon"; "the site icon"
OTHER FUNCTIONALITY	1.7%	
Make a bookmark	1.1%	"A link to easily create a shortcut."
Drag the URL	0.6%	"THat's where I click whan I want to drag the URL..."
DON'T KNOW	7.1%	
Don't know	7.1%	"i just dont know."; "no idea"
NO MEANING	0.9%	
No meaning	0.9%	"nothing"; "...It mean nothing."
INCORRECT THEORIES	2.4%	
Bookmark indicator	0.6%	"I think it signifies that the page is saved as a bookmark."
Page loading	0.4%	"The page is loaded."; "The page hasn't loaded entirely."
Trouble loading	0.4%	"Trouble loading page"
SECURE	1.5%	
Secure page	1.3%	"secure site"; "Th url is safe"

Table 2: The percentage of responses that fell into each category, and representative quotes. Percentages do not add up to totals because some responses received multiple codes. Responses are verbatim, except as indicated by ellipses.

NECTION was the most-mentioned category, applying to 40.1% of responses. IDENTITY, at 13.4%, was the fourth most-mentioned. An expert would ideally mention both.

The CONNECTION category is the most unambiguously correct category. Responses within this category fell into five sub-codes, four of which explicitly mention connection security: *Encrypted connection*, *Secure connection*, *Private connection*, and *Connection in general*. A fifth code, *Safe to enter data*, was assigned to responses that did not explicitly mention the connection but indicated that the data exchanged with the server could not be intercepted.

The IDENTITY category is more complex. With an HTTPS connection, the browser verifies the server's identity to make sure the client isn't accidentally talking to a man-in-the-middle attacker. Some IDENTITY codes correctly refer to this process by talking about a *Valid certificate*, *Authority/root CA/chain of trust*, or how the *Identity applies only to domain*. However, HTTPS alone does not provide any guarantees that the website is trustworthy or the right website for the user's task. Some respondents mentioned identity but incorrectly said that HTTPS vouched for the website's trustworthiness (*Verified or authenticated*, *Trusted site*). This is an unfortunate misconception, although it was rare (about 3% of the total).

Protocol. A third of responses (34.4%) correctly mentioned the protocol. These responses mentioned *HTTPS*, *SSL*, *TLS*, or a *Secure form of HTTP*, which demonstrates an association between the indicator and protocol. However, we cannot tell whether a respondent understands what HTTPS is just by mention of the name, so these codes do not necessarily indicate an understanding of the protocol.

Security. The second most-mentioned category at 35.7%, SECURITY, included responses that mentioned security in a general sense, without necessarily mentioning the TLS guarantees or any of the protocols. Some responses in this category mentioned *security or safety in general*, while others mentioned security or safety in the context of a site or page.

Icon appearance, Don't know, Incorrect theories. The last three categories were rarely assigned for the HTTPS indicator. Responses in ICON APPEARANCE mentioned the literal appearance of the icon, namely that it depicts a *lock*. Responses in DON'T KNOW explicitly stated that respondents did not know what the HTTPS indicator meant. Responses in INCORRECT THEORIES suggested miscellaneous incorrect meanings for the indicator.

5.2.2 HTTP survey

Codes for the HTTP survey reflect a greater variety of responses than we observed for the HTTPS survey, and respondents displayed less knowledge about HTTP. Table 2 shows results from the HTTP survey.

We grouped responses into 12 categories, ordered by decreasing technical correctness and completeness: NOT SECURE, PROTOCOL, ABOUT SECURITY, REGULAR WEBPAGE, CONTEXT MENU ITEMS, ICON APPEARANCE, FAVICON, OTHER FUNCTIONALITY, DON'T KNOW, NO MEANING, INCORRECT THEORIES, and SECURE. We explain categories, codes, and corresponding results for the HTTP study below.

Not secure. About a fifth of responses (21.2%) correctly say that the security guarantees of TLS are not in place. Most of the NOT SECURE responses indicated that something (the page, the site, or no subject at all) was *not secure in general*. Others more specifically named the connection and noted that it was *not encrypted* or *insecure*.

Protocol and About security. As with our HTTPS survey, many responses mentioned a protocol or talked about security in general (17.4% for PROTOCOL, 7.1% for ABOUT SECURITY). Within the PROTOCOL responses, people talked about HTTP using various synonyms, and the ABOUT SECURITY responses touched generally on connection security or identity. Unfortunately, a few of the PROTOCOL responses incorrectly suggested *HTTPS* was in use.

Context menu items and other functionality. Surprisingly, the most popular topic was about what the HTTP icon can do if clicked or dragged. 23.8% of responses talk about the CONTEXT MENU ITEMS that appear when someone clicks on the icon, and another 1.7% talk about OTHER FUNCTIONALITY. We did not see these types of responses for the HTTPS indicator, even though it has the same behavior when clicked or dragged. One potential explanation is that respondents who were unfamiliar with the HTTP icon clicked on it after reading our question, and then told us what they found.

Regular webpage. 8.4% of responses called HTTP websites "regular" or "normal." This reflects the prevalence of HTTP on the web.

Don't know, no meaning, and icon appearance. Some respondents simply didn't know what the HTTP indicator means. 7.1% responses said they DON'T KNOW, 1% said the icon has NO MEANING, and 5.3% simply described the ICON APPEARANCE without commenting on its functionality or meaning. These types of responses were more common than for the HTTPS survey.

Incorrect responses and secure. A small but still too-large number of respondents provided incorrect descriptions of the HTTP indicator. 9% of respondents thought the indicator was the default FAVICON, rather than a security indicator, and 2.4% had miscellaneous other incorrect theories. Unfortunately, 1.5% of responses thought that the HTTP indicator meant the opposite: that the page is SECURE.

6. EVALUATING NEW ICONS

With our survey (Section 5), we learned that even tech-savvy people hold incomplete or incorrect beliefs about Chrome's HTTP indicator. Since we see shortcomings in other browsers' security indicators as well (Section 4), we decided to create new security indicators. We began by searching for icons for our proposal and evaluating them with Google Consumer Surveys (GCS) [15]. Our goal was to determine which icon shape and color best represented secure and insecure connections to websites. We ultimately selected three shapes: a green lock, a black circle, and a red triangle.

In our analysis, we performed thirteen tests for statistical significance. To account for multiple testing, we adjusted our levels of significance using the Holm-Bonferroni method.

Figure 3: The candidate indicator shapes, split between positive (top) and negative (bottom).

6.1 Candidate icons

We began with forty candidate icons. They varied in three dimensions: shape, historical connotation, and color.

Shape. We selected eight shapes (Figure 3) that are commonly used in road signs or Google products to communicate safety information. They are all simple shapes that scale, and their profiles can be distinguished from one another.

Connotation. Four of the shapes have historically been used to communicate safety, and four of the shapes have historically been used to communicate danger. We considered the former to be candidates for a security icon, and the latter to be candidates for an insecure icon.

Color. We chose five colors: black, blue, green, orange, and red. We produced each shape in five colors.

6.2 Survey method

Questions. We ran two sets of surveys in September 2015 to evaluate which icons best represent a secure connection or insecure connection. The questions were, respectively:

- *Imagine each of the icons below next to a URL in your browser address bar. Which of the icons best represents a connection to the website that IS secure?*
- *Imagine each of the icons below next to a URL in your browser address bar. Which of the icons best represents a connection to the website that is NOT secure?*

To answer the question, respondents had to pick an icon from a pair. The two icons were different shapes but the same color. Each respondent answered the same question five times, once for each color. For example, a respondent might have to pick between a green lock and a green shield, then pick between a blue triangle and a blue checkmark, and so on. A screenshot in Appendix C.1 shows what the pairwise comparison looked like.

Recruitment. GCS surveys are published on news, reference, and entertainment websites. Respondents answer the survey questions to gain access to free content, in lieu of subscribing or upgrading. We did not directly pay respondents. Google paid the publisher for the responses.

Sample. Five hundred participants answered each variant of each question, which yielded a total of 7,000 responses from 1,000 respondents. We did not ask any demographic or personal questions, although Appendix D contains inferred demographics. All of our respondents were physically located in the United States at the time of the survey.

	Positive icons				*Negative icons*			
...IS secure?								
Black	**23%**	20%	18%	13%	8%	8%	5%	5%
Blue	20%	**21%**	17%	17%	7%	7%	5%	6%
Green	**23%**	20%	16%	12%	8%	10%	6%	4%
Orange	19%	**20%**	18%	18%	6%	9%	6%	4%
Red	19%	**20%**	19%	18%	7%	7%	5%	5%
...is NOT secure?								
Black	4%	8%	10%	6%	19%	14%	**21%**	19%
Blue	5%	8%	7%	8%	**21%**	19%	16%	16%
Green	3%	10%	7%	8%	19%	17%	**20%**	16%
Orange	6%	8%	9%	7%	**19%**	17%	17%	16%
Red	7%	6%	7%	6%	**21%**	18%	16%	19%

Table 3: How often each icon "won" when the respondent answered, *Which of the icons best represents a connection to the website that...* N=1000

Google Consumer Surveys are typically representative of the Internet-using population in the United States [15].

6.3 Survey results

Although respondents exhibited strong associations between icon shape and (in)security, no individual shape-color combination stood out. Table 3 shows our results.

Preconceived beliefs. We hypothesized that respondents would have preconceived beliefs about the icon shapes based on past experiences, and our data substantiates this hypothesis. Prior to running the experiment, we categorized our icon shapes as "positive" or "negative" based on how they are used in existing products. The "positive" icons were more likely to be considered secure than insecure, and the "negative" icons were more likely to be considered insecure than secure. We found a significant difference between between the positive icons' scores across the secure and insecure questions ($\chi^2 = 57.06, df = 3, p < 0.01$). Similarly, we found a significant difference between the negative icons' average scores in the secure and insecure questions ($\chi^2 = 42.91, df = 3, p < 0.01$).

Secure connection. Respondents did not have a clear favorite for a color-shape combination that represents a secure connection. The "positive" icons won at similar rates for the secure connection question, although the shield or lock won the most across colors.

Insecure connection. Respondents also did not have a clear favorite for a color-shape combination that represents an insecure connection. There "negative" icons won at similar rates for the insecure connection question, although the triangle placed either first or second across the five colors.

6.4 Icon selection

We had hoped that three clear winners would emerge from the forty icons: an icon strongly associated with a secure connection, an icon strongly associated with an insecure connection, and an icon moderately associated with an insecure connection. Although that did not happen, we can still look at the pairwise rankings to identify candidates.

Secure connection. The shield and lock consistently performed well across all colors, which suggests that either shape should be meaningful to people who are colorblind. We break the tie by considering that many browser users have already been taught to look for locks, and our tech-savvy extension survey respondents related it to security (Section 5). Over ten years ago, Whalen and Inkpen cautioned against changing the lock shape because their interviewees had begun to expect it [23]. Thus, we propose to continue using a green lock for HTTPS.

Insecure connection. The triangle and slash both tested as viable candidates. They jointly won all of the insecurity comparisons, and the slash ranked among the lowest on the security question. We break the tie by considering scalability and contrast; the blockier triangle will be easier to recognize at small scale on different backgrounds. Thus, we propose to use a red triangle for insecure connections.

Slightly insecure connection. To represent HTTP, we want to choose an icon from the "negative" group that is not strongly associated with either end of the spectrum. The circle with an exclamation point fits that criteria and also appears similar to the ISO symbol for information. We hope that the similarity would encourage people to click on it to find out more information about connection security. Thus, we propose to use a black circle with an exclamation point for connections over HTTP.

7. EVALUATING NEW TEXT

We hope that text can aid user comprehension of security indicators, particularly for new Internet users who do not have preexisting expectations of icons. But which strings should we use? Using Google Consumer Surveys, we tested a set of strings to see which helped comprehension the most.

7.1 Candidate strings

We paired each of the three icons with seven strings. The strings are simple phrases that convey slightly different threat models. The sets of candidate strings are:

- For the green lock: *"https," "private," "secure," "safe," "encrypted," "secure and private," "secure site"*
- For the black circle: *"http," "not private," "not secure," "not safe," "not encrypted," "not secure, not private," "site not secure"*
- For the red triangle: *"~~https~~," "not private," "not secure," "not safe," "not encrypted," "not secure, not private," "site not secure"*

Two designers selected the strings in consultation with security experts. Their simplicity should make them (relatively) easy to translate correctly. The black circle and red triangle strings are similar because they are both conveying insecure states, of different degrees of severity.

7.2 Method

Questions. We asked three GCS questions in November 2015 about website safety, each intended to capture a different aspect of security indicators. We wanted to understand how respondents perceive the safety of the page, threat model, and desired action given different security indicators. Our questions were:

1. *If you saw this browser page, how safe would you feel about the current website?*
   Not at all safe
   A little safe
   Somewhat safe
   Very safe
   Extremely safe
2. *If you saw the below icon and message in the browser's address bar, that would be that someone might...*
   Try to put a virus or malware on your PC
   Modify the content of the page
   Have created a technical bug on the site
   Steal the things you read and type
   None of the above
3. *If you came across a site in your browser and saw this in the address bar, how would you most likely proceed?*
   I'd browse normally
   I'd leave the site
   I wouldn't enter any credit card details
   I'd look for more information about the site
   I'd browse quickly, then leave

Each question was accompanied by a mock browser screenshot that included an icon, string, and blurred URL. We made 21 variants of each question because we had 21 combinations of icons and strings. An individual respondent answered all three questions for the same icon-string pair. For Q2, respondents could select multiple choices *or* "None." Responses were either randomly flipped (Q1) or randomly ordered (Q2 and Q3). Appendix C.2 shows an example question.

Q3 asks respondents how they would react to an indicator. Since this is self-reported data, it likely does not reflect actual behavior in the field. However, it gives us insight into how respondents *perceive* the indicators' calls to action.

Recruitment. GCS surveys are published on news, reference, and entertainment websites. Respondents answer the survey questions to gain access to free content, in lieu of subscribing or upgrading. We did not directly pay respondents. Google paid the publisher for the responses.

Sample. Three hundred respondents took each of our twenty-one variants, each of which consisted of three questions. This yielded 19,386 responses from 6,462 respondents. We did not ask any demographic or personal questions, although Appendix D contains inferred demographics. All of our respondents were physically located in the United States at the time of the survey.

7.3 Results

Respondents had different perceptions of page safety, threat models, and calls to action depending on the strings. Table 4 shows the full results.

7.3.1 Valid HTTPS

We find that "secure" and "https" are the most promising companions to a green lock icon.

Q1:		Not at all	A little	Somewhat	Very	Extremely
	https	23%	9%	**32%**	26%	10%
	Private	24%	16%	**35%**	18%	7%
	Secure	12%	19%	**40%**	24%	5%
🔒	Safe	20%	16%	**34%**	20%	10%
	Encrypted	23%	12%	**42%**	19%	4%
	Secure and private	20%	19%	**36%**	21%	4%
	Secure site	18%	17%	**32%**	24%	8%
	http	**40%**	20%	27%	11%	3%
	Not private	**60%**	17%	15%	4%	4%
	Not secure	**58%**	14%	19%	6%	4%
❗	Not safe	**61%**	12%	16%	7%	4%
	Not encrypted	**52%**	19%	18%	5%	6%
	Not secure, not private	**57%**	17%	18%	6%	2%
	Site not secure	**63%**	14%	14%	6%	3%
	~~https~~	**63%**	16%	12%	5%	4%
	Not private	**68%**	14%	11%	3%	5%
	Not secure	**61%**	22%	11%	2%	4%
⚠	Not safe	**65%**	14%	14%	5%	3%
	Not encrypted	**53%**	18%	19%	6%	5%
	Not secure, not private	**64%**	20%	11%	3%	2%
	Site not secure	**64%**	15%	12%	6%	4%
Q2:		**Malware**	**Steal**	**Bug**	**Modify**	**None**
	https	15%	10%	12%	14%	**64%**
	Private	24%	22%	16%	14%	**51%**
	Secure	15%	12%	12%	13%	**65%**
🔒	Safe	24%	19%	16%	14%	**54%**
	Encrypted	22%	15%	12%	16%	**56%**
	Secure and private	23%	18%	15%	17%	**53%**
	Secure site	18%	12%	9%	14%	**60%**
	http	30%	24%	22%	27%	**41%**
	Not private	41%	**48%**	29%	26%	21%
	Not secure	**51%**	37%	29%	24%	22%
❗	Not safe	**53%**	39%	29%	22%	25%
	Not encrypted	**36%**	38%	24%	23%	32%
	Not secure, not private	**50%**	42%	32%	26%	22%
	Site not secure	**48%**	40%	30%	25%	24%
	~~https~~	**47%**	34%	30%	26%	23%
	Not private	46%	**49%**	30%	27%	21%
	Not secure	**54%**	46%	32%	33%	20%
⚠	Not safe	**61%**	39%	25%	23%	20%
	Not encrypted	**43%**	39%	23%	28%	26%
	Not secure, not private	**50%**	37%	25%	23%	23%
	Site not secure	**61%**	43%	35%	29%	22%
Q3:		**Leave site**	**More information**	**No credit card**	**Normally**	**Quickly**
	https	20%	12%	12%	**51%**	5%
	Private	**28%**	19%	18%	25%	9%
	Secure	17%	15%	18%	**41%**	9%
🔒	Safe	26%	14%	14%	**37%**	10%
	Encrypted	28%	14%	18%	**33%**	7%
	Secure and private	25%	15%	20%	**31%**	9%
	Secure site	23%	16%	14%	**40%**	8%
	http	**38%**	10%	22%	21%	10%
	Not private	**53%**	13%	17%	9%	8%
	Not secure	**58%**	10%	16%	9%	7%
❗	Not safe	**66%**	6%	16%	7%	5%
	Not encrypted	**49%**	8%	25%	9%	9%
	Not secure, not private	**51%**	13%	21%	8%	7%
	Site not secure	**59%**	8%	20%	7%	7%
	~~https~~	**60%**	12%	14%	6%	8%
	Not private	**60%**	11%	15%	7%	7%
	Not secure	**54%**	12%	17%	11%	7%
⚠	Not safe	**68%**	9%	14%	6%	4%
	Not encrypted	**53%**	11%	21%	10%	5%
	Not secure, not private	**59%**	12%	17%	5%	8%
	Site not secure	**64%**	8%	15%	7%	7%

Table 4: Responses to the three GCS questions with both icons and strings. N=6462

Safety. Respondents associated different levels of safety with different strings, based on comparing all of the different outcomes to Q1 (chi-square = 101.30, $df = 24$, $p < .01$). "Secure" yielded the highest number of respondents who felt that the website was at least somewhat safe and the lowest number of participants who felt not safe at all.

Threat. Respondents were most likely to trust a page with the "https" and "secure" strings. The strings influenced the number of respondents who chose "none of the above" (vs. any other response) when asked what kinds of risks might exist on the page (chi-square = 68.23, $df = 6$, $p < .01$). Additionally, across all strings, respondents were unlikely to think that a website with a green lock might try to install malware. This suggests that our indicators are broadly perceived as security indicators, not specifically as connection security indicators.

Action. Respondents claimed they would take different actions depending on the strings (chi-square = 17.40, $df = 6$, $p < .01$), with "https" resulting in the highest number of respondents browsing normally and "secure" having the fewest respondents who would leave the website. We do not assume that respondents would necessarily take these actions, but this demonstrates differing perceptions of the strings.

7.3.2 Invalid HTTPS

We find that "not secure" and "site not secure" are the most promising companions to the red triangle.

Safety. There was a significant difference in how respondents perceived the safety of the website across all strings (chi-square = 71.62, $df = 24$, $p < .01$). Respondents viewed "~~https~~," "not secure", and "not encrypted" as the least safe.

Threat. For invalid HTTPS, "none of the above" is not a desirable answer. We compared how many respondents answered "none of the above" (vs. any other response) and observed a significant difference between the strings (chi-square = 18.51, $df = 6$, $p < .01$). The "not secure" and "site not secure" strings yielded the most respondents who believed at least one of the negative actions could occur.

Action. When faced with an insecure connection, the ideal user behavior is to leave the website. As a result, we compared the ratio of respondents who chose "I'd leave the site" to the total of the other options. The chi-square reveals a significant difference (chi-square = 35.40, $df = 6$, $p < .01$), with "not safe" and "site not secure" ranking highest.

7.3.3 HTTP

Our HTTP security indicator needs to communicate a state that is mildly insecure, but not as insecure as invalid HTTPS or a known malware page. Using "http" would yield the least alarming indicator, and "site not secure" the most alarming.

Safety. Respondents felt at least somewhat safe with the "http" string, whereas "not private" and "site not secure" had the lowest percentage of respondents who felt at least somewhat safe. The differences between strings were statistically significant (chi-square = 116.59, $df = 24$, $p < .01$).

Threat. Respondents were most likely to select "none of the above" (vs. any other response) with the "http" string, which we interpret to mean they felt safest with the "http" string. On the other hand, they were most likely to choose at least one negative consequence with "not private." The differences between the set of strings was statistically significant (chi-square = 110.68, $df = 6$, $p < .01$).

Action. When using an HTTP page, we want respondents to seek more information and/or avoid entering their credit card. Across the strings we observed a significant difference in responses when comparing the number of respondents who say they would perform one of the actions compared to browsing normally (chi-square = 63.08, $df = 6$, $p < .01$), with respondents most likely to browse normally with "http."

8. DISCUSSION

We draw out the implications from our extension survey, Google Consumer Surveys, and prior work.

8.1 Shortcomings of HTTP indicators

We want indicators to teach people that HTTP is less secure than HTTPS. Conveying the threat of a network attacker with an icon and three words is challenging, and we don't think that browsers are currently succeeding.

Most of our extension survey respondents did not relate Chrome's HTTP indicator to connection security, despite their tech-savvy demographics. It was a disappointing but unsurprising finding. We can't say *why* they failed to mention connection security; it could be lack of knowledge, or that it did not come to mind at the moment of the survey. Either way, indicators are supposed to be immediately recognizable and understandable without significant thought.

Although we did not test other browsers' security indicators, we would not expect them to fare better at explaining HTTP. Edge and Safari don't display any indicator at all for HTTP, and UCWeb browsers don't distinguish between HTTP and HTTPS. Firefox's globe is neutral, so we suspect people would view it much like Chrome's neutral page icon. This means that we do not think Chrome can solve its problem by copying other browsers' HTTP indicators.

We did learn, however, that understanding security icons is not impossible for non-experts. Nearly all of our extension survey respondents associated Chrome's green lock with HTTPS and security. Their beliefs — particularly around identity — were not always complete or correct, but they still understand the general concept of the indicator. Although these respondents were tech-savvy, they were not security experts, which makes us hopeful that others will also learn the meanings of indicators with sufficient nudges.

8.2 Proposed connection security indicators

We propose three security indicators, shown in Figure 4. The strings should smoothly collapse or re-appear, depending on the page state and device screen size.

Section 6.4 describes how we narrowed down our icon choices to the lock, circle, and triangle. After testing, we modified the circle icon to more closely resemble the ISO Information Symbol; we hope that it will attract clicks from curious users seeking further information about the website.

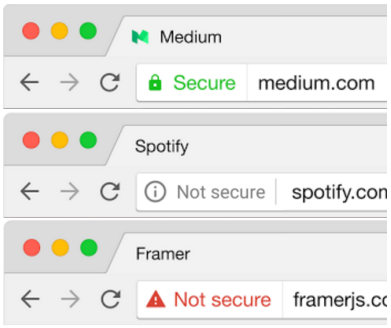


Figure 4: Proposed connection security indicators.

We chose strings after selecting the icons. For the positive security state, "secure" and "https" performed well across all three metrics (Section 7.3). Between the two, we preferred "secure" because it is less technical. We chose "not secure" for the neutral and negative states because it performed reasonably well and has a pleasing symmetry with "secure."

Chrome will launch our proposed connection security indicators with Chrome 53. However, we hope that our indicators are not limited to Chrome's URL bar. We would like to see other products that convey connection security adopt similar shapes to reinforce the meaning of the indicators. All of the icons are free to use as part of Material Design.[3]

Although we believe our changes are an improvement, open questions about HTTP remain. Our extension survey respondents did not connect HTTP with a lack of connection security. Despite our desire to teach people that HTTP is not secure, we do not want to frighten people from using the Internet. We therefore plan to gradually ease into the "not secure" label to avoid panicking people, beginning with private browsing mode because users are presumably performing privacy-sensitive tasks. Whether this is too conservative (or too aggressive) remains to be seen.

8.3 Malware security indicators

We can easily imagine why some end users do not distinguish between connection security indicators and website trustworthiness indicators. It is confusing, even to an expert, that clicking through a malware warning does not yield a negative security indicator in most browsers. In the extension survey, many tech-savvy people mistakenly believed that HTTPS identity guarantees pertain to website trustworthiness. Many GCS respondents similarly did not distinguish between the threat models.

Edge displays a negative security indicator for malware and phishing websites (Table 2). We recommend that other browsers, including Chrome, also use a negative security indicator for known malware and phishing websites.

8.4 Future work

Internationalization. One of our primary goals is to help new Internet users learn the meaning of security indicators. We added strings to the indicators specifically for this demographic. However, we have not yet tested the indicators in countries with many new Internet users; we only tested our icons with English-speaking Americans. Translation, cultural differences, or prior computing experiences might cause our results to not hold across countries. We need to do further work to find out whether we have achieved our full set of goals, although we expect that this will require a longitudinal field study to see whether people learn the meanings of the indicators over time. Thus, our next step is to test these indicators outside of the United States.

Repeat the survey. Once people have had time to acclimatize to the new icons, we should repeat the extension survey to see whether results remain the same. Will people be more likely to understand the HTTP indicator?

Attention. How might we draw users' attention to security indicators at the right time? (And when is the right time?) People sometimes ignore security indicators at crucial moments, or — worse — look within the content area of the website for the indicators [7]. Even if we were to train people to only look for security indicators in trusted browser UI, there are exceptions. Websites can add favicons to tabs, extensions can add icons near the URL bar, and so on. How might we teach people to look — and look in the right place?

9. CONCLUDING SUMMARY

We surveyed 1,329 people about Google Chrome's security indicators using a custom Chrome extension. Although our moderately tech-savvy respondents could relate Chrome's green lock to security, they had varying thoughts on the meaning of Chrome's neutral page icon. This motivated the need for new security indicators. Since existing security indicators from other browsers didn't entirely meet our design constraints, we set out to create new indicators.

We evaluated forty icons and seven complementary strings by surveying thousands of Google Consumer Survey respondents. Ultimately, we selected and proposed three indicators: Secure for HTTPS, Not secure for HTTP, and Not secure for invalid HTTPS. Our proposed indicators have been adopted by Chrome, and we hope to motivate others to update their security indicators as well. Our next step is to evaluate the indicators internationally, once they have been in use for several months.

10. ACKNOWLEDGMENTS

We thank Emily Stark and Lucas Garron for their help coding the extension survey responses, Chris Palmer for his input into the security icon redesign, and the rest of the Chrome security team for their feedback and support.

11. REFERENCES

[1] Inferred demographics. `https://support.google.com/consumersurveys/answer/6218151`.

[2] SmartScreen Filter: Frequently asked questions. `http://windows.microsoft.com/en-us/windows/smartscreen-filter-faq#1TC=windows-7`. Accessed February 2016.

[3] C. Amrutkar, P. Traynor, and P. C. van Oorschot. Measuring ssl indicators on mobile browsers: extended life, or end of the road? In *Information Security*, pages 86–103. Springer, 2012.

[4] R. Biddle, P. C. van Oorschot, A. S. Patrick, J. Sobey, and T. Whalen. Browser interfaces and Extended Validation SSL Certificates: An Ampirical Study. In

[3] `https://design.google.com/icons/`

Proceedings of the ACM Cloud Computing Security Workshop, 2009.

[5] C. Bravo-Lillo, S. Komanduri, L. F. Cranor, R. W. Reeder, M. Sleeper, J. Downs, and S. Schechter. Your attention please: designing security-decision uis to make genuine risks harder to ignore. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, page 6. ACM, 2013.

[6] G. Developers. SafeBrowsing API. `https://developers.google.com/safe-browsing/`. Accessed February 2016.

[7] R. Dhamija, J. D. Tygar, and M. Hearst. Why phishing works. In *Proceedings of ACM CHI*, 2006.

[8] B. Friedman, D. Hurley, D. C. Howe, E. Felten, and H. Nissenbaum. Users' conceptions of web security: A comparative study. In *Proceedings of ACM CHI*, 2002.

[9] I. Grigorik. HTTPS navigations in Chrome. `https://plus.google.com/+IlyaGrigorik/posts/7VSuQ66qA3C`, November 2014.

[10] C. Jackson, D. R. Simon, D. S. Tan, and A. Barth. An evaluation of extended validation and picture-in-picture phishing attacks. In *Proceedings of the International Conference on Financial Cryptography and International Conference on Usable Security*, 2007.

[11] F. Lardinois. Google launches Chrome extension to solicit user feedback about its browser. TechCrunch. `http://techcrunch.com/2015/05/11/google-launches-chrome-extension-to-solicit-user-feedback-about-its-browser`, May 2015.

[12] E. Lin, S. Greenberg, E. Trotter, D. Ma, and J. Aycock. Does domain highlighting help people identify phishing sites? In *Proceedings of CHI*, 2011.

[13] K. M. Help improve Chrome with this extension. `https://productforums.google.com/forum/#!category-topic/chrome/sUwEcHPygFU`, May 2015.

[14] M.-E. Maurer, A. De Luca, and T. Stockinger. Shining chrome: Using web browser personas to enhance SSL certificate visualization. In *Human-Computer Interaction – INTERACT*, 2011.

[15] P. McDonald, M. Mohebbi, and B. Slatkin. Comparing Google Consumer Surveys to existing probability and non-probability based Internet surveys. Technical report, Google Inc., 2015.

[16] S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer. The emperor's new security indicators: An evaluation of website authentication and the effect of role playing on usability studies. In *Proceedings of IEEE Symposium on Security and Privacy*, 2007.

[17] L. T. Sharpe, A. Stockman, H. Jagle, and J. Nathans. Opsin genes, cone photopigments, color vision and color blindness. In K. R. Gegenfurtner and L. T. Sharpe, editors, *Color Vision: From Genes to Perception*. Cambridge University Press, 1999.

[18] J.-E. Sneddon. Google's latest Chrome extension wants to ask you stuff. OMG!Chrome! `http://www.omgchrome.com/chrome-user-experience-surveys-extension/`, May 2015.

[19] J. Sobey, R. Biddle, P. C. van Oorschot, and A. S. Patrick. Exploring user reactions to browser cues for Extended Validation certificates. In *Proceedings of ESORICS*, 2008.

[20] A. Sotirakopoulos, K. Hawkey, and K. Beznosov. On the challenges in usable security lab studies: lessons learned from replicating a study on ssl warnings. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, page 3. ACM, 2011.

[21] J. Sunshine, S. Egelman, H. Almuhimedi, N. Atri, and L. F. Cranor. Crying wolf: An empirical study of ssl warning effectiveness. In *USENIX Security Symposium*, pages 399–416, 2009.

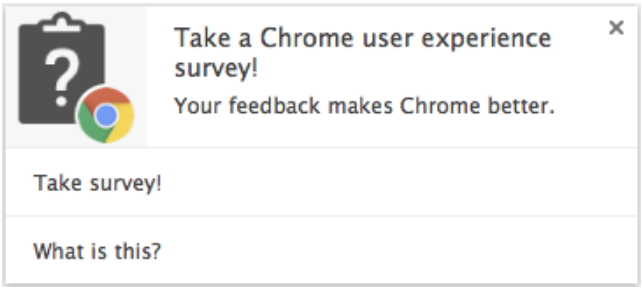
[22] R. Wash and E. Rader. Too much knowledge? Security beliefs and protective behaviors among United States Internet users. In *Proceedings of SOUPS*, 2015.

[23] T. Whalen and K. M. Inkpen. Gathering evidence: Use of visual security cues in web browsers. In *Proceedings of Graphics Interface*, 2005.

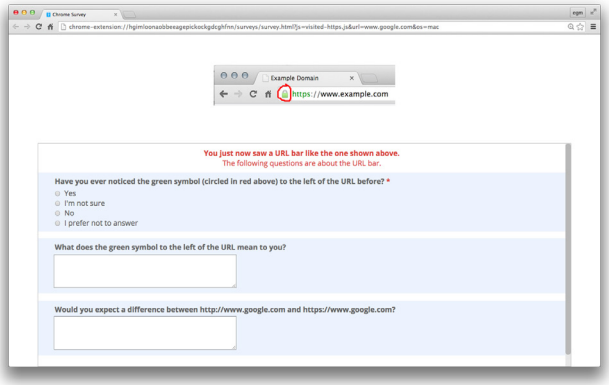
APPENDIX

A. EXTENSION SCREENSHOTS

When the survey criteria were met, the extension would generate a notification that looked like:



After clicking on the notification, the respondent would see a survey that looked like:



B. EXTENSION QUESTIONS

The full list of questions for the extension survey.

B.1 HTTP survey questions

1. Have you ever noticed the white symbol (circled in red above) to the left of the URL before?
   Yes
   I'm not sure
   No
   I prefer not to answer

2. What does the white symbol to the left of the URL mean to you? [Short answer]
3. Would you expect a difference between http://www.example.com and https://www.example.com? [Short answer]

B.2 HTTPS survey questions

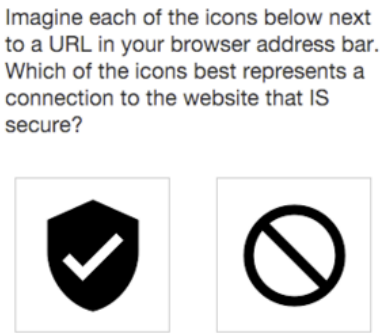
1. Have you ever noticed the green symbol (circled in red above) to the left of the URL before?
   Yes
   I'm not sure
   No
   I prefer not to answer
2. What does the green symbol to the left of the URL mean to you? [Short answer]
3. Would you expect a difference between http://www.example.com and https://www.example.com? [Short answer]

C. GCS SURVEY QUESTIONS

Examples of what the questions looked like to respondents.

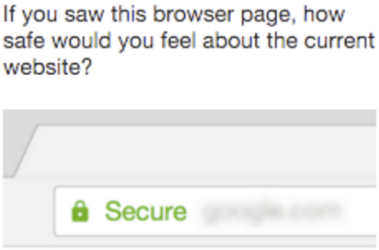
C.1 Icon questions

An example pairwise icon question:



C.2 Text questions

An example text question:



D. GCS SURVEY DEMOGRAPHICS

For completeness, we provide the *inferred* demographics of our survey respondents as provided by the GCS platform. We urge caution in interpreting inferred demographics. GCS assigns demographic characteristics to respondents based on their browsing history, which is an imperfect process [1].

D.1 Icon questions

	N	% of total
Male	496	49.6%
Female	349	34.9%
Unknown	155	15.5%
Age 18-24	131	13.1%
Age 25-34	178	17.8%
Age 35-44	157	15.7%
Age 45-54	132	13.2%
Age 55-64	109	10.9%
Age 65 or over	54	5.4%
Age Unknown	239	23.9%
Income $0-$24,999	80	8.0%
Income $25,000-$49,999	545	54.5%
Income $50,000-$74,999	250	25.0%
Income $75,000-$99,999	64	6.4%
Income $100,000-$149,999	24	2.4%
Income $150,000+	9	0.9%
Income Unknown	28	2.8%

D.2 Text questions

	N	% of total
Male	3006	46.5%
Female	2186	33.8%
Unknown	1270	19.7%
Age 18-24	918	14.2%
Age 25-34	1283	19.9%
Age 35-44	956	14.8%
Age 45-54	724	11.2%
Age 55-64	642	9.9%
Age 65+	312	4.8%
Unknown	1627	25.2%
Income $0-$24,999	576	8.9%
Income $25,000-$49,999	3421	52.9%
Income $50,000-$74,999	1624	25.1%
Income $75,000-$99,999	434	6.7%
Income $100,000-$149,999	151	2.3%
Income $150,000+	51	0.8%
Income Unknown	205	3.2%

A Week to Remember

The Impact of Browser Warning Storage Policies

Joel Weinberger
Google, Inc.
jww@chromium.org

Adrienne Porter Felt
Google, Inc.
felt@chromium.org

ABSTRACT

When someone decides to ignore an HTTPS error warning, how long should the browser remember that decision? If they return to the website in five minutes, an hour, a day, or a week, should the browser show them the warning again or respect their previous decision? There is no clear industry consensus, with eight major browsers exhibiting four different HTTPS error exception storage policies.

Ideally, a browser would not ask someone about the same warning over and over again. If a user believes the warning is a false alarm, repeated warnings undermine the browser's trustworthiness without providing a security benefit. However, some people might change their mind, and we do not want one security mistake to become permanent.

We evaluated six storage policies with a large-scale, multi-month field experiment. We found substantial differences between the policies and that one of the storage policies achieved more of our goals than the rest. Google Chrome 45 adopted our proposal, and it has proved successful since deployed. Subsequently, we ran Mechanical Turk and Google Consumer Surveys to learn about user expectations for warnings. Respondents generally lacked knowledge about Chrome's new storage policy, but we remain satisfied with our proposal due to the behavioral benefits we have observed in the field.

1. INTRODUCTION

An HTTPS error warning might be the last defense between an activist and an active network attacker. As a community, we need security warnings to be effective: clear, trustworthy, and convincing. Prior research has focused on warning comprehension, design, and performance in the field (e.g., [2, 9, 10, 17, 18, 19]). We look at a new angle: storage policies.

Network attackers and benign misconfigurations both cause HTTPS errors. Users may *perceive* warnings as false positives if they do not believe they are under attack. In such a situation, Alice can override the warning and proceed to the website. What happens the next time Alice visits the same website with the same warning? She won't see the warning again until her error exception expires, the length of which depends on her browser's *storage policy*. E.g., Edge saves exceptions until the browser restarts.

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22–24, 2016, Denver, Colorado, USA.

A browser's exception storage policy has profound usability and security effects. On one hand, consider a user under attack who overrides a warning because she incorrectly believes it to be a false positive. Saving the exception forever puts that user at risk of a persistent compromise. On the other hand, consider a user who repeatedly visits a website with an expired but otherwise valid certificate. Showing the second person a warning every time they visit the site would undermine the warning's trustworthiness without providing much security benefit. Over time, that user will become jaded and might override a real warning.

We are unaware of any prior research into the effects of exception storage policies. As a result, browser engineers have selected storage policies without knowing the full trade-offs. In the case of Google Chrome, the original storage policy was chosen entirely for ease of implementation. Our goal is to provide user research and a security analysis to inform browsers' storage policies in the future.

In this paper, we evaluate various storage policies. Ideally, an optimal policy should maximize warning adherence and minimize the potential harm that could come from mistakenly overriding a warning. We ran a multi-month field experiment with 1,614,542 Google Chrome warning impressions, followed by Mechanical Turk and Google Consumer Surveys (GCS). Based on our findings, we propose a new storage policy that has been adopted by Google Chrome 45.

Contributions. We make the following contributions:

- To our knowledge, we are the first to study warning storage policies. We define the problem space by identifying goals, constraints, existing browser behavior, and metrics for evaluating storage policies.
- Using a large-scale, multi-month field experiment, we demonstrate that storage policies substantially affect warning adherence rates. Depending on the policy, adherence rates ranged from 56% to 70%.
- We propose a new storage policy, grounded in our experimental results and a security analysis of available policies. We implemented and deployed this strategy as part of Google Chrome 45.
- We ran surveys about storage policies. Respondents generally did not have strong beliefs or accurate intuitions about storage policies, suggesting that changing a browser's policy would not negatively surprise users.

- We show that researchers need to account for storage policies when comparing adherence rates across browsers. We find that Chrome's adherence rate could be significantly higher or lower than Firefox's depending solely on the selected storage policy.

2. BACKGROUND

We explain the role of HTTPS errors and why the false alarm effect is a concern for HTTPS warnings.

2.1 Purpose of HTTPS errors

HTTPS ensures that web content is private and unalterable in transit, even if a man-in-the-middle (MITM) attacker intercepts the connection. In order to do this, the browser verifies the server's identity by validating its public-key certificate chain. Browsers show security warnings if the certificate chain fails to validate.

Threat model. MITM attackers range in skill level and intent. An attacker could be a petty thief taking advantage of an open WiFi hotspot, or it might be a wireless provider trying (fairly benignly) to modify content for traffic shaping [15]. On the more serious end, governments are known to utilize MITM attacks for censorship, tracking, or other purposes [7, 12, 13]. The attacker might be *persistent*, meaning the target user is continuously subject to attack over a long period of time. Governments and ISPs are examples of entities that have the technical means for persistent attacks.

False positives. Many HTTPS errors are caused by benign misconfigurations of the client, server, or network [1]. When people encounter these situations, they want to ignore the error. Although the actual attack rate is unknown, we believe that false positives are much more common than actual attacks. Unfortunately, false positives and actual attacks seem very similar to non-expert end users.

Warnings. If there is an HTTPS error, the browser will stop the page load and display an HTTPS error warning (for examples, see Figure 1). Typically, users are able to override the warning by clicking on a button, although this may be disabled if the website serves the HTTP Strict Transport Security (HSTS) or HTTP Public Key Pinning (HPKP) headers [1]. If the error is caused by an actual attack, overriding the warning allows the attack to proceed.

Storage policy. Once a user has overridden an HTTPS warning on a website, the browser must persist the user's exception for some amount of time. The browser's *storage policy* determines how long the exception is saved for.

[1]The HSTS header specifies that the host is only loaded over valid HTTPS. HPKP allows the server to specify a set of public keys of which at least one is required to be in the certificate chain on any future loads of the server in the browser. User agents generally assume that the presence of these headers imply stronger requirements by the server about the importance of valid HTTPS, and thus HTTPS warnings on such sites are generally made non-overridable.

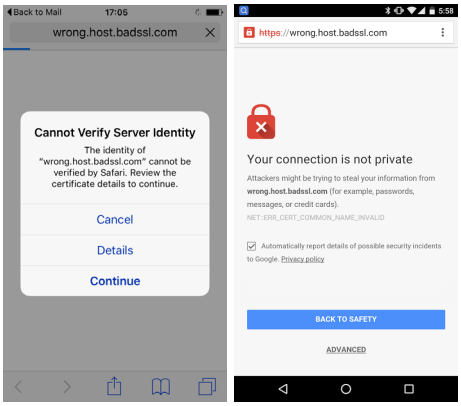


Figure 1: HTTPS error warnings in Safari for iOS (left) and Chrome for Android (right).

2.2 The false alarm effect

"Each false alarm reduces the credibility of a warning system," cautioned Shlomo Breznitz in 1984 [6]. "The credibility loss following a false alarm episode has serious ramifications to behavior in a variety of response channels. Thus, future similar alerts may receive less attention... they may reduce their willingness to engage in protective behavior." Breznitz was describing the *false alarm effect*, a theory that humans heed warnings less after false alarms. The false alarm effect is long known to decrease attention and adherence to non-computer warnings (e.g., [14, 20]).

Many prior researchers have observed evidence of the false alarm effect for computer security warnings. Nearly all of these researchers have urged industry vendors to decrease their false positive rates to mitigate the effect. Unfortunately, HTTPS errors are still commonly false alarms [1].

In one study of simulated spear phishing, researchers observed a correlation between recognizing a warning and ignoring it [8]. For example, one of their participants said the phishing warning that would have protected him/her "looked like warnings I see at work which I know to ignore" [8]. In a similar study of PDF download warnings, "55 of our 120 participants mentioned desensitisation to warnings as a reason for disregarding them" [16]. Bravo-Lillo et al. found, in two related studies, that participants quickly learned to ignore spurious security dialogs [5, 4].

The false alarm effect is a psychological process that can happen quickly. Anderson et al. watched participants view repeated security dialogs in an fMRI machine [3]. Their participants did less visual processing of the dialogs after only one exposure, with a large drop after thirteen exposures.

Outside of the lab, researchers have seen evidence of the false alarm effect in Chrome users in the field. Chrome users clicked through 50% of SSL warnings in 1.7 seconds or less, which "is consistent with the theory of warning fatigue" [2].

3. PROPOSAL

We argue that a browser's exception storage policy should be chosen with care (rather than for ease of implementation) because it affects end user security and warning effectiveness. We propose a new policy based on desired usability and security properties of HTTPS warnings.

3.1 Goals

Our goals for a storage policy are:

- Reduce the false alarm effect by avoiding unnecessary warnings. The longer the storage policy, the less likely it is that a user will see a repeat warning that they consider a false alarm. In the long run, this should yield increased attention to actual attacks.
- Reduce the cost of a mistake if someone misidentifies an actual attack as a false alarm. If a user fails to heed a warning during an actual attack, we do not want that user to be permanently compromised. The shorter the storage policy, the less time that an attacker has to intercept the client's connection.
- Avoid unpleasantly surprising users.

A keen reader may notice that the first two goals are diametrically opposed. To reduce the false alarm effect, we should *increase* how long exceptions are stored; to reduce the cost of a mistake, we should *decrease* how long exceptions are stored. In Section 4, we perform a study to find a policy that satisfies both constraints as much as possible, while acknowledging that neither can be completely satisfied.

3.2 Analysis of existing options

There is no industry consensus for how long HTTPS error exceptions should be stored, and existing browser exception storage policies do not meet our goals. We tested browser storage policies as of February 2016 (Figure 2). With the notable exception of Firefox, browser vendors appear to have selected their storage policies based on ease of implementation, which sometimes results in the same browser having inconsistent policies across platforms.

Browser session. The most common storage policy is to save exceptions until the browser restarts, either by closing the browser or closing all window instances of the current profile. A session-based policy yields unpredictable but typically short storage lengths. Although a browsing session can last anywhere from five minutes to a month, we know that the average Chrome browsing session lasts slightly less than a day. False alarms could therefore still be daily occurrences for people who need to interact with misconfigured websites.

Browser	OS	Storage policy
Chrome 44	Windows	Browser session
Safari 9	Mac	Browser session
UC Browser 10	Android	Browser session
Edge 20	Windows	Browser session
Firefox 44	Windows	User choice (browser session or permanent)
Safari 9	iOS	Permanent
UC Browser 2	iOS	Permanent
UC Mini 10	Android	Overriding not allowed

Figure 2: Browser exception storage policies. This covers Google Chrome, Apple Safari, Microsoft Edge, and Mozilla Firefox, as well as UC Web's three browsers, which are popular in South and East Asia.

From a technical perspective, this is the simplest policy to implement: the user's decision is saved as an in-memory map of hostnames to exception state. For Chrome, engineers chose this policy in large part because it was very easy to implement and the trade-offs associated with the storage policy were unknown; we guess the same decision making process might have been used by other browsers.

Permanently. The next most common storage policy is to always save exceptions permanently. This policy is also easy to implement in browsers that store other per-website preferences permanently in a preferences file. Permanently storing exceptions reduces the false alarm effect, but the cost of a mistake is also permanent.[2]

A choice. Firefox is the only browser to explicitly give users a choice between two storage policies. By default, an exception is stored permanently. However, the user has the option to store the exception only until browser restart. Firefox users choose the shorter option 21% of the time [2]. Although we like the idea of a choice, browser vendors still need to decide what the options and default are.

Not applicable. UC Mini for Android doesn't let users override HTTPS errors, so there is no need for storage. This prevents people from accessing misconfigured websites at all.

3.3 Our proposal

In contrast to the above existing options, we propose a new, time-based storage policy:

- Store exceptions for a fixed amount of time that is not forever. The amount of time should empirically minimize the cost of a mistake and false alarm frequency.
- Delete stored exceptions when we think users will expect it, for example when clearing browser history or closing a private browsing session.
- If a user ever encounters a valid certificate chain for a website, forget any previously stored exceptions for that website. This can occur when someone proceeds through a warning in the presence of a transient attacker and then later reconnects from a safe network. Forgetting the exception in this situation should reduce harm without increasing the false alarm rate.

In Section 4, we test this policy with several different configurations and compare it to other, existing strategies.

[2]Browsers that provide a "permanent" storage strategy do generally provide a way to remove an exception once granted, but the difficulty in undoing this decision depends on the browser. In Firefox 44, for example, it requires going to a special "Certificates" menu several levels into Preferences under "Advanced" settings. Then one must manually curate a list of server certificates to find the one for which there an exception was earlier granted, and then the user must explicitly choose to delete it.

4. FIELD EXPERIMENT

We ran a large-scale field experiment to determine whether the time-based storage policy has merit and, if so, the ideal length of time for a time-based storage policy.

4.1 Measurement

We want to know whether there is a length of time that minimizes both the false alarm effect and cost of a mistake. We cannot measure either property directly because we do not know which HTTPS errors are false alarms or mistakes. However, we can use the warning adherence rate and regret rate as proxies of our desired properties.

Adherence rate. Chrome already uses telemetry to record important warning metrics in aggregate, including adherence. *Adherence* is the rate at which people heed the warning's advice to not proceed to the page. We desire high adherence rates. A low adherence rate is a sign that users are experiencing warnings that they consider false alarms.

Regret rate. How often do users change their mind about whether it's safe to override a warning? If someone repeatedly overrides the same warning, then we should stop showing them that warning. On the other hand, consider someone who overrides a warning on Tuesday but then adheres to that warning on Thursday. We view this as an indication of regret — that the user's original decision to override the warning was a mistake. We don't want to store mistakes for any longer than necessary. If a long time period has a high regret rate, it is inferior because it perhaps is preventing users from changing their minds sooner. We acknowledge that our regret rate is an imperfect metric because it does not actually measure users' *feelings*. However, it is meaningful when applied as a comparison tool across experimental conditions because it allows us to see changes in behavior.

Thus, we deem a storage policy as superior if it has a high adherence rate and low regret rate. A strictly superior strategy would be one that did not change the regret rate at all, but increased the adherence rate.

4.2 Experiment structure

Groups. We tested six policies: one session-based policy, three short time periods (one day, three days, one week), and two long time periods (one month and three months). In the first round of our experiment, we tested only the session-based and short time policies. After that round was successful, we added the long time periods. Our groups and metrics only apply to overridable HTTPS error warnings. Errors that cannot be overridden (due to HSTS or HPKP) are excluded from our experiment.

Assignment. We set the number of Chrome users in each experimental group to the same small percentage. The experiment was done across all Chrome platforms[3]. Clients were randomly assigned into experimental groups, and their pseudonymous telemetry data was tagged with the group name. Telemetry data was collected only from Chrome users who opted in to Chrome user metrics.

Length. Regret rates cannot be measured until exceptions begin to expire, which happens at the time determined by the policy length. So to collect useful data, we let the experiment run for three months on Chrome's stable release channel, but discarded the data. This warm-up gave the longest strategy time for decisions to expire initially so we could measure changes in user behavior and regret rates. At this point, we collected warning impressions for 28 days.

[3]Windows, Mac, Linux, ChromeOS, Android, and iOS.

```
host string : {
    fingerprint string
    decision_expiration_time uint64
    guid string
}
```

Figure 3: Decision memory structure

4.3 Implementation

We describe how we implemented the storage policies.

4.3.1 Session storage policy

Chrome's original implementation is an in-memory map from hostname to a map of certificate and policy decision. The decision is an **enum** of 3 possible values: **ALLOWED**, **DENIED**, **UNKNOWN**. They respectively represent a certificate error that was allowed by the user, one that was denied, or one in an unknown state. In practice, the saved state is either **ALLOWED** or **DENIED**.

If a certificate error is encountered, the networking stack asks the warning manager for the user's preference. If the user has already allowed the error for this particular host, the warning manager tells the networking stack to allow the connection to continue. Otherwise, a warning is shown. If the user overrides the warning, the warning manager will add the decision to the map.

All profiles receive their own map, so decisions do not carry between profiles. Since this map is in-memory, it is reset when Chrome completely shuts down. On restart, users will be re-asked for any decisions they previously granted.

4.3.2 Time-based storage policies

With a time-based storage policy, exceptions need to persist through browser restarts. This means that exceptions need to be saved on disk. We used an internal Chrome API named "Content Settings," which stores persistent preferences on a per-profile basis. For example, user preferences about geolocation use and plug-ins are stored in Content Settings. We consider an error exception for a website to be a type of website preference.

Content Settings are stored and retrieved by hostname. We thus map a given hostname to a set of structures containing metadata about individual exceptions granted by the user. Figure 3 shows the Content Setting structure for storing certificate exceptions. It contains:

- `host` is the hostname where the exception was made.
- `fingerprint` is a SHA-256 hash of the certificate and full certificate chain that contained the error. We save the hash to individually identify each error.
- `decision_expiration_time` is the Epoch Time in seconds of when the decision expires. If the certificate in question is checked again, we check the expiration time to see whether the previous exception is still valid.
- `guid` is a globally unique identifier set at browser session start. We use this to address a complexity that arose from sharing code between the time-based policies and the session-based policy. The Content Settings API doesn't allow the caller to know whether a setting was made this session or a previous session. Since Chrome may not cleanly shutdown (for example, if it is force killed or if the machine resets), the per-session exceptions cannot be reliably cleaned up when the session is over. It is tempting to clean up the settings on browser start, but this is problematic for measuring the regret rates: if the settings are cleaned up, and then the same certificate is received, there is no way to know if the user previously made an exception. The solution is to create a per-session globally unique identifier (GUID) which is stored with all the exceptions stored in Content Settings. Then, if the browser session restarts for any reasons, all of the old exceptions are still stored, but they will reflect an old GUID, so it is known that they are "expired" (i.e. they were created in a previous session).

4.3.3 History

When stored on disk, exceptions contain hostnames and certificate fingerprints. They are potentially privacy-sensitive since they reveal information about the user's browsing habits. Our implementation therefore needs to take care with how exceptions are handled.

Profiles are the mechanism for storing settings, state, and history of the current user. However, Incognito profiles, also known in other browsers as "Private Browsing," do not record state about the user past the current session. In general, Chrome does a best effort to not store user history in permanent storage. Thus, Chrome does not save any Content Settings for Incognito profiles to disk.

Additionally, there is a general expectation that history-resetting activities should delete site visiting activities. Since this is an indirect type of history recording, it is necessary to make sure certificate exception Content Settings are erased when history is cleaned up. Chrome internally provides a `BrowsingDataRemover` API which is called whenever browsing data or history is cleaned up. The certificate exceptions Content Settings are cleared whenever this API is invoked.

```
EXPIRED_AND_PROCEED
EXPIRED_AND_DO_NOT_PROCEED
NOT_EXPIRED_AND_PROCEED
NOT_EXPIRED_AND_DO_NOT_PROCEED
```

Figure 4: Events when a certificate exception is encountered, used to calculate the regret rate.

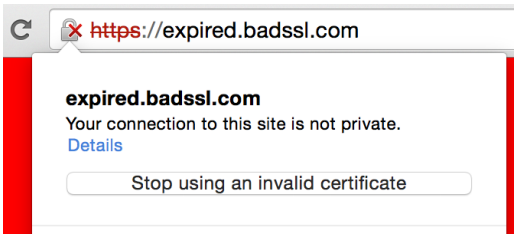


Figure 5: The button we added to let experiment participants revoke exceptions.

4.3.4 Analytics

To do the study, we must measure the adherence and regret rates. The adherence rate is already recorded by Chrome, so we had to add the regret rate. This is calculated by recording (a) the decision made when a certificate error is encountered, and (b) the prior state of that certificate exception. Figure 4 shows the recorded events. The `EXPIRED_*` events indicate that the identical certificate error had been encountered in the past, while the `NOT_EXPIRED_*` events indicate the opposite. The `*_PROCEED` events indicate an ultimate decision to create an exception for the error, while the `*_DO_NOT_PROCEED` indicate the opposite. These measurements are all taken in relation to the user's interaction with the HTTPS warning page.

4.4 Ethics

Running a security field experiment inherently has risks, as do all security engineering changes. In this case, the primary risk was that adherence or regret rates could suffer in undesirable ways. A secondary risk was that saving a preference to a local file might have an impact on user expectations of local privacy. However, all of our experimental treatments fell within the bounds of other browsers' behavior (since other major browsers have both very short and very long storage policies). We believed the small risk was worth the potential benefit of a new, improved policy.

Still, we were cautious. We took steps to limit any potential harm that could come from the experiment:

- We monitored key statistics as we ran the experiment. For example, we monitored the average number of warning impressions to watch for any sudden large increases, which could be an indication of accidentally desensitizing users to Chrome's warnings. We also observed how often users utilized the "Revoke" button in the page info bubble to make sure users were not explicitly changing their minds often. We could have immediately stopped the experiment via server-side controls if we had believed it necessary.
- We slowly rolled out the experiment to progressively larger groups of users, beginning with pre-release ver-

sions of Chrome. Pre-release Chrome users are developers, power users, and other people willing to trade inconvenience for cutting edge features. When we progressed to stable, we slowly ramped up group sizes.

- We started with three short time periods that were similar to the average Chrome user session: one day, three days, and one week. Initially, we did not do long groups in case the regret rates were too high. Once we saw that regret rate changes were small in the first three groups, we added the two longer groups.
- Previously, users could force Chrome to revoke an exception by restarting. We didn't want to take away this control, so we added a button to the page info bubble (Figure 5) to let users revoke an exception. Additionally, it resets all socket connections for the current browser session to make sure that any exceptions already granted in the networking layer are reset.
- Our implementation provides an additional local history entry for websites with exceptions, but it is similar to regular history. According to our proposal, clearing history now also deletes error exceptions.
- Incognito mode should not persist anything new to disk, so exceptions made in Incognito mode are forgotten once the last Incognito tab is closed.

We did not debrief study participants. Given the low level of risk and our cautious experimental rollout, we did not feel that debriefing was necessary. Furthermore, debriefing notices are infeasible for small, low-risk field experiments. We run many in-product experiments in the course of improving and rolling out new security features, so debriefing notices would be frequent and tiresome. Instead, we prefer to design our experiments to be low-risk. If we had felt that the potential for harm was great enough to merit debriefing, we would have run a lab study instead of a field experiment.

Our experiment was internally reviewed prior to launch in a process that included security experts, a privacy expert, and an experimental research expert.

4.5 Limitations

We believe that our data is representative and well-defined. However, there are limitations and potential sources of bias.

Sample bias. Since our metrics were collected via Chrome's user metrics analytics (UMA) opt-in program, our data is biased towards users who have chosen to have events anonymously collected. This could mean, for example, that there is a bias towards users who are more or less privacy sensitive, affecting the rate that they adhere to warnings as compared to the general population. However, a large fraction of the population opts in, and we examined millions of warning impressions during the full course of the experiment.

Metrics. We are using adherence and regret rates as proxies for actual human desires and intent. It is impossible for our large-scale metrics to precisely capture actual human meaning. For example, imagine that Alice views a warning, gets distracted by her dog, and then overrides a new impression of the same warning once she returns to her task. The initial adherence does not mean the warning worked. In the reverse, imagine that Alice overrode a warning on Tuesday but got distracted when she saw it again on Thursday. She is not actually expressing regret for her Tuesday action. This same limitation holds true across all of our conditions, and we expect the same amount of noise for all conditions.

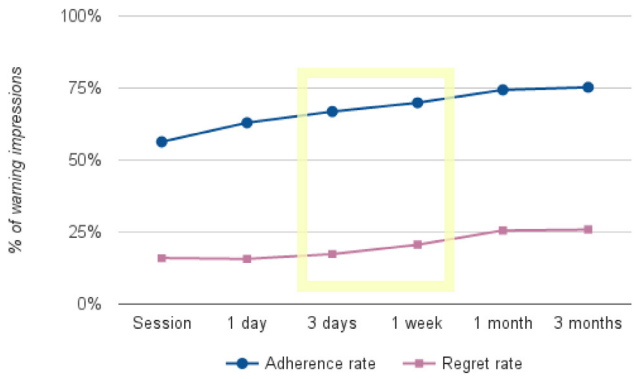


Figure 6: Results of different storage policies.

Continuous measurement. Once we choose a strategy for deployment, we can keep our metrics in place to continuous measure adherence and regret to look for unexpected changes. However, we cannot continuously run a full experiment for *all groups* to know if our initial experimental results permanently hold. Because real users are affected, we must choose a system that we think is safest and most usable for our users. Unfortunately, this means that while we can see if our initial results remain for our chosen strategy, we cannot know if they would remain for the other groups.

Over-representation. We do not differentiate between users who see many warnings and users who see few warnings. Our statistics are averaged across all users within a treatment group, so users who see many warning impressions will be over-represented when averaging across impressions. Given the scale of our experiment, we do not expect a confound because different types of users should be evenly distributed across experimental groups.

4.6 Results

Table 1 shows the impact of different storage policies on 1,614,542 warning impressions. Our experimental data substantiates two hypotheses:

- Storage policies matter. We see large differences in adherence and regret rates across policies.
- Storage length correlates with both adherence rates and regret rates.

We see the biggest difference by comparing the two extremes. Participants in the three-month group saw an increase in adherence from 56% to 75%, as compared to the session-based policy. At the same time, the three-month group's regret rate increased from 16% to 26%.

While we were pleased to see the adherence rate increase with the longer storage policies, we recognize the cost. The longer Chrome stores exceptions, the more likely the user is to reverse their decision once the exception expires.

	Session (baseline)	One Day	Three Days	One Week	One Month	Three Months
Adherence rate	56.35%	62.96%	66.82%	69.88%	74.38%	75.28%
Regret rate	15.98%	15.67%	17.35%	20.59%	25.56%	25.86%
Difference in regret from baseline	-	-0.31	1.37	4.61	9.58	9.88

Table 1: Results of different exception storage policies. The difference in regret from baseline is simply the baseline's regret rate subtracted from the policy's regret rate.

4.7 Choosing a new policy

Following the experiment, we needed to select a new policy for Chrome. Figure 6 highlights the most promising candidates. We ultimately chose the one-week policy.

Our main aim is to raise the adherence rate for Chrome's HTTPS error warnings. With only this constraint in mind, we would select the three-month policy. The results show, as expected, that the longer the policy, the greater the adherence. However, the increase in adherence also brings an increase in regret rate: the three-month policy yielded a 9.88 point increase in the regret rate (Table 1). We are not willing to accept such a large increase to the regret rate.

To strike a balance between the two conflicting constraints, we decided that we would accept up to a 5-point increase from the baseline's regret rate. Of the policies that meet this requirement, the one-week policy has the greatest adherence gains. Both the one-month and three-month policies have much larger regret rate increases, while the one-day and three-day policies have lower adherence rate gains.

We do not assert that this is the objectively best choice for a storage policy. All of the policy choices require a trade-off, and different companies may weigh adherence and regret rates differently. For example, someone who is willing to tolerate a higher regret rate would likely choose the three-month policy. Going forward, as we monitor Chrome's metrics, we plan to re-evaluate this trade-off.

4.8 Deployment

We launched the one-week policy as part of Google Chrome 45, in September 2015. Post-launch, the policy is working well for the general population. Looking at 9,318,975 warning impressions, we see an adherence rate of 71.79% and a regret rate of 18.20%. To our pleasant surprise, the policy yielded a slightly higher adherence rate and slightly lower regret rate for the general population as compared to the one-week experimental group.

5. USER EXPECTATIONS

One of our initial goals was to avoid unpleasant surprises, which requires understanding user expectations. Several months after Chrome adopted our week-long storage policy in Chrome, we collected user feedback to either confirm or question our decision. Do users *have* expectations? Does our newly adopted proposal meet those expectations?

5.1 Method

We surveyed 1,327 people about Chrome's exception storage policy. First, we asked 100 Mechanical Turk workers to tell us about the storage policy in their own words. Based on those responses, we designed multiple-choice questions and gathered 1,227 Google Consumer Survey responses.

5.1.1 Mechanical Turk

Questions. The survey contained three questions, which intentionally did not mention security:

1. *Which Internet browsers do you use at least once a week?*
2. *Imagine that you saw this error page while trying to open a website in Chrome. [Image.] If you clicked 'Proceed' on the error page, how long would Chrome remember your decision for?*
3. *Have you ever seen this error page before, in Chrome?*

A screenshot of the questions is available in Appendix A.1.

Screening. We limited the survey to Mechanical Turk workers in the US, and we screened for Chrome usage. The survey was advertised as "Chrome users - Survey about error pages, takes about 4 minutes," with the goal of attracting survey respondents who use Chrome. We ran the survey until we collected 100 responses from people who said they use Chrome at least weekly according to the first question. We paid other respondents but discarded their responses. We did not receive any nonsense or garbage responses.

Coding. One researcher coded the short answer responses. The researcher did one round of open coding, developed a codebook, and then applied a fixed codebook to the responses. Since the responses are short and straightforward, we did not have a second researcher duplicate the codes.

Payment. We paid respondents $0.80 to complete a survey that took between one and four minutes. This amount was chosen to reflect a minimum hourly wage of $12.

5.1.2 Google Consumer Surveys

Questions. We ran two questions as separate surveys, each accompanied by an image of an HTTPS warning:

- *If you clicked 'Proceed' on this error page, how long would Chrome remember that decision for?* (Response options: Once, while I'm using the website; A week; Until I clear my history; Forever; I don't know)
- *If you clicked 'Proceed' on this error page, how long would you WANT Chrome to remember that decision for?* (Response options: Once, while I'm using the website; A few hours or days; Until I clear my history; Forever)

Response options were randomly reversed, with "I don't know" pinned as the bottom answer for the first question. Screenshots of the questions are available in Appendix A.2.

Session	58%
Period of time	19%
Browser cleared	5%
Forever	13%
Don't know	5%

Table 2: *How long would Chrome remember your decision for?* Mechanical Turk short answers.

	AU	US
Once, while I'm using the website	20%	9%
A few hours or days	6%	4%
Until I clear my history	16%	17%
Forever	9%	10%
I don't know	49%	60%

Table 3: *How long would Chrome remember your decision for?* GCS multiple choice responses.

Screening. We requested 600 responses for each question, split evenly between Australian and American respondents. We received 300 for each category except for the first one from Australia, where we received 327 responses.

Payment. Respondents were not directly paid. Google Consumer Surveys on desktop are displayed on websites in lieu of paywalls. Respondents received free access to website content after completing the survey.

5.2 Ethics

We did not ask for any personally identifiable or sensitive information. Participants were compensated for their time in a way suitable for each survey platform.

5.3 Results

We conclude that respondents do not have strongly held beliefs about Chrome's exception storage policy, and preferences are split between session-based and longer policies.

5.3.1 Beliefs about current behavior

Categories. The short answer responses fell into five categories, which we used for the multiple choice questions:

- *Session.* The response specified a period of time that's similar to a session-based policy. This includes saving it once, for a very short period of time, until restarting, or until closing the window.
- *Period of time.* The response talked about a period of time that lasts longer than a typical browsing session on a website. For example, "a week" or "30 days."
- *Browser cleared.* The respondent mentioned "clearing" something (for example, "until your browsing history is cleared," or "until you clear your cache").
- *Forever.* A synonym of "forever," like "always."
- *Don't know.* The respondent couldn't answer the question. For example, "not sure," or "I don't know."

Correctness. Few respondents correctly identified Chrome's current storage policy, even though it had been in place for several months. We find that most respondents lack preconceived beliefs about Chrome's storage policy, although they can make reasonable guesses when incentivized.

	AU	US
Once, while I'm using the website	58%	51%
A few hours or days	12%	13%
Until I clear my history	22%	18%
Forever	8%	18%

Table 4: *How long would you want Chrome to remember your decision for?* GCS responses.

A majority of the Mechanical Turk respondents incorrectly said the storage policy is session-based (Table 2). Although incorrect, this is a reasonable guess; Chrome exhibited this behavior until several months prior to the survey, and other browsers have session-based storage policies. 95% of the Mechanical Turk responses matched feasible potential policies. We therefore conclude that non-expert browser users are capable of reasoning about exception storage policies — when paid to pay attention to a survey.

In reality, browser users are not paid to pay attention to our question. Warnings interrupt people who are trying to complete another task. As a result, their attention is split between the warning and the other task. Consumer Surveys are similar because they interrupt respondents en route to a desired website. In this context, people struggled to answer the question. Approximately half of GCS respondents said they didn't know the answer, and the response rate was low (2.4% in Australia, 6.8% in the United States). This suggests to us that respondents found this question too difficult to answer quickly, meaning they have no strongly held, preconceived belief about current exception storage policies.

Defining a session. Mechanical Turk respondents had varying definitions of a "session." We assigned one or more secondary codes to the session-related responses, depending on the type of session the respondent described. Of the 58 session-related responses:

- 26 referred to storing the exception once ("1 time")
- 14 explicitly used the word "session" ("that session only")
- 10 talked about the lifetime of a tab ("till I close the window")
- 6 listed very short time periods
- 5 mentioned restarting ("until I restarted the browser")

All of these responses relate to the lifetime of a browsing session, but they each have different properties in practice. For example, tab lifetimes are generally much shorter than the time between browser restarts.

5.3.2 Policy preferences

We asked GCS respondents to choose their preferred storage policy, and their answers were split (Table 4). Half of respondents preferred a session-based storage policy, but the other half expressed a preference for the current time-based

strategy or longer. This leaves us with no clear consensus, although favoring the previous session-based policy.

Notably, a fifth of respondents expected clearing their history to revoke an exception. Chrome's previous strategy did not do this, nor do other browsers. We are glad we added it because it appears to be a common expectation.

6. IMPLICATIONS

We discuss the main lessons learned from our experiment and surveys, and give suggestions for future work.

6.1 Storage policies matter

Changing a warning's storage policy has almost as large an effect on adherence as completely changing the warning's UI. In our experiment, we saw a 19 percentage point difference between different storage policies (56% to 75%), which is huge! For comparison, Chrome researchers raised adherence by 25 percentage points with a full text and design overhaul [10]. Our work demonstrates that exception storage policies are an important part of warning interaction design, and we see enormous potential in this line of research.

We hope to motivate further research into storage policies. Historically, research into warning effectiveness has primarily focused on the warning's content or effectiveness [10, 17, 19, 18]. Storage policies have received little attention aside from brief mentions in two of our recent projects [2, 11]. Other warnings' storage policies might also benefit from changes, or there might be more clever storage policies that outperform the ones we tested.

6.2 Our proposed policy works

Our proposed storage policy reduced the number of likely-unnecessary warnings. Warnings should be meaningful, justified, and rare. If the browser knows with a good degree of certainty that a user will not adhere to a warning, and there is a reasonable chance that the user's decision is correct, then the browser should not show the warning. When we reduced the number of unnecessary warnings, the overall adherence rate improved significantly with little cost to the regret rate.

We believe in removing unnecessary warnings because it increases the salience and trustworthiness of the warnings that remain. Over time, we hope that showing fewer unnecessary warnings will mitigate the false alarm effect and increase confidence in HTTPS error warnings.

We encourage other browser vendors to experiment with and adopt similar policies for storing (and forgetting) certificate error decisions. We would be interested to learn whether other browsers find similar benefits and side effects.

6.3 Warning adherence across browsers

Researchers need to account for storage policies when testing browser UI or otherwise comparing adherence rates. It is tempting to attribute differences in adherence to obvious differences in UI across browsers or experimental treatments. However, our findings demonstrate that one must first control for differences in storage policies.

Consider our efforts to improve Chrome's HTTPS error warning. In 2013, we learned that Firefox users were twice as likely to adhere to warnings as Chrome users (66% vs 30%) [1]. We initially attributed this to the obvious differences in UI between the browsers and thus began experimenting with design changes. Was Firefox's text easier to understand? Did the background color matter?

We were partly right: the design did matter. However, it was not the only factor. Chrome's HTTPS warning adherence rate remained lower than Firefox's even after a full redesign [10]. In fact, Chrome's adherence rate remained lower even when we tried using Firefox's exact warning UI in Chrome [11]. At the time, we hypothesized that this surprising finding might be due to demographics or storage policies [2, 11]. Our findings now support the storage policy hypothesis; Firefox's longer storage policy should give it a higher adherence rate. As Table 1 shows, Chrome's adherence rate could be higher or lower than Firefox's depending on our choice of storage policy.

Our findings also have two implications for laboratory studies. First, the effect of storage policies makes it difficult to compare adherence rates in the field to rates in a laboratory setting. A controlled laboratory study will include a fixed number of repeat warning exposures over a short period of time, whereas field data might include an unknown number of repeat warning exposures over a long period of time. This is not an apples-to-apples comparison. Second, we also recommend that researchers control for the number of repeat exposures when comparing experimental treatment groups, either across experiments or within the same experiment.

6.4 Storage policies are confusing

Our survey respondents were not familiar with Chrome's storage policy. Few of the survey respondents correctly identified Chrome's current storage policy, and most Google Consumer Survey respondents couldn't guess at all.

We find this confusion unsurprising. The eight browsers that we examined (Section 3.2) have four different storage policies, and we added a fifth policy. Browsers made by the same company have different policies across platforms, so people who use multiple devices will see different behaviors over time. Furthermore, storage policies are not well documented. Firefox is the only browser to mention the storage policy in the warning UI, and the policies are not mentioned on most browsers' help pages. Given this, how would people learn about storage policies?

We do not conclude that browser vendors should immediately embark on an education campaign. End users are not responsible for learning all of the technical details of their browsers. Instead, we think that browsers should act in the user's best interest and try to meet user expectations as much as possible without explicitly teaching people about storage policies. However, future work could explore whether people's behavior changes once they learn about different storage policies. (For example, people might be more cautious if they learn that exceptions last forever.) If that were the case, then comprehension of storage policies might be important enough to merit UI changes.

7. CONCLUDING SUMMARY

How long should a browser store a user's decision to override an HTTPS error warning? There is no clear industry consensus — eight major browsers have four different policies — and little research exists to guide the choice. We defined the usability and security requirements of an ideal policy,

and then we proposed a policy that meets our constraints.

We performed a large-scale field experiment to test different storage policies. After comparing adherence and regret rates between experimental groups, we concluded that error exceptions should be forgotten after one week. A one-week storage policy raised the adherence rate from 56% to 70% with little cost to the regret rate. Google Chrome 45 adopted our proposal, which brought the overall adherence rate to 72% as of February 2016.

To learn more about user beliefs and preferences, we ran Mechanical Turk and GCS surveys that asked about Chrome's storage policy. Most respondents did not know Chrome's current storage policy, and preferences were split between Chrome's old policy and our proposal. We remain satisfied with our proposal because respondents did not appear to have strong enough opinions to negate the clear benefit that we observed in our field experiment.

We encourage future work into the usability and security of different error storage policies. Would other browsers benefit from our policy? Are there changes to our policy that would improve it? How important is comprehension?

8. ACKNOWLEDGMENTS

We thank Ryan Sleevi and Chris Palmer for their expert input into the security trade-offs and experimental design. We also thank SOUPS reviewers for their suggestions.

9. REFERENCES

[1] D. Akhawe, B. Amann, M. Vallentin, and R. Sommer. Here's my cert, so trust me, maybe? Understanding TLS errors on the Web. In *World Wide Web Conference (WWW)*, 2013.

[2] D. Akhawe and A. P. Felt. Alice in warningland: A large-scale field study of browser security warning effectiveness. In *Proceedings of Usenix Security*, 2013.

[3] B. B. Anderson, C. B. Kirwan, J. L. Jenkins, D. Eargle, S. Howard, and A. Vance. How polymorphic warnings reduce habituation in the brain: Insights from an fMRI study. In *Proceedings of CHI*, 2015.

[4] C. Bravo-Lillo, L. F. Cranor, S. Komanduri, S. Schechter, and M. Sleeper. Harder to ignore? Revisiting pop-up fatigue and approaches to prevent it. In *Proceedings of SOUPS*, 2014.

[5] C. Bravo-Lillo, S. Komanduri, L. F. Cranor, R. W. Reeder, M. Sleeper, J. Downs, and S. Schechter. Your attention please: Designing security-decision UIs to make genuine risks harder to ignore. In *Proceedings of SOUPS*, 2013.

[6] S. Breznitz and C. Wolf. *The psychology of false alarms*. Lawrence Erbaum Associates, NJ, 1984.

[7] P. Eckersley. A Syrian man-in-the-middle attack against Facebook. `https://www.eff.org/deeplinks/2011/05/syrian-man-middle-against-facebook`. Accessed June 2016.

[8] S. Egelman, L. F. Cranor, and J. Hong. You've been warned: An empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of CHI*, 2008.

[9] S. Egelman and S. Schechter. The importance of being earnest [in security warnings]. In *Financial Cryptography and Data Security, Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2013.

[10] A. P. Felt, A. Ainslie, R. W. Reeder, S. Consolvo, S. Thyagaraja, A. Bettes, H. Harris, and J. Grimes. Improving SSL warnings: Comprehension and adherence. In *Proceedings of CHI*, 2015.

[11] A. P. Felt, R. W. Reeder, H. Almuhimedi, and S. Consolvo. Experimenting at scale with Google Chrome's SSL warning. In *Proceedings of CHI*, 2014.

[12] E. Hjelmvik. Analysis of Chinese MITM on Google. `http://www.netresec.com/?page=Blog&month=2014-09&post=Analysis-of-Chinese-MITM-on-Google`. Accessed June 2016.

[13] E. Hjelmvik. Forensics of Chinese MITM on GitHub. `http://www.netresec.com/?page=Blog&month=2013-02&post=Forensics-of-Chinese-MITM-on-GitHub`. Accessed June 2016.

[14] S. Kim and M. S. Wogalter. Habituation, dishabituation, and recovery effects in visual warnings. In *Proceedings of Human Factors and Ergonomics Society Annual Meeting*, 2009.

[15] A. Kingsley-Hughes. Gogo in-flight wi-fi serving spoofed ssl certificates. `http://www.zdnet.com/article/gogo-in-flight-wi-fi-serving-spoofed-ssl-certificates/`, January 2015.

[16] K. Krol, M. Moroz, and M. A. Sasse. Don't work. Can't work? Why it's time to rethink security warnings. In *Proceedings of the International Crisis on Risk and Security of Internet and systems (CRiSIS)*, 2012.

[17] S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer. The emperor's new security indicators: An evaluation of website authentication and the effect of role playing on usability studies. In *Proceedings of IEEE Symposium on Security and Privacy*, 2007.

[18] H. K. Sotirakopoulos, A. and K. Beznosov. On the challenges in usable security lab studies: Lessons learned from replicating a study on SSL warnings. In *Proceeings of SOUPS*, 2011.

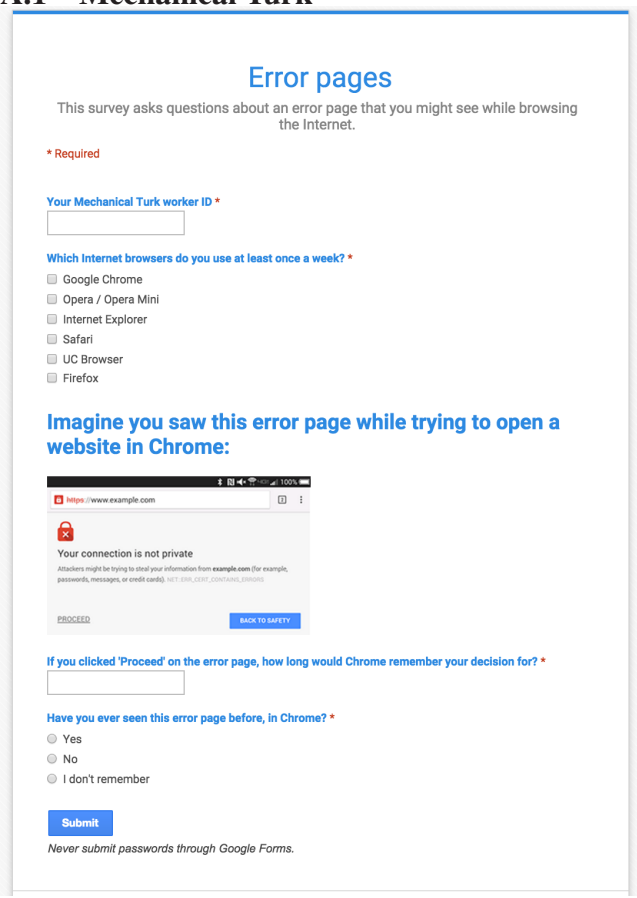
[19] J. Sunshine, S. Egelman, H. Almuhimedi, N. Atri, , and L. F. Cranor. Crying wolf: An empirical study of SSL warning effectiveness. In *Proceedings of USENIX Security*, 2009.

[20] P. Thorley, E. Hellier, and J. Edworthy. Habituation effects in visual warnings. *Contemporary Ergonomics*, 2001.

APPENDIX

A. SURVEY SCREENSHOTS

A.1 Mechanical Turk



Error pages

This survey asks questions about an error page that you might see while browsing the Internet.

* Required

Your Mechanical Turk worker ID *

Which Internet browsers do you use at least once a week? *

- Google Chrome
- Opera / Opera Mini
- Internet Explorer
- Safari
- UC Browser
- Firefox

Imagine you saw this error page while trying to open a website in Chrome:

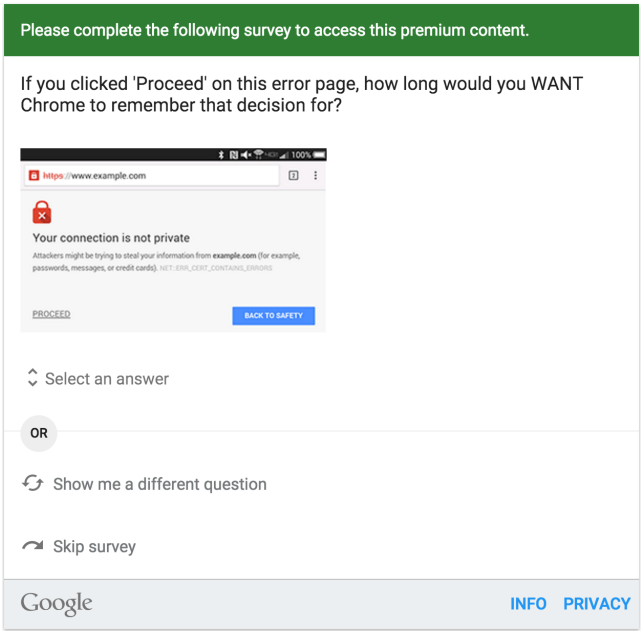
If you clicked 'Proceed' on the error page, how long would Chrome remember your decision for? *

Have you ever seen this error page before, in Chrome? *

- Yes
- No
- I don't remember

Submit

Never submit passwords through Google Forms.



Please complete the following survey to access this premium content.

If you clicked 'Proceed' on this error page, how long would you WANT Chrome to remember that decision for?

Select an answer

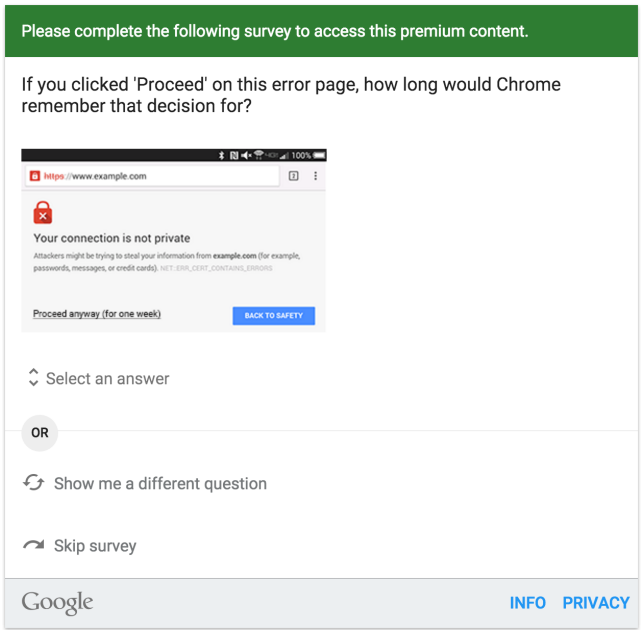
OR

Show me a different question

Skip survey

Google INFO PRIVACY

A.2 Google Consumer Survey



Please complete the following survey to access this premium content.

If you clicked 'Proceed' on this error page, how long would Chrome remember that decision for?

Select an answer

OR

Show me a different question

Skip survey

Google INFO PRIVACY

Follow My Recommendations: A Personalized Privacy Assistant for Mobile App Permissions

Bin Liu,* Mads Schaarup Andersen, Florian Schaub, Hazim Almuhimedi
Shikun Zhang, Norman Sadeh,* Alessandro Acquisti, Yuvraj Agarwal
Carnegie Mellon University
Pittsburgh, PA, USA
{bliu1, manderse, fschaub, hazim, shikunz, sadeh, yuvraj.agarwal}@cs.cmu.edu
acquisti@andrew.cmu.edu

ABSTRACT

Modern smartphone platforms have millions of apps, many of which request permissions to access private data and resources, like user accounts or location. While these smartphone platforms provide varying degrees of control over these permissions, the sheer number of decisions that users are expected to manage has been shown to be unrealistically high. Prior research has shown that users are often unaware of, if not uncomfortable with, many of their permission settings. Prior work also suggests that it is theoretically possible to predict many of the privacy settings a user would want by asking the user a small number of questions. However, this approach has neither been operationalized nor evaluated with actual users before. We report on a field study (n=72) in which we implemented and evaluated a Personalized Privacy Assistant (PPA) with participants using their own Android devices. The results of our study are encouraging. We find that 78.7% of the recommendations made by the PPA were adopted by users. Following initial recommendations on permission settings, participants were motivated to further review and modify their settings with daily "privacy nudges." Despite showing substantial engagement with these nudges, participants only changed 5.1% of the settings previously adopted based on the PPA's recommendations. The PPA and its recommendations were perceived as useful and usable. We discuss the implications of our results for mobile permission management and the design of personalized privacy assistant solutions.

1. INTRODUCTION

Mobile app ecosystems such as Android or iOS compete in part based on the number, and the quality, of apps they offer. To attract developers and help generate more apps, these platforms have exposed a growing number of APIs. These APIs provide access to smartphone functionality (e.g., GPS, accelerometer, camera) and user data (e.g., unique identifiers, location, social media accounts), much of which is privacy-sensitive.

*Main contacts: Bin Liu and Norman Sadeh.

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22–24, 2016, Denver, Colorado.

While the Android and iOS platforms both rely on permission-based mechanisms and allow users to control access to sensitive data and functionality, the end result is an unwieldy number of app-permission decisions that users are expected to make. Estimates indicate that users, on average, have to make over one hundred permission decisions (95 installed apps on average per user [48]; 5 permissions on average per app [37]). Prior work has shown that users are often unaware of – if not uncomfortable with – many of the permissions they have ostensibly consented to at some point (e.g., [6, 8, 16, 17, 21, 24]).

To help overcome the burden associated with managing such a large number of decisions, prior research suggests that – despite the diversity of users' privacy preferences – it is theoretically possible to predict many of a user's permission settings by asking the user a small number of questions [28, 29]. These approaches suggest that, using machine learning, it may be possible to reduce user burden when it comes to configuring mobile app permission settings. However, this approach has not been fully operationalized so far.

We propose a practical solution that operationalizes privacy preference modeling in a personalized privacy assistant (PPA) by (1) developing privacy profiles for users, (2) determining which of these profiles is the best match for a given user, and (3) configuring many of the user's permissions based on the selected profile. This paper is the first to report on the implementation and field evaluation of a personalized privacy assistant (PPA) for mobile app permissions.

We propose a methodology to learn privacy profiles for permission settings and leverage these profiles in a personalized privacy assistant that actively supports users in configuring their permission settings. In a field study we collected permission settings from 84 Android users with rooted smartphones who received privacy nudges designed to motivate them to interact with their permission settings. Mobile app permission settings collected from these users were organized along three dimensions: app categories, app permissions and purposes associated with each permission (e.g., supporting an app's core functionality versus advertising). The resulting data was used to identify clusters of like-minded users and to generate recommended permission settings (or "profiles") for users in each cluster. Our results indicate that despite relying on app permission settings collected from a small number of users (n=84), our learned privacy profiles can accurately recommend mobile app permission settings that users are likely to adopt.

Our personalized privacy assistant uses information about the apps installed on a user's smartphone to elicit the user's privacy preferences and offer recommendations on how to configure associated

permission settings. We designed an interactive profile assignment dialog, in which the PPA relies on dynamically-generated decision trees to generate questions that help match users to the privacy profile that best aligns with their preferences, which is then used to provide recommendations on which permissions to deny. The PPA gives the user the option to accept multiple recommended settings at once and the ability to modify them as needed.

We show the effectiveness and usability of a profile-based PPA through a field study. The profiles built using permission settings collected from the first set of users (n=84) were used by our PPA, which we evaluated in a second between-subjects field study with different participants (n=72). This enabled us to evaluate the effectiveness and usability of the PPA on participants' own (rooted) Android smartphones. Our results show that 78.7% of the recommendations made by the PPA were accepted by participants in the treatment group, and only 5.1% of recommended permission settings were later revised by participants, despite being exposed to privacy nudges designed to motivate them to revisit their earlier decisions. Participants in the treatment group also converged faster on their settings and reported satisfaction with the recommendations and the PPA functionality.

Our results provide rich insights on the interaction design of personalized privacy assistants, permission managers, mobile privacy nudges, and their interplay. These insights are relevant for developers of mobile platforms, privacy tools, and mobile apps.

2. RELATED WORK

Our work relates to research on mobile privacy, mobile app permissions, privacy awareness, and building privacy profiles for users.

2.1 Mobile App Privacy

Prior work has shown that many mobile apps access sensitive functionality and data for purposes that are not limited to the delivery of their core functionality [5, 13, 27, 49]. Sensitive resources and data commonly accessed by mobile apps, whether on iOS or Android, include unique device identifiers (e.g., IMEI), user location, contacts list, camera, texting, and much more. Many apps share sensitive personal information with advertising networks and analytics companies, which in turn use the data to build extensive user profiles [1,34,47,49]. Research shows that users are often unaware of the extent of these practices and that many will express reservations and concern when they learn about them [18, 23, 25, 27, 45].

2.2 App Privacy Management

Functionality that enables users to manage mobile app permissions has evolved quite significantly in recent years – for both iOS and Android. While early versions of iOS only allowed users to control access to their location, the number of such permissions has increased in each new version of iOS. In iOS 9, 11 categories of permissions exist with settings enabling users to grant or deny individual permissions on an app-by-app basis, at the time the permission is requested by an app. Until recently, the user privacy controls provided by Android were fairly limited. They mainly involved displaying a list of permissions to the user when installing an app and asking the user to confirm that they consent to grant all the requested permissions. In Android 6.0, this has changed, with both Android and iOS now offering very similar control over mobile app permissions to their users. While this increase in control is a positive development, it also exposes users to a large number of privacy settings.

Prior work has shown that mobile app permission screens at install time are largely ineffective in helping users make informed privacy decisions, because most users do not pay close attention to the permissions screen and do not understand what the permissions mean or entail [16, 23]. Alternative designs that highlight privacy implications (e.g., how personal information is shared with advertisers [24] or unexpected data collection practices [27]) have been more effective in helping users avoid what they perceive as intrusive apps [9, 21, 24, 27, 35, 50]. Instead of assisting decisions about whether to install an app, our work focuses on helping users manage their privacy for apps already installed on their devices.

In Android 6.0, Google replaced install-time permission screens with just-in-time permission requests and a permission manager [7], reminiscent of iOS' permission management approach. Prior work has explored the utility and usability of such permission managers showing how users employ them to limit app access to personal information [6, 19]. Fisher et al. found that the majority of iOS users in their study prevented a third of their apps from accessing the users' location [19]. Similarly, Almuhimedi et al. found that 65% of Android users in their study utilized the permission manager to control how apps access personal information [6]. However, they also showed that the permission manager alone is not sufficient for users to reach satisfying levels of privacy protection because the permission manager does not provide enough information to assist users in making informed privacy decisions [6]. To account for such a limitation, we enrich the permission manager in our study with additional information such as the purpose and access frequency information for specific permissions.

Both iOS and Android 6.0 encourage app developers to specify a purpose in permission request dialogs in order to enable users to make informed privacy decisions. Tan et al. evaluated the prevalence of such developer-specified explanations in iOS apps (only 19% of permission requests had explanations) and observed that while users did not really understand them they were still more likely to grant requests if an explanation was provided [46]. Using experience sampling, Shih et al. find an opposite effect: participants shared more when permission requests did not contain explanations, whereas vague explanations decreased users' willingness to grant permission requests [44]. Instead of relying on developer-specified explanations, we notify users of the likely purpose of an app's permission request, based on static code analysis results from PrivacyGrade [2,27,28]. Prior work indicated that purpose explanations play an important role in making privacy decisions [6,27,44].

A number of recent studies explored approaches to help users manage their privacy for apps they already installed on their devices [6, 8, 20]. Fu et al. showed in a field study that a full-screen and interruptive privacy notification is more effective than an uninterruptive icon in the notification area in informing users when apps access their location [20]. However, users were annoyed by the full-screen notifications, especially when apps accessed location frequently [20]. Using just-in-time notifications when personal information is accessed and a summary of how frequently apps access users' information, Balebako et al. showed that users are in general unaware of data collection practices by apps and that users are surprised at how frequently apps access their personal information [8]. Both Fu et al. and Balebako et al. did not provide users with tools to exercise control over how apps access users' personal information. In contrast, we enabled our users to manage their app privacy settings through an enhanced permission manager. To explore whether interventions can motivate users to review their app privacy settings, Almuhimedi et al. designed "privacy nudges" that inform users of how frequently apps access personal information (e.g., location), and also enable users to adjust their app settings [6].

They found that nudges indeed increase awareness of apps' behaviors and motivate users to review and adjust their app permissions.

In this paper, we build on some of the ideas proposed by prior work. In particular, in addition to showing frequency of access to private data, we also show the inferred purpose of the access using the public PrivacyGrade dataset [2]. Second, while we build upon the idea of privacy nudges, we extend it to elicit user preferences on a set of privacy-related questions to build privacy profiles with machine learning. Finally, we build on prior work on using privacy profiles to reduce user burden in terms of decisions, but we extend it to use privacy nudges to help users review their settings after profile assignment to ensure that profile-based settings match users' actual preferences. Most importantly, our PPA app integrates these aspects in an end-to-end system to evaluate their effectiveness in real-world settings.

2.3 Privacy Profiles and Preference Modeling

Privacy controls, such as permission managers, enable users to configure their privacy settings. However, the growing number of configurable privacy settings makes it difficult for users to align their privacy settings with their actual preferences [6, 32] Agarwal and Hall [5] and Rashidi et al. [39] proposed crowd-powered and expert-powered systems to recommend settings to users. However, users' app privacy settings are diverse [29], rendering one-size-fits-all solutions insufficient to accurately capture users' diverse preferences.

Researchers have proposed modeling and predicting users' privacy preferences. Collaborative filtering has been proposed for location sharing preferences [53, 54]. However, the proposed approaches were only evaluated in simulations. In real-world scenarios for mobile apps, the collaborative filtering solutions would suffer from data sparsity and the cold-start problem, where the model requires sufficient user feedback before giving accurate recommendations. Ismail et al. [22] proposed a collaborative-filtering-based recommender for security configurations of mobile apps. They determined a sufficiency threshold for user input before providing recommendations. And they pre-determined diverse scenarios for users to ensure informativeness of the training input.

Privacy profiles, which are collections of related privacy and sharing rules that correspond to privacy preferences of similar-minded users [11, 15, 26, 28, 29, 40, 51, 52], can provide decision support if one can identify a privacy profile that matches a new user. In the context of online social networks, Fang and LeFevre suggested using active machine learning to design a "privacy wizard" to assist Facebook users in managing their complex privacy settings [15]. The authors evaluated the privacy wizard using real data from 25 Facebook users and showed that the privacy wizard can predict users' privacy settings with high accuracy (above 90%) and minimal effort by users (only labeling 25 friends) [15]. In the context of mobile app privacy, recent work has explored utilizing related approaches. Lin et al. [28] generated privacy profiles for app privacy settings, taking into consideration purpose information and users' self-reported willingness to potentially grant access, elicited in a scenario-based online study. However, the privacy paradox suggests that self-reported preferences may not necessarily reflect actual privacy behavior [10, 31]. In contrast, Liu et al. identified six privacy profiles based on 239K real users using only their app privacy settings [29]. However, prior work shows that permission settings alone might not reflect users' actual privacy preferences, because users may be unaware of many apps' data collection practices occurring in the background [6]. In contrast, we built privacy profiles from users' real-world permission settings collected in a field study using permission settings, purpose information as well as app categories to obtain a diverse set of profiles from a comparatively smaller dataset. We further use privacy nudges to make users aware of unexpected data practices and thus elicited privacy settings likely better aligned with users' privacy preferences.

In contrast to prior work, we evaluated the effectiveness of our privacy profiles with actual users in a field study, thereby, demonstrating the practical impact of privacy profiles on mobile privacy configuration. Few others have evaluated privacy profiles on real users' phones in the field. Wilson et al. studied privacy profiles in the context of a location-sharing system [51]. They found that privacy profiles impacted users' privacy decisions and satisfaction level. However, they evaluated their privacy profiles based on simulated location requests, whereas we evaluated our privacy profiles based on real permission requests on participants' own smartphones.

3. PPA OVERVIEW

We designed and implemented a profile-based personalized privacy assistant (PPA).[1] Specifically, the PPA uses apps on the user's smartphone to engage in a dialog and elicit a small set of preferences pertaining to whether or not the user feels comfortable granting some permissions to apps from certain categories. Using these answers, the PPA identifies a privacy profile that best matches the user's preferences and, based on this profile, recommends a number of permission settings changes to the user. The user is given the option to accept or change recommendations individually or in bulk. The specific set of questions the PPA asks a user is determined by the user's installed apps and dynamically adapts as the user answers questions.

Developing and deploying our PPA involved multiple steps. We first collected users' app privacy preferences using an enhanced permission manager on rooted Android devices to develop mobile app privacy preference profiles. We organized users into clusters of like-minded people, and developed profiles for each cluster to capture typical user preferences. Next, a field study was conducted where we deployed the PPA to newly recruited users, also with rooted Android devices. In this study, the PPA used its profiles to engage in dialogs with users and assign them to a particular cluster. The profiles were finally used to recommended specific mobile app permission settings to users. This is further detailed below.

Enhanced Android Permission Manager

For the purpose of accurately capturing users' privacy preferences from their privacy settings, we assume that users are comfortable with a restrictive permission setting they chose, if they keep the setting and do not change it back to a permissive setting. To increase users' awareness and engagement, so that they review their permission settings if they find a setting they do not agree with, we made a number of modifications and enhancements to the Android permission manager App Ops [12], which we describe below.

Simplified controls. In the permission manager, we organized permission settings into six groups of privacy-related permissions: Location, Contacts, Messaging, Call Log, Camera, and Calendar. As a result, multiple permissions are represented as a single permission, reducing the overall number of permissions users have to consider. For example, `READ_CONTACTS` and `WRITE_CONTACTS` are represented as "Contacts." This grouping is partially based on results by Lin et al. [27] and Felt et al. [16]. Users can directly allow or deny

[1] Our personalized privacy assistant app is publicly available at: www.privacyassistant.org

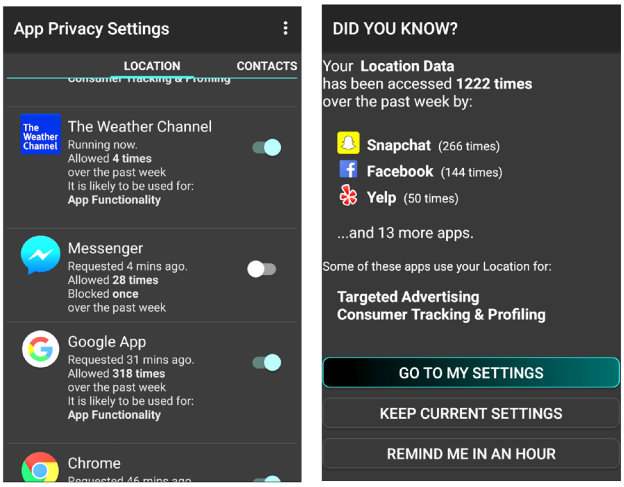


Figure 1: Permission manager (*left*) and a daily privacy nudge (*right*), which include the access frequency and purpose information.

each permission while reviewing them in the permission manager.[2]

Enhanced Awareness. We extended the permission manager to show not only an app's most recent access requests, but also how often the app requested access over the last seven days, as shown in Figure 1. We further included purpose information from PrivacyGrade [2,28] for apps for which it was available. Using Androguard static analysis [27], PrivacyGrade identifies the likely purpose(s) of an app's permission requests by analyzing its third-party libraries (e.g., app functionality, targeted advertising, consumer tracking & profiling, or sharing with social network services).

Privacy Nudges. Nudges have been found to be effective at increasing users' privacy awareness and motivating them to review and adjust their permissions [6, 9]. We adopt a similar nudging strategy to get users to reflect on their permissions and engage with our permission manager to adjust their settings, in order to collect rich permission settings from each user. Our privacy nudge, shown in Figure 1, includes access frequency for the given permission [6], other apps that accessed the same permission, and, if known, the likely purpose of the access for that permission. From the nudge, users can open the permission manager to change their settings, keep the current settings and close the nudge, or postpone managing their privacy.

Building Profiles

After deploying our enhanced permission manager to users, we collect their real-world permission settings. For each permission setting, we collect the likely purpose of the permission request from PrivacyGrade [2], and the category of the requesting app from the Google Play store. We use app categories as features, rather than individual apps, to reduce over-fitting caused by less popular apps and limited training samples. Using this training data, we build user profiles by applying hierarchical clustering [43] on the feature vectors generated from a set of features. We describe the process of building privacy profiles from real users' privacy settings in more detail in Section 4.

[2]Coincidentally, Google announced similarly grouped permissions for Android 6.0 shortly after we conducted our first field study.

Assigning users to privacy profiles

In order to assign new users to the generated privacy profiles, we ask them a small number of tailored questions about their privacy preferences. To generate these questions, we first aggregate user preferences in the training data set by (a) each permission; (b) each (permission, app category) pair; and (c) each (permission, purpose) pair. Each aggregated feature represents a potential question to ask a new user. However, we first check whether users have apps installed that fit the particular question. For example, to be asked a question about preferences for (location, advertisement), the user must have at least one app installed that accesses location for advertisement purposes. We then train a C4.5 decision tree [38] on the set of questions applicable to a particular user, and generate an ordered list of questions. Users are asked 5 questions at most to be assigned to a profile. Note that with our method the set of questions is dynamically personalized for each user, so that the questions can be contextualized using the apps each user has installed on their phones.

Generating recommendations

On the server side, we train a scalable SVM classifier (LibLinear [14]) using the permission settings we collected from the profile-building procedure mentioned above. The PPA app will pass the user's features to the classifier to generate recommendations for privacy settings learned from the training data. The features we include are the user's assigned profile, app category, permission, and purposes. Even though our model can make recommendations for each (category, permission, purpose) tuple, Android's permission model does not support granular control by purposes. Therefore, our personalized privacy assistant provides privacy recommendations to deny access based on permission and app categories, while we use purpose information to further explain our recommendations. Note that we only provide recommendations to deny access, as permissions were allowed by default once an app was installed prior to Android 6.0.

Next, we discuss our process for building privacy profiles in Section 4, followed by a discussion of the design of our personalized privacy assistant in Section 5.

4. BUILDING PRIVACY PROFILES

To obtain real users' permission settings from which to build privacy profiles, we conducted a first field study in which we deployed our enhanced permission manager to actual Android users.

4.1 Privacy Settings Dataset Collection

Since permission management requires system privileges, this study (as well as the later evaluation of our PPA) had to be conducted with users of rooted Android phones. Importantly, our participants installed our app on their own rooted Android phones – namely the phones they use in their regular daily activities. In previous online surveys and studies using dialogs on simulated phone screens [28, 50], settings selected by participants were not applied to devices actually used by these participants. In contrast, our approach allows us to collect real settings stemming from user behavior, rather than aspirational responses that don't match users' behavior [31]. While users of rooted Android phones may constitute a biased population, this approach still allows us to evaluate the practicality of building privacy settings profiles, and using a PPA, on real users. Assuming it will be possible to customize permission management in future versions of mobile platforms, the same approach can be adopted to build privacy profiles representative of the general population's privacy settings.

Our study was approved by Carnegie Mellon University's Institutional Review Board. We recruited Android phone users (>1 month use) who used a rooted Android phone (4.4.X or 5.X; Android 6.X had not been released at the time of the study) with a data plan. Considering that our target population is limited to users of rooted Android phones, we recruited participants from multiple online communities related to Android in general or rooted Android in particular on Facebook Groups, Google+ communities, Reddit subreddits, and tech forums. We disclosed that the study app collected and managed Android app privacy settings as it would have root access to participants' phones. All participants had to be 18 years or older. We asked participants to complete an initial screening survey to verify that they matched the above criteria and to collect demographic information. Participants who qualified were sent a download link for our permission manager and a user name to activate it.

In the first week of the study, participants could use the permission manager to selectively deny or allow permissions. Our app also collected the frequencies of permission requests for installed apps, which were shown in the permission manager. In the second week, the participants received a privacy nudge once a day, between 12pm and 8pm. Figure 1 shows both the permission manager (left) and the nudge dialog (right). We waited one week before showing daily nudges to allow participants to familiarize themselves with the enhanced permission manager and to ensure that the privacy nudge messages contained meaningful access frequencies based on the behavior of participants' installed apps. The privacy nudges provided information about one of six permissions available in the enhanced permission manager. The selection of which nudge was shown was randomized to counter order effects. If a particular permission had never been accessed by apps on the participant's device (access frequency would be zero), another permission would be selected to be shown in the nudge instead.

After participants completed the study, we asked them to fill an exit survey online, consisting of the 10-item IUIPC scale on privacy concerns [30] and an 8-item scale on privacy-protective behavior [36]. They were compensated with a $15 giftcard afterwards. We further invited all participants to an optional interview, in which we explored their reasons for restricting or allowing different permissions, their comfort level concerning their permission settings, and the usability of the enhanced permission manager and privacy nudges. Those who participated in the optional interview received an additional $10 giftcard.

4.2 Dataset Analysis

In total, we collected data and survey responses from 84 Android users, and interviewed 10 of them. The 84 participants originated from North America (66; 62 U.S.), Europe (10), Asia (7), and South America (1). Given the target population of rooted phone users, we expected our study population to skew towards young, tech-savvy males. Indeed, the majority of our participants were male (78 male, 6 female) and 18–54 years old (median 23). Among them, 8 had a graduate degree, 22 a Bachelor's degree, and 5 had an Associate's degree; 30 attended some college, and 19 had a high school degree or lower. Most commonly reported occupations were student (35), computer engineer or IT professional (8), service (5), and unemployed (5). Participants exhibited relatively high privacy concerns, scoring high on the IUIPC [30] scales for control (median 6.33, mode 6.33, min 2.33, max 7), awareness (median 6.67, mode 7, min 4, max 7), and collection (median 6, mode 7, min 1.25, max 7). They also took more measures to protect their online privacy compared to the general population [36], as shown in Table 1. This suggests, that our participants' privacy settings may be more conservative than those of the general population.

In total, we obtained 4,197 permission settings from 84 participants, reflecting their allow and deny settings of the 6 permissions in the enhanced permission manager. We filtered the dataset to only analyze permission settings for apps available in the Google Play Store. Because Android permission requests of installed apps are set to allow by default,[3] we analyzed only those permission settings for which the corresponding app had been launched in the foreground at least once during the study, or if users explicitly denied or allowed an app's permissions. After filtering, our dataset consisted of 3,559 individual permission settings for 729 distinct apps.

Of the 3,559 permission settings, 2,888 were allowed (81.15%, mean: 34.38 per user), which is the default choice, and 671 (18.85%, mean: 7.99 per user) were denied by participants. Call Log requests were denied the most (41.33%), while Camera access was allowed the most (95.07%). Of the permissions participants changed explicitly, 7.58% were re-allows of permissions they had previously denied. In the interviews, we asked participants why they did not deny certain apps, in cases where they re-allowed or just never changed an app's permission. The main reason for re-allowing a permission, as mentioned by two interviewees, was that denying it broke or might break app functionality. P6 noted "The moment I turned it off I realized that it wasn't gonna send me any messages." Nine interviewees reported not denying permissions, because they were required for the app to function. Two interviewees noted that they trusted the app or the app provider. P2 stated "This fitness app is made by Google and I trust it so I allowed it."

We fitted the users' settings data to a random effect logistic regression model grouped on users' allow/deny decisions on app permissions. The independent variables include major features that could be obtained in our dataset such as user demographics and app category. App category information was retrieved from the Google Play store. The detailed logistic regression results are shown in Table 2 in Appendix A. App category and the type of permission are significant predictors for an individual's allow or deny decision, whereas demographics, privacy concerns, the app name, access frequency and purpose information were not significant.

Participants largely agreed on permission settings for certain app categories. For example, apps in the "Books & Reference" category were always denied access to Contacts and Call Log, while "Photography" apps were always allowed access to Camera, as is to be expected. Participants' aggregated settings on app categories are somewhat diverse (average SD=0.388, if we define allow=0, deny=1). The detailed effect size (odds ratios) can be found in Table 2. Eight interviewees mentioned that they denied access based on app functionality, e.g., when the use of the permission was not clear or when they thought that an app would not need it. P4 stated: "I do not use Facebook for any calendar function so I denied it access to my calendar." Four interviewees mentioned denying apps when they did not use them, especially pre-installed apps they did not uninstall.

Nine interviewees (out of ten) confirmed the usefulness of access frequency information; four stated it was as a reason to deny a permission, five mentioned it was useful in the nudge, and two stated

[3] All participants use Android 4.4.X or 5.X phones, where app permissions were granted by default when an app is installed. Android 6 prompts users to grant or deny permission requests, thus making this pre-processing unnecessary.

it was useful in the permission manager. For example, P1 stated: "Didn't notice that the app had actually accessed the location that many times. It is pretty crazy." However, despite reported usefulness, we did not find significant impact of access frequency on users' decision of permission settings (see Table 2).

The logistic regression model indicates that purpose information was not a significant predictor for whether a permission is denied in our dataset. A likely reason is the sparsity of purpose information compared to app category and permission type which are always available. Our purpose information stems from PrivacyGrade's dataset [2], which covers popular free apps on Google Play. During the study, purpose information was shown for 8.6% of apps requesting Location access, 35.1% for Contact, and 42.5% for Camera requests. Of the daily privacy nudges, 60.4% contained purpose information; 31.45% of those nudges showed purposes other than required for app functionality. Participants denied less if any purpose(s) were shown (13.53% compared to 19.95%; Chi-square=10.1793, df=1, p=0.0021, effect size(odds ratio)=0.6784), which matches Tan et al.'s results [46]. However, none of the purposes had significant impact on users' decisions (see Table 2). Participants further agreed on some specific cases. For instance, 100% allowed Contacts for Social Network Services and 95.63% allowed Camera for App Functionality. Nine interviewees mention that purpose information was useful; three as a reason to deny, seven as useful in the nudge, and three as useful in the permission manager. Three interviewees mentioned a trade off when applications had more than one purpose stated. They wanted the app's main functionality that needed a permission, but did not like that it was being used for other purposes. P3 stated "Snapchat is a tradeoff. Although I'm not happy they access my contacts for tracking I think I will allow them to access my contacts because of the function they provide." Participants' choices were typically permissive in such cases. This suggests that the additional purpose information is useful to participants and it would be desirable to provide it for more apps. However, it seems some purposes also caused confusion. P3 had problems understanding the meaning of "Consumer Tracking / Profiling." Thus, more research is needed to reliably determine purposes of permission requests, convey this information to users, and enable users to make access decisions for specific purposes. We discuss these aspects in more detail in Section 7.2.

4.3 Generating Privacy Profiles

From the collected dataset, we obtained users' detailed app permission settings as a collection of rows in the form of (user, app, permission, decision). We collected app category information from the Google Play store. Purpose information is based on PrivacyGrade data [2], which provides an indication of the purposes an app may use requested data for, but does not provide purpose information for all apps or permission requests.

4.3.1 Clustering Approach

We quantify each user's preferences as a three-dimensional tensor of aggregated preferences of (app category, permission, purpose). For each cell, we define the value as the tendency of the user to allow or deny permissions requested by apps from a specific category with a corresponding purpose: from -1 (100% deny) to 1 (100% allow), and N/A if we do not have the user's settings data for a cell. To estimate similarities among participants' feature tensors, we impute the missing values in the tensors. In order to impute without biasing any dimension, we apply weighted PARAFAC Tensor factorization [3]. We put 1-weight on all known data cells and 0-weight on unknown data cells in the tensor. Thus, we optimize the overall error of the imputed tensor in Frobenius norm using only the values known from the data. Using the users' feature vectors reshaped from the imputed tensor, we build user profiles by applying hierarchical clustering [43] on the feature vectors. We choose hierarchical clustering since it is not sensitive to the size or density of clusters and allows non-Euclidean distances.

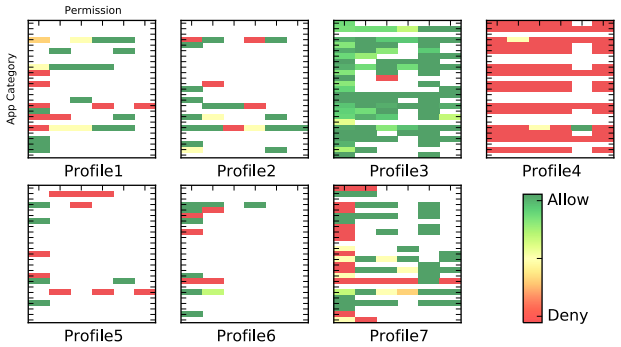


Figure 2: Privacy profiles learned from collected app privacy settings. Profile 1 is more protective on Location and Productivity apps than other profiles. Profile 2 denies phone call log permission more. Profile 3 is generally permissive. Profile 4 denies most permission requests. Profile 5 generally denies contacts, message, phone call log and calendar access, with only location and camera allowed for some apps. Profile 6 denies location and contact access of Social apps and Finance apps. Profile 7 is stricter regarding Social apps and location access in general.

4.3.2 Generating Recommendations

The profile-based recommended settings are generated by a scalable SVM Classifier (LibLinear [14]) on the decision of each permission request. The features of the classifier consist of the user's assigned profile, the category of the corresponding app, the permission requested, and the likely purpose(s) of the permission request. The classifier is pre-trained using the permission settings data we collected when building privacy profiles, with the profile assignment information of the users in the dataset.

4.3.3 Resulting privacy profiles

We applied a grid-search of the parameters for the hierarchical clustering and the SVM classifier to choose the ones that have better cross-validated F-1 scores of the accuracy of the recommended items to deny. We tried Manhattan, Euclidean, and Cosine distances in the grid search of parameters for hierarchical clustering, and tried Gamma={0,1e-3, 1e-4} and C={1e-4, 1e-3, ..., 1e3} for the linear-kernel SVM. With 5-fold cross-validation on the dataset described in Section 4.2, we found the optimized mode for the dataset (hierarchical clustering: K=7, complete linkage, cosine distance, Silhouette Coefficient=0.2079; classifier: Gamma=1e-3, C= 1e3, hinge loss) with a cross-validated F-1 score of 90.02%. In contrast, if we train a global model for all users without splitting them into profiles, the best F-1 score would be 74.24%, much lower than the profile-based optimized model.

Figure 2 shows the permission preferences in each profile aggregated by app categories. It provides an overview of the diversity in privacy preferences among the different profiles. Profile 3 contains 67 of the 84 participants (79.8%), who are generally permissive. Profile 4 contains 2 participants (2.4%), who denied most permission requests. Note that the majority of participants were grouped in the most permissive profile (profile 3) despite our privacy-conscious and tech-savvy participant population. The remaining profiles (15 participants, 17.8%) express variations in privacy preferences depending on app category and permission of ac-

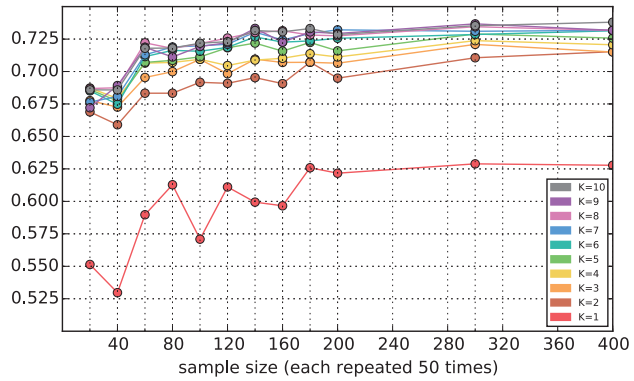


Figure 3: Down-sampling simulation on Lin et.al's dataset [28] (F-1 score). With 5 profiles or more training on data from just 80 users provides reasonable F-1 score (> 70%). When training on 400 users, the accuracy improves, but only marginally.

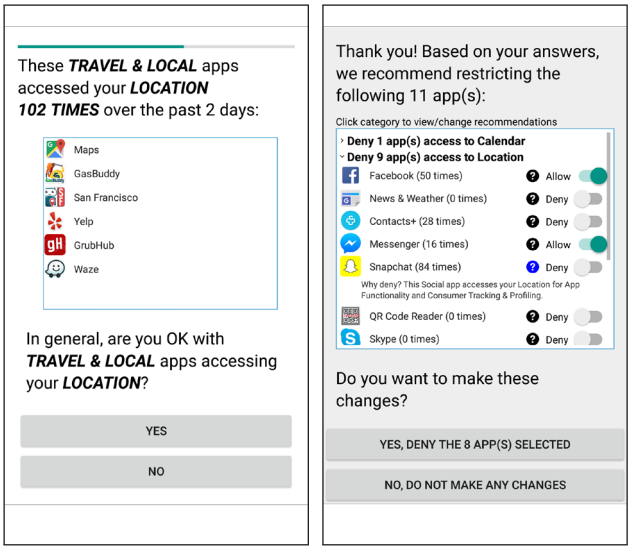


Figure 4: Profile assignment dialog: After answering up to 5 questions (*left*) users may receive personalized recommendations (*right*). Users can review and customize the recommended deny settings.

cess. Profile 1 (3 participants) is more protective on Location and on apps in the category of Productivity comparing to other profiles. Profile 2 (4) denies phone call log permission more. Profile 5 (1) generally denies contacts, message, phone call log and calendar permission access to all apps, with only location and camera allowed for some. Profile 6 (3) denies location and contact access of Social apps and Finance apps. Profile 7 (4) is restrictive for Social apps and location access in general.

Lin et al. [28] identified similar profiles. Their "unconcerned" profile corresponds to our profile 3, their "conservative" profile to profile 4, and their "fence-sitter" and "advanced users" profiles align with our more specialized profiles (profiles 1, 2, 5, 6, 7).

4.3.4 *Downsampling comparison*

Given the relatively small number of 84 participants in our dataset, a potential concern is whether our profiles are expressive enough to cover privacy preferences of a larger user population, and whether we can provide useful recommendations. To explore the utility of our profiles, we applied our approach for building profiles to Lin et al.'s considerably larger dataset [28]. This dataset has 21,657 records in total, consisting of 725 MTurkers' self-reported preferences of 540 apps accessing permissions for specific purposes, whereas our dataset consists of 3,559 permission settings by 84 participants for 729 apps. To compare the effects of different dataset sizes, we down-sample their dataset by removing randonmly-selected users to create smaller datasets, ranging from 20 to 400 users in size, which is more than half of the entire dataset. Figure 3 shows F-1 scores for 1–10 profiles.

The results show that with as little as 80-100 users, which corresponds to our sample size (n=84), the F-1 score can already reach 0.725, only slightly different from the larger sample sizes, which get best F-1 scores around 0.73. Obviously, with training data from more users our recommendation accuracy is likely to increase, but this experiment suggests that learning profiles from 84 participants already results in profiles sufficiently stable to be used in practical applications.

5. PROVIDING RECOMMENDATIONS

Our PPA app elicits a user's privacy preferences with an interactive dialog to provide the user with personalized recommendations. Thus, the PPA's recommendation process consists of two main components: (a) First, the PPA shows a series of dynamically-generated questions to elicit the user's app privacy preferences and assign the user to a privacy profile. (b) Then, the PPA provides profile-based recommendations according to the user's privacy profile and installed apps. The user can review and adjust recommended settings before applying them.

5.1 Interactive Profile Assignment

The profile-assignment questions elicit a user's preferences for (1) individual permissions, (2) permission and app category pairs, and (3) permission and purpose pairs. Each question has a Yes/No response. For a new user, the PPA dynamically generates a decision tree that uses input from a question to determine the next question to ask and eventually assign the user to one of our privacy profiles. Users are asked 5 questions at most to be assigned to a profile. The decision tree is generated based on profile assignments and aggregated preferences from the dataset used to build the privacy profiles, as well as the user's installed apps. Considering installed apps allows us to contextualize the decision tree by excluding questions for which the user has no apps installed. For example, if the user has no Game app installed, the PPA would not ask if the user would generally allow Game apps to access location.

To contextualize the questions in the profile assignment dialog, installed apps that fit the particular question are listed in the dialog with their access frequency for the respective permission, inspired by Almuhimedi et al.'s privacy nudges [6]. Figure 4 shows an example of an assignment dialog question. In this example, installed apps from the Travel & Local category have accessed the Location permission 102 times over the past 2 days. A progress bar at the top shows how many questions have been completed.

5.2 Profile-based Recommendations

After a user has responded to the questions, the PPA assigns a privacy profile to the user, which is used to determine which recommendations to show. For each permission requested by apps on the user's phone, the PPA applies the classifier trained with the profiles (see Section 4.3.2) to generate an allow/deny decision for the user. The PPA will then display a list of recommended restrictive permission changes to the user.

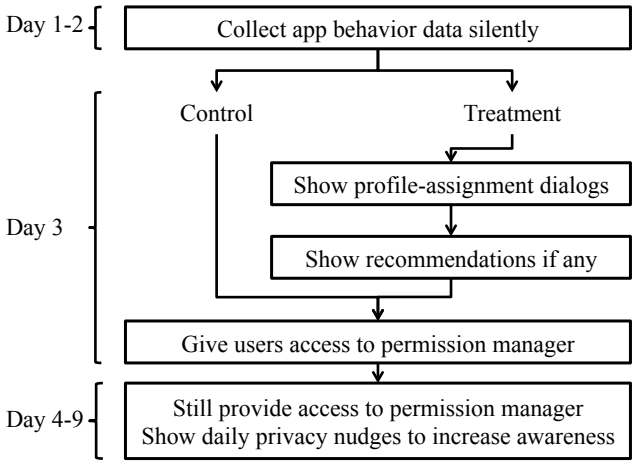


Figure 5: Overview of the study protocol for the two conditions.

Recommendations are grouped by permission (e.g., Calendar, Location); these groups can be expanded to view individual apps, as shown in Figure 4. For each app, clicking the question mark reveals an explanation for this specific recommendation, referencing the user's responses to the profile assignment questions. For instance, in Figure 4 the explanation for denying Snapchat location access is shown. The user can review and adjust recommendation settings. With toggle buttons users can selectively "allow" specific permissions the PPA suggested to deny. The user can accept all shown recommendations, accept some of them by making selective changes, or reject all recommendations.

Thus, based on the privacy profiles generated from real users' privacy settings, our personalized privacy assistant can assign a new user to one of those profiles based on their responses to the profile-assignment dialog. Once a user has been assigned to a profile, we generate recommendations about which permissions a user may want to restrict, personalized to the user's installed apps, by using a classifier with input of the user's profile and the apps' characteristics, such as its category and the purpose of permission requests.

6. FIELD STUDY: EVALUATING THE PPA

We conducted another field study with a second group of Android users with rooted devices to evaluate the effectiveness of our privacy profiles in the context of our PPA. In this study, we collected empirical data on how participants interacted with our PPA app and how they modified their permission settings. The study was conducted as a between-subjects experiment with two conditions: (a) the treatment condition in which participants interacted with the PPA, including profile assignment and recommendations; and (b) a control condition without profile-based support. Participants in both conditions had access to our enhanced permission manager and received privacy nudges.

6.1 Study Procedure

We wanted to evaluate the effectiveness of the profile-based PPA with participants from the same population the privacy profiles were based on. Hence, we followed the same recruitment approach as in the data collection study. We extended the screening survey to exclude individuals with prior experience using other Android permission or privacy managers. We also excluded any participants from our first study. After qualifying for the study, the newly-recruited participants received a user id and instructions for installing the study client.

Our study protocol is summarized in Figure 5. During day 1 and 2 of the study, the PPA silently collected permission access frequency statistics for installed apps. Participants did not have access to the permission manager at that time.

On the third day, the PPA initiated a dialog with participants. In the treatment condition, the app showed an introduction screen, and then initiated the profile assignment dialog, in which participants were asked up to five questions about their privacy preferences, as described in Section 5.1. Users were assigned to a profile and personalized recommendations were generated, as described in Section 5.2. If recommendations could be made, the recommendation screen was shown, and if the PPA did not recommend any changes (i.e., the user was assigned to profile 3), the user was presented with a message saying that it was recommended to keep the current permission settings. The user could review the recommended permission changes and make adjustments as needed. After accepting all, some, or none of the recommendations, participants were asked to rate how comfortable they were with the recommendations on a 7-point Likert scale, followed by a question on why they accepted all, some, or none of the recommendations. After the recommendations and follow-up questions, the PPA opened our permission manager to allow participants to further revise their permission settings.

In the control condition, the app only showed an introduction screen explaining that users could now change their settings, followed by opening our permission manager. This way, the control and treatment conditions were identical in all aspects, except for the omission of the profile assignment dialog and permission recommendations in the control condition.

Starting on day 4, participants in both conditions started receiving one privacy nudge per day for six days, following exactly the same approach as in the first field study. The goal was to get users to reflect on their privacy settings and thus evaluate whether the profiles match their preferences or if they make additional restrictive changes or re-allow any permissions that were restricted based on recommendations. During this phase, we used probabilistic experience sampling (ESM) with single-question dialogs in order to better understand why they denied or allowed permissions, or closed the permission manager without making changes. ESM enabled us to elicit responses from a wider range of participants than would typically agree to participate in exit interviews. ESM dialogs were always consistent with a participant's prior action (e.g., denying permissions). They were shown with 0.66 probability after a user action, to avoid overwhelming users with too many additional dialogs.

At the end of the study, participants were asked to complete an exit survey, which focused on their experience with the profile assignment dialog, perception of the received recommendations, and utility of the additional nudges. After completing the survey, participants were issued a $15 gift certificate. The study received IRB approval.

6.2 Results

We received valid screening survey responses from 138 participants. We excluded 4 participants who had participated in the first study and 3 participants who had prior experience with another app privacy manager. Of 131 initial participants, 72 successfully completed the study (49 treatment, 23 control). Participants were randomly assigned to the two conditions in a 2:1 ratio, as the first study suggested that many participants may have permissive privacy attitudes, in which case they may be assigned to profile 3 (most permis-

Table 1: Privacy protective measures of our study populations compared to the general population. Questions and general population results are based on a Pew survey [36].

Population	Pew Survey	Data Coll. Study	PPA Field Study
Used a temporary username or email address	30.86%	90.00%	92.75%
Added a privacy-enhancing browser plugin (e.g., DoNotTrackMe, Privacy Badger)	11.11%	67.09%	57.35%
Given inaccurate or misleading information about oneself	28.57%	83.75%	78.79%
Set browsers to disable or turn off cookies	44.16%	61.54%	63.24%
Used a service that allows to browse the Web anonymously (e.g., proxy, Tor, or VPN)	11.84%	81.01%	83.82%
Decided not to use a website because it asked for real name	29.49%	66.67%	54.84%
Used a public computer to browse anonymously	15.00%	49.35%	44.92%
Used a search engine that doesn't keep track of search history	22.39%	71.25%	63.64%

sive) and thus would not receive restrictive recommendations and, hence, would not interact with the recommendation screen (shown on the right in Figure 4). Thus, we increased the number of treatment participants to account for these considerations.

6.2.1 *Demographics*

Our sample population was recruited from the same population as for the data collection study and exhibited similar characteristics. Most participants were male (66 male, 5 female, 1 did not disclose) and originated from North America (56, 52 U.S.), Europe (7), South America (3) and Asia (2). Among them, 5 had graduate, 17 Bachelor, and 4 Associates degrees; 23 attended some college, 23 had a high school degree or lower. Commonly reported occupations were student (37), computer engineer or IT professional (12), engineer in other fields (6), service (5) and unemployed (3). Participants in this study also exhibited high privacy concerns (IUIPC [30]): control (mean 6.33, median 6, min 4, max 7), awareness (mean 6.67, median 7, min 5, max 7), and collection (mean 6, median 7, min 2.33, max 7). The participants' measures to protect their online privacy compared to the general online population [36] are shown in Table 1.

6.2.2 *Effectiveness of recommendations*

In the treatment group, the number of received recommendations depended on the privacy profile participants were assigned to and their installed apps. Of the 49 participants in the treatment group, 22 were recommended to keep their current settings. Among them 21 answered "YES" (allow) to most profile assignment questions and got assigned to Profile 3, the most permissive profile. Another participant was assigned to Profile 2 but did not have any of the apps installed that were denied in the assigned privacy profile.

Majority of recommendations were accepted. The 27 participants who received recommendations to deny certain permissions accepted 196 out of 249 individual app recommendations provided (78.7%). Of the 27 participants, 15 accepted all recommendations (they were from profile 1 (4 of them), 2(3), 3(6) and 7(2)), 9 accepted some (they were from profile 1(2), 2(2), 5(3) and 7(2)), and 3 accepted none (all from profile 3; they were shown only one recommendation). Figure 6 shows the number of accepted and rejected recommendations for each of these participants.

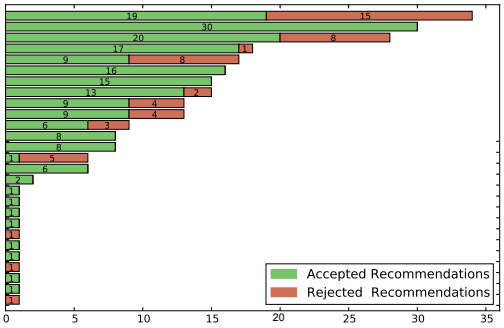


Figure 6: The numbers of recommendations accepted or rejected by participants receiving them. Overall, users accept 78.7% of all recommendations.

The 15 participants that accepted all recommendations primarily stated that they did so because the recommendations matched their preferences (11) or that they trusted the PPA (8). Note that participants could provide multiple reasons. The 3 participants that accepted no recommendations stated that it would have restricted app features (3) or broken app functionality (1), or that the recommendations did not reflect their preferences (2). The 9 participants who accepted some recommendations also stated restricted (6) or broken (4) app functionality as a reason for non-acceptance; 4 stated the recommendations did not reflect their preference, while only 1 responded that they did not like that the PPA wanted to change so many settings automatically.

Participants kept most of the accepted recommendations. During the remaining six days of the study after the recommendation dialog (days 4-9), we showed daily privacy nudges to remind users of actual app permission accesses to increase their awareness and engagement. However, only 10 of the previously accepted recommended permission restrictions (5.10% of all accepted recommendations) were re-allowed. This indicates that the privacy choices made based on the recommendations tended to be accurate, and hence the recommendations were effective (high precision).

Recommendations helped users converge more quickly on settings. The average numbers of permissions changed by participants per day of the study are shown in Figure 7. Among the 383 permission settings changes made by the treatment group, the participants made 316 (82.51%) of them during day 3, which is the day when they received profile-based recommendations and the first day when they had access to the permission manager. In contrast, the control group only made 68.42% (104 of 152) of their permission settings on day 3. The difference of the treatment and the control condition has significant effect on whether participants made changes on day 3 (logistic regression with user ids, Odds Ratio=1.72, StdErr.=0.36, z=2.56, p=0.010).

On days 4–9, the treatment group made 67 additional changes to permissions settings (per participant mean 1.39, SD 2.03), and the control group 48 (per participant mean 2.09, SD 2.63). The difference between conditions was not significant. We have 43 respective ESM responses from the treatment group and 23 from the control group. Participants gave the following reasons for making restrictive changes: "I don't use the app's features that require this permission" (treatment: 10, control: 6), "I don't want this app to use this permission" (21, 18), "The app doesn't need this permission to function" (16, 11), and "Don't know" (4, 0). This suggests that

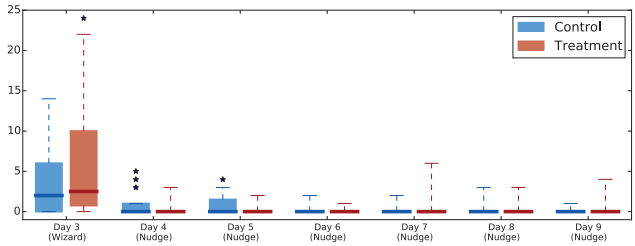


Figure 7: Number of permission changes in the control and treatment groups on the different days of the study. On day 3, the treatment group got recommendations; and both groups were given access to the permission manager.

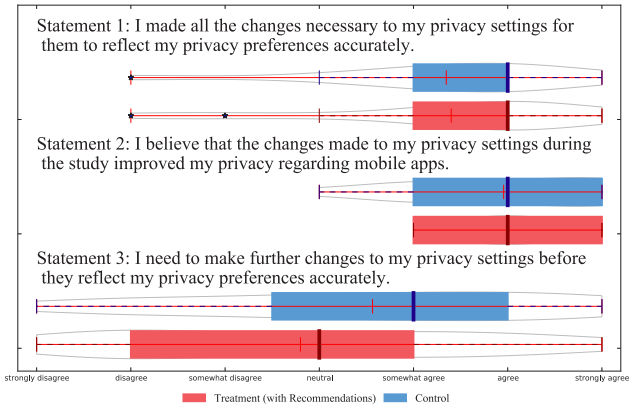


Figure 8: Participants' responses about their privacy settings in the exit questionnaire. Participants who received recommendations felt slightly less of a need to make further changes to their settings.

reasons for restricting permissions were similar across conditions, but the control group had to make more overall changes to arrive at satisfactory settings, whereas the recommendations provided in the treatment group were effective at reducing configuration effort for participants.

In both conditions, few permissions were restricted and later re-allowed (treatment: 18, mean .62, SD 1.37; control: 11, mean .48, SD .73), with no significant difference between conditions (Mann-Whitney U: U=548.5, z=0.1751, p=0.8572). Participants gave the following reasons for re-allowing: "I want to use a feature of the app that requires this permission" (treatment: 3, control: 1), "I am OK with this app using this permission" (4, 1), "The app didn't work as expected when access was restricted" (2, 1), and "Don't know" (0, 1).

Most participants remain in the same profile. We collected the participants' app permission settings at the end of the study and compared them to their responses in the profile-assignment dialogs. For this purpose, we re-ran the profile assignment process with their final permission settings to check their assigned profile, and then compare the two assignments for each participant. Of the 49 treatment group participants, 35 (71.43%) remained in the same privacy profile they were assigned to initially. For the other 14 participants (28.57%), their permission settings changes during the study resulted in a different profile being a better fit for them. Two participants switched from profile 1 to profile 2, which generally allows Location access but denies Call Log access. One participant switched from profile 5 to profile 6, which allowed Camera access more. One switched from Profile 7 to Profile 1, loosening the restrictions on Social apps. The remaining 10 were re-assigned to Profile 3, which is the most permissive one. A likely explanation is that participants' preferences are more restrictive, but that the lack of ability to control for which purposes permissions are granted forced them to be more permissive than desired, i.e., they lack the capabilities to regulate privacy as desired.

Participants are comfortable with provided recommendations. We also collected participants' self-reported comfort with the recommendations and the privacy settings they made during the study. Directly after they accepted recommendations, we asked them to rate their comfort level with the received recommendations on a 7-point Likert scale. Participants felt very comfortable with the provided recommendations (median 6, mode 7, min 3, max 7).

In the exit survey, we asked participants whether they felt that their permission settings changes during the study had improved their privacy, whether they made all necessary changes, and whether they felt more settings changes were needed. The results are shown in Figure 8. We did not find significant differences between the control group and the treatment group (n.s., Mann-Whitney U tests). Participants in both groups felt that their privacy had improved and that they made all the changes necessary for their privacy settings to accurately reflect their privacy preferences. We also did not find significant differences in participants' feelings of a need to make further changes before the settings would reflect their preferences.

6.2.3 *Usability of the personalized privacy assistant*

To evaluate the PPA's usability, we asked Likert-scale and open-response questions to learn what participants found useful or problematic about the PPA, and how it could be improved. We further asked them about the usefulness of the provided recommendations.

Permission manager is useful to monitor apps. Participants in both conditions stated that they especially liked the ability to monitor apps with our enhanced privacy manager (22 treatment, 12 control). That the PPA was helpful in monitoring apps was also confirmed by treatment group participants when asked about the additional nudges (16). Participants also noted the app's general usability (20 treatment, 11 control).

Nudge timing and delivery is important. When asked about what they liked the least, participants from both conditions identified timing of the nudges as an issue (18 treatment, 13 control). Asked how we could improve the PPA, participants from both groups suggested to turn the nudge into an Android notification (9 treatment, 7 control). Treatment participants also indicated that they would have liked more configuration options (7), mainly to influence the timing of nudges. Note that for study purposes, we purposefully displayed the nudge as a modal dialog to force explicit interaction with the nudge. Finally, it should be stressed that the nudges are not an essential component of the PPA evaluated in this study. They were introduced as part of our empirical protocol to evaluate the stability of settings adopted by participants based on the PPA's recommendations.

Recommendations are helpful. Of the 49 treatment participants, 27 were shown recommendations, of whom 24 completed the exit survey. Most participants found the recommendations useful (median 5.5, mode 6, min 2, max 7). This was corroborated by free text answers where 13 responses stated that the recommendations provided useful configuration support (11) and decision support (3). P20 stated: "It made what would have taken 10-20 clicks through menus looking to change these settings done in one click." and P10 stated: "It provides you with recommendations using your prefer-

ences so you can quickly change the settings without have to do much yourself." P4 and P38 found recommendations useful, but would have preferred to set permissions manually. Four participants found recommendations less useful (3) or useless (1), stating that they prefer to manage settings themselves (1) or that some recommendations would have impaired app functionality (3). Overall, this indicates that recommendations were mostly useful, but also points at the issue that users are forced to make trade-offs when apps crash without permission access. In addition, permissions are currently binary choices: either an app has access to a resource for any purpose or not at all, restricting permissions for specific purposes is not possible in today's commercial mobile platforms.

Bulk recommendations are useful. We also asked questions in the exit survey to assess the usability and utility of the different parts of the recommendation screen, such as the timing and amount of information displayed. Participants found that it was useful that all recommendations were listed on one screen (median 6, mode 6, min 3, max 7). This was corroborated by participants disagreeing that it was annoying that they had to click the categories to see details (median 2, mode 2, min 1, max 5). Participants reported their preference for seeing recommendations right after answering each question (median 4, mode 5, min 1, max 6). Participants reported that they somewhat preferred to see the PPA directly after installation (median 5, mode 5, min 3, max 7).

Question dialogs were usable. Question dialogs were shown to all treatment participants. We asked them to rate on a 7-point Likert scale how easy or difficult the three question types were to answer. All three question types were reported to be easy to answer (permission only: median 7, mode 7, min 3, max 7; permission/purpose: median 6, mode 6, min 3, max 7; permission/category: median 6, mode 7, min 4, max 7). Participants also reported that the app list (median 6, mode 7, min 4, max 7) and access frequency (median 6, mode 6, min 1, max 7) were useful. The app list helped create awareness of how installed apps used permissions (29) and helped to identify apps with undesired permissions (17). Access frequency also helped improve awareness (36) and was mentioned by 6 participants as an important decision factor.

7. DISCUSSION

Our results suggest that personalized privacy assistants can indeed help users better manage their mobile app permission settings. They provide evidence based on deployment with actual users that profile-based recommendations can help users configure their mobile app permissions. Below, we first discuss limitations of our work, followed by insights gained about the development and interaction design of personalized privacy assistants.

7.1 Limitations

Because manipulating people's mobile app permission settings requires root access, the target population available for recruitment for this study was limited. As a result, the sample populations in both filed studies skew young, male, tech-savvy, and privacy-conscious. Accordingly, one might expect the privacy settings and permission profiles obtained for this population to be more conservative (namely, more restrictive) than those of the general population. But one cannot be entirely sure: rooted users are also more technically sophisticated and possibly more daring. In fact, a relatively large number of our participants selected rather permissive privacy settings. It is important to understand that the objective of this work was not to identify the "ultimate" privacy profiles for the general population. Rather our main objective was to evaluate (1) a practical approach for collecting permission data and learning profiles, and (2) a method for using the resulting profiles in the context of personalized privacy assistants. The work presented herein is particularly important because it relies on the collection of permission data and the validation of personalized privacy assistants in field studies, in which participants used their regular phones in their daily activities. A similar study could be conducted with other target populations, including the general population, given the ability to reliably collect and manage privacy settings on non-rooted phones. Developers who have access to the necessary functionality (whether on smartphones or in other contexts, such as a web browser or a permission manager for a social network) could leverage our approach to learn profiles and provide their users with personalized privacy recommendations. Mobile platform providers, such as Google, Samsung, or Apple, could implement our approach (or provide APIs for researchers and developers) and support functionality similar to the one evaluated in this study.

In contrast to prior work, we learned privacy profiles from a relatively small dataset, which could be viewed as a limitation. We overcame this potential limitation by collecting rich, real-world permission data and aggregating obtained permission settings along three dimensions, namely app category, permissions, and purpose information. Our second field study validates the effectiveness of the learned profiles and recommendations. Three-quarters (78.7%) of the provided recommendations were accepted, and only a small number of recommendations to restrict permissions were later re-allowed (5.1%) – primarily because the restrictive permissions impaired some app functionality, rather than participants having privacy preferences that differed from those in the assigned profiles. Participants further reported high comfort with their privacy settings at the end of the study.

A potential limitation is the relatively short length of our study. It is possible that participants may not have fully converged on stable privacy settings. We believe that the likelihood that this was the case is fairly low because of our use of daily privacy nudges. These nudges were effective at getting participants to review and adjust their permission settings. This approach enabled us to elicit permission settings for a large number of apps (729) and permissions (3,559) in a relatively short time from 84 participants. This data was used to learn privacy profiles and provide participants in the second study with privacy recommendations to support initial configuration. The low number of subsequent permissions changes (see Figure 7) furthers support the notion that PPA users had converged on stable settings by the end of the study. In future work, we plan to explore longitudinal interactions with personalized privacy assistants over longer periods of time and further study continuous privacy decision making processes.

7.2 Privacy Profiles and Recommendations

Our results show the feasibility of learning privacy profiles from a relatively small number of users. These profiles are effective at supporting users in configuring their permission settings and helping them make privacy decisions. In the second field study, which evaluated the profile-based PPA, participants reviewed and accepted 78.7% of our recommendations. Additionally, very few recommended restrictive permission settings were changed back by participants (5.1%). However, some participants restricted additional permissions based on information shown in the privacy nudges and the permission manager. This suggests that our classifier could possibly be tuned to provide more aggressive recommendations. It is also likely that having access to a larger corpus of permission settings would enable us to build profiles with higher predictive power. Finally, the ability to directly adjust recommended settings and the

option to make additional changes in the permission manager was perceived as useful by most participants, as it helped them reflect on their privacy settings and bootstrap the configuration.

Our recommendations could further be improved with enhanced filtering techniques to exclude core system apps and services, as well as apps that crash when restricted. App crashes were sometimes reported as a reason for re-allowing permissions. The introduction of a selective permission model in Android 6.0 suggests that in the future most apps will likely continue to work properly even when requested permissions are denied, as is already the case in iOS, since app developers will adapt and add exception handling for denied permissions.

A general issue that emerged was a conflict between restrictive privacy preferences and permissions required by an app to properly function. This happens when apps require permissions for multiple purposes (e.g., both to support their core functionality and to support advertising). Multiple participants reported that they would have liked to deny certain permissions (e.g., location) for specific purposes (e.g., tracking and profiling), but that they could not do so, as it would have broken essential features of the application. This suggests that current permission models would benefit from allowing users to grant and deny permissions for specific purposes, rather than forcing users to deny or accept the combination of all purposes. While iOS and Android 6.0 support developer-specified purposes in permission requests [44, 46], once access is granted, apps can currently use the corresponding resource for any purpose. The current permission model also fails for system services, such as Google Play Services, that provide resource access to multiple apps (e.g., location). Because it is unclear how many apps depend on sensitive resources provided by a service like Google Play Services, it is effectively impossible for users to make meaningful decisions about granting or denying Google Play access to a permission such as location. A substantial challenge in mobile computing and other domains will be to shift permission models from resource-centric fine-grained access control (e.g., multiple permissions to read, write SMS) to purpose-centric controls that better align with users' privacy decision making. While these finer-grained models could increase user burden, our research suggests that they may in fact lend themselves to the learning of more powerful predictive models, which in turn could actually help reduce user burden by providing a larger number of more accurate recommendations.

For future personalized privacy assistants, we envision to assist users with privacy monitoring, configuration, and decision support beyond initial permission configuration. Settings recommendations could be provided when installing new apps or as part of just-in-time permission requests. Ultimately, privacy assistants should further adapt to users by learning their privacy preferences over time, for instance by engaging with them in a continuous, yet unobtrusive, dialog. Micro-interactions initiated at opportune times and tailored to the user's context [41, 42] could help increase the usability of privacy nudges by better integrating them into a user's interaction flow. This also requires enhancing machine learning techniques to appropriately account for the uncertainty, contextual nature, and malleability of privacy preferences [4].

7.3 Designing Personalized Privacy Assistants

Our two field studies provided extensive insights on how users interact with different mobile privacy tools: our enhanced permission manager, privacy nudge interventions, privacy profile assignment dialogs, and profile-based recommendations. Our results show that all these tools play important, yet different, roles in supporting users with privacy configuration and decision making, and should therefore be taken into consideration when designing personalized privacy assistants and the associated user experience.

Profile assignment is an integral part of our personalized privacy assistant. We use a small number of privacy preference questions to assign users to a profile and provide them with privacy recommendations personalized to their installed apps. We found that participants felt confident answering all three types of questions asked. Contextualizing the questions with apps that would be affected by the user's response was perceived as useful, and access frequency also helped most users. In addition to using access frequency of the installed apps, we plan to explore the utility of creating statistical models of how often specific apps access certain resources in order to be able to provide permission recommendations without a training phase. This information could in addition be added to an app's app store information, enabling users to use frequency in decision making even before installing an app.

Privacy recommendations introduce a degree of automation to privacy configuration. Automation can potentially impact technology acceptance [33]. Our results indicate that we have achieved a good balance, given that participants reviewed and edited recommendations while reporting high levels of comfort and usability. In future work, we plan to further investigate the impact of different levels of automation on the acceptance of personalized privacy assistants.

Our results show that the enhanced privacy manager – including both information on permission access frequency and purpose – helped participants monitor app behavior and manage their privacy settings effectively. A further improvement, motivated by participants' responses, would be to include more information about how privacy and app functionality would be affected by allowing or denying specific permissions. Furthermore, many participants mentioned the nudge's timing and modality as an issue. However, the use of modal dialogs was a conscious choice to force interaction with the nudge messages in our study. In the public release version of our PPA, we implemented nudges as standard Android notifications to make them less obtrusive.

While our results and insights pertain primarily to mobile interaction, we expect that personalized privacy assistant approaches can also be applied to support privacy decision making in other domains where privacy configuration or awareness is an issue. For instance, in the context of websites, where privacy policies are often difficult to understand, or the Internet of Things (IoT), where secondary channels will have to be utilized for privacy management, because most IoT devices have small or no screens [41].

8. CONCLUSION

In this paper, we demonstrated how users can benefit from a personalized privacy assistant that provides them with recommendations for privacy configuration. Our personalized privacy assistant is based on privacy profiles learned from real-world permission settings. Our proposed approach is practical and can learn representative privacy profiles even from a relatively small number of users (n=84). We evaluated the effectiveness of the privacy profiles by conducting a field study (n=72), in which we deployed our personalized privacy assistant on participants' own smartphones (rooted Android devices). Our results show that 78.7% of recommendations were accepted by users and that only 5.1% of settings were changed back during the study. Overall, the assistant led to more restrictive permission changes without sacrificing users' comfort with these settings.

9. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under grants CNS-1012763 and SBE-1513957, as well as by DARPA and the Air Force Research Laboratory, under agreement number FA8750-15-2-0277. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright notation thereon. Additional funding has also been provided by Google through a Google Faculty Research Award and the Google Web of Things Expedition and in part through a grand from the CMU-Yahoo! InMind project, as well as by the Carlsberg Foundation. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, the Air Force Research Laboratory, the National Science Foundation, the U.S. Government, Google or Yahoo!

The authors would like to thank the anonymous reviewers and our shepherd Marian Harbach for their constructive feedback.

10. REFERENCES

[1] Android Flashlight App Developer Settles FTC Charges It Deceived Consumers. `https://goo.gl/Zf18jI`, 2013. Accessed: 2016-02-01.

[2] PrivacyGrade: Grading The Privacy Of Smartphone Apps. `http://privacygrade.org`, 2015. Accessed: 2016-02-01.

[3] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup. Scalable tensor factorizations with missing data. In *SDM*, pages 701–712. SIAM, 2010.

[4] A. Acquisti, L. Brandimarte, and G. Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, Jan. 2015.

[5] Y. Agarwal and M. Hall. ProtectMyPrivacy: detecting and mitigating privacy leaks on iOS devices using crowdsourcing. In *Proc. MobiSys*, 2013.

[6] H. Almuhimedi, F. Schaub, N. Sadeh, I. Adjerid, A. Acquisti, J. Gluck, L. Cranor, and Y. Agarwal. Your location has been shared 5,398 times! a field study on mobile app privacy nudging. In *Proc. CHI*. ACM, 2015.

[7] arstechnica. Android M Dev Preview delivers permission controls, fingerprint API, and more. `http://goo.gl/NdmOx1`, 2015. Accessed:2016-02-01.

[8] R. Balebako, J. Jung, W. Lu, L. F. Cranor, and C. Nguyen. Little brothers watching you: Raising awareness of data leaks on smartphones. In *Proc. SOUPS*, 2013.

[9] E. K. Choe, J. Jung, B. Lee, and K. Fisher. Nudging people away from privacy-invasive mobile apps through visual framing. In *Proc. INTERACT*, 2013.

[10] K. Connelly, A. Khalil, and Y. Liu. Do i do what i say?: Observed versus stated privacy preferences. In *Proc. INTERACT 2007*, pages 620–623. Springer, 2007.

[11] J. Cranshaw, J. Mugan, and N. Sadeh. User-controllable learning of location privacy policies with gaussian mixture models. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[12] EFF. Awesome Privacy Tools in Android 4.3+. `https://www.eff.org/deeplinks/2013/11/awesome-privacy-features-android-43`, 2013. Accessed: 2015-2-17.

[13] W. Enck, P. Gilbert, B.-G. Chun, L. P. Cox, J. Jung, P. McDaniel, and A. N. Sheth. Taintdroid: an information flow tracking system for real-time privacy monitoring on smartphones. *Comm. ACM*, 2010.

[14] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[15] L. Fang and K. LeFevre. Privacy wizards for social networking sites. In *Proc. WWW '10*. ACM, 2010.

[16] A. Felt, E. Ha, S. Egelman, A. Haney, E. Chin, and D. Wagner. Android Permissions: User Attention, Comprehension, and Behavior. In *Proc. SOUPS '12*, 2012.

[17] A. P. Felt, E. Chin, S. Hanna, D. Song, and D. Wagner. Android permissions demystified. In *Proc. CCS '11*, pages 627–638. ACM, 2011.

[18] A. P. Felt, S. Egelman, and D. Wagner. I've got 99 problems, but vibration ain't one: a survey of smartphone users' concerns. In *Proc. SPSM*, 2012.

[19] D. Fisher, L. Dorner, and D. Wagner. Short paper: location privacy: user behavior in the field. In *Proc. SPSM '12*, pages 51–56. ACM, 2012.

[20] H. Fu, Y. Yang, N. Shingte, J. Lindqvist, and M. Gruteser. A field study of run-time location access disclosures on android smartphones. In *Proc. USEC*, 2014.

[21] M. Harbach, M. Hettig, S. Weber, and M. Smith. Using personal examples to improve risk communication for security & privacy decisions. In *Proc. CHI*, 2014.

[22] Q. Ismail, T. Ahmed, A. Kapadia, and M. K. Reiter. Crowdsourced exploration of security configurations. In *Proc. CHI '15*, pages 467–476. ACM, 2015.

[23] P. G. Kelley, S. Consolvo, L. F. Cranor, J. Jung, N. Sadeh, and D. Wetherall. A conundrum of permissions: installing applications on an android smartphone. In *Proc. FC '12*. Springer, 2012.

[24] P. G. Kelley, L. F. Cranor, and N. Sadeh. Privacy as part of the app decision-making process. In *Proc. CHI*, pages 3393–3402. ACM, 2013.

[25] J. King. How come i'm allowing strangers to go through my phone? smartphones and privacy expectations. In *Proc. SOUPS*, 2013.

[26] B. P. Knijnenburg. Information disclosure profiles for segmentation and recommendation. In *SOUPS2014 Workshop on Privacy Personas and Segmentation*, 2014.

[27] J. Lin, S. Amini, J. I. Hong, N. Sadeh, J. Lindqvist, and J. Zhang. Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing. In *Proc. UbiComp*, 2012.

[28] J. Lin, B. Liu, N. Sadeh, and J. I. Hong. Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings. In *Proc. SOUPS*, 2014.

[29] B. Liu, J. Lin, and N. Sadeh. Reconciling mobile app privacy and usability on smartphones: could user privacy profiles help? In *Proc. WWW '14*. ACM, 2014.

[30] N. K. Malhotra, S. S. Kim, and J. Agarwal. Internet users' information privacy concerns (iuipc): The construct, the scale, and a causal model. *Information Systems Research*, 15(4):336–355, 2004.

[31] P. A. Norberg, D. R. Horne, and D. A. Horne. The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs*, 41(1):100–126, 2007.

[32] L. Palen and P. Dourish. Unpacking "privacy" for a networked world. In *Proc. CHI '03*, pages 129–136. ACM, 2003.

[33] R. Parasuraman, T. Sheridan, and C. D. Wickens. A model

for types and levels of human interaction with automation. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 30(3):286–297, May 2000.

[34] Path official blog. We are sorry. http://blog.path.com/post/17274932484/we-are-sorry, 2012. Accessed:2016-02-01.

[35] A. Paturi, P. G. Kelley, and S. Mazumdar. Introducing privacy threats from ad libraries to android users through privacy granules. In *Proc. USEC '15*. Internet Society, 2015.

[36] Pew Research Center. Internet project/GFK privacy panel. http://www.pewinternet.org/files/2015/05/Privacy-and-Security-Attitudes-5.19.15_Topline_FINAL.pdf, 2014. Accessed:2016-02-01.

[37] Pew Research Center. An Analysis of Android App Permissions. http://www.pewinternet.org/2015/11/10/an-analysis-of-android-app-permissions/, 2015. Accessed:2016-02-01.

[38] J. R. Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[39] B. Rashidi, C. Fung, and T. Vu. Dude, ask the experts!: Android resource access permission recommendation with recdroid. In *Integrated Network Management (IM), 2015 IFIP/IEEE International Symposium on*, pages 296–304, May 2015.

[40] R. Ravichandran, M. Benisch, P. G. Kelley, and N. M. Sadeh. Capturing social networking privacy preferences. In *Proc. PET '09*, pages 1–18. Springer, 2009.

[41] F. Schaub, R. Balebako, A. L. Durity, and L. F. Cranor. A design space for effective privacy notices. In *Proc. SOUPS '15*, pages 1–17, Ottawa, July 2015. USENIX Association.

[42] F. Schaub, B. Konings, and M. Weber. Context-adaptive privacy: Leveraging context awareness to support privacy decision making. *Pervasive Computing, IEEE*, 14(1):34–43, Jan 2015.

[43] Scikit-Learn. Scikit-learn manual. http://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html. Accessed:2016-02-01.

[44] F. Shih, I. Liccardi, and D. J. Weitzner. Privacy tipping points in smartphones privacy preferences. In *Proc. CHI*. ACM, 2015.

[45] I. Shklovski, S. D. Mainwaring, H. H. Skúladóttir, and H. Borgthorsson. Leakiness and creepiness in app space: Perceptions of privacy and mobile app use. In *Proc. CHI*, 2014.

[46] J. Tan, K. Nguyen, M. Theodorides, H. Negrón-Arroyo, C. Thompson, S. Egelman, and D. Wagner. The effect of developer-specified explanations for permission requests on smartphone user behavior. In *Proc. CHI*. ACM, 2014.

[47] The Guardian. Uber faces FTC complaint over plan to track customers' locations and contacts.

[48] The Next Web. Android users have an average of 95 apps installed on their phones, according to Yahoo Aviate data. http://thenextweb.com/apps/2014/08/26/android-users-average-95-apps-installed-phones-according-yahoo-aviate-data/#gref, 2014. Accessed:2016-02-01.

[49] S. Thurm and Y. I. Kane. Your apps are watching you. http://www.wsj.com/articles/SB10001424052748704368004576027751867039730, 2010. Accessed: 2016-02-01.

[50] N. Wang, B. Zhang, B. Liu, and H. Jin. Investigating effects of control and ads awareness on android users' privacy behaviors and perceptions. In *Proc. MobileHCI '15*. ACM, 2015.

[51] S. Wilson, J. Cranshaw, N. Sadeh, A. Acquisti, L. F. Cranor, J. Springfield, S. Y. Jeong, and A. Balasubramanian. Privacy manipulation and acclimation in a location sharing application. In *Proc. UbiComp '13*, pages 549–558. ACM, 2013.

[52] P. Wisniewski, B. P. Knijnenburg, and H. Richter Lipford. Profiling facebook users' privacy behaviors. In *SOUPS2014 Workshop on Privacy Personas and Segmentation*, 2014.

[53] J. Xie, B. P. Knijnenburg, and H. Jin. Location sharing privacy preference: Analysis and personalized recommendation. In *Proc. IUI '14*, pages 189–198. ACM, 2014.

[54] Y. Zhao, J. Ye, and T. Henderson. Privacy-aware location privacy preference recommendations. In *Proc. Mobiquitous '14*, 2014.

APPENDIX

A. LOGISTIC REGRESSION RESULTS

Results of the random effect logistic regression are shown in Table 2.

Table 2: Random effect logistic regression on users' allow/deny decisions grouped by users (Likelihood ratio test of $\rho = 0$: $\bar{\chi}^2 = 338.10$, $P >= \bar{\chi}^2 : 0.000$).

Factors		Odds Ratio	StdErr	z	P>|z|
Age		1.024816	.0619711	0.41	0.685
Gender		.6941319	.6480886	-0.39	0.696
Education	Associate	6.351436	6.536207	1.80	0.072
	Bachelor	.3252345	.2102106	-1.74	0.082
	Graduate	2.265247	2.258762	0.82	0.412
	High School	.9914089	.5819914	-0.01	0.988
	No High School	1			
	Some College	1			
Occupation	Administrative	5.442226	8.371201	1.10	0.271
	Art/Writing/Journalism	1			
	Business/Management/Finance	1			
	Computer/IT	1.364362	1.553644	0.27	0.785
	Decline to answer	5.775118	6.803399	1.49	0.137
	Education	.0920523	.1597209	-1.37	0.169
	Engineer in other fields	16.96705	31.93771	1.50	0.133
	Homemaker	1.134727	3.123314	0.05	0.963
	Legal	.1008037	.1688665	-1.37	0.171
	Medical	.633246	.8901533	-0.33	0.745
	Other	1.804592	2.601707	0.41	0.682
	Scientist	1.903118	2.983608	0.41	0.681
	Service	1.962722	2.268031	0.58	0.560
	Skilled labor	.7758243	1.22502	-0.16	0.872
	Student	2.534309	2.248981	1.05	0.295
	Unemployed	1			
IUIPC Scale	Control	.6704036	.3212597	-0.83	0.404
	Awareness	.6779195	.381246	-0.69	0.489
	Collection	1.810677	.4923613	2.18	**0.029**
App Category	Books & Reference	12.19531	9.009827	3.39	**0.001**
	Business	11.00032	6.011878	4.39	**0.000**
	Communication	4.464244	1.614809	4.14	**0.000**
	Education	5.988742	6.630343	1.62	0.106
	Entertainment	7.792989	3.563787	4.49	**0.000**
	Finance	3.490802	1.561327	2.80	**0.005**
	Game	8.974919	4.578022	4.30	**0.000**
	Health & Fitness	4.637063	2.497553	2.85	**0.004**
	Libraries & Demo	2.107152	2.378477	0.66	0.509
	Lifestyle	4.278822	1.932977	3.22	**0.001**
	Media & Video	5.627252	3.56555	2.73	**0.006**
	Medical	1			
	Music & Audio	14.15537	7.885298	4.76	**0.000**
	News & Magazines	6.177335	3.068304	3.67	**0.000**
	Personalization	.6819545	.5712842	-0.46	0.648
	Photography	1.099871	.8050647	0.13	0.897
	Productivity	2.107637	.8318742	1.89	0.059
	Shopping	4.381211	1.813481	3.57	**0.000**
	Social	7.208478	2.76813	5.14	**0.000**
	Sports	25.32193	17.04635	4.80	**0.000**
	Tools	3.562823	1.293064	3.50	**0.000**
	Transportation	.8090313	.530982	-0.32	0.747
	Travel & Local	1			
	Weather	1			
Permission	Location	2.620968	1.041181	2.43	**0.015**
	Contacts	.7826907	.3259032	-0.59	0.556
	Messages	3.870752	1.591046	3.29	**0.001**
	Call Log	2.39916	1.127688	1.86	0.063
	Camera	.1410928	.0698829	-3.95	**0.000**
	Calendar	1			
log(Frequency+1)		.9541353	.0317826	-1.41	0.159
Purpose	App functionality	1.296318	.2925215	1.15	0.250
	Targeted advertising	1.235337	.5431015	0.48	0.631
	Consumer tracking & profiling	1.123383	.6212463	0.21	0.833
	Social networking services	.2956021	.3464561	-1.04	0.298
(Constant)		.0275754	.0780506	-1.27	0.205
Logged variance of random effect		.7827504	.2309066		
StdEv. of random effect		1.479013	.170757		
ρ (Intraclass correlation)		.3993685	.0553883		

"They Keep Coming Back Like Zombies": Improving Software Updating Interfaces

Arunesh Mathur
amathur@umd.edu

Josefine Engel
jme4yg@virginia.edu

Sonam Sobti
sonam.sobti9@gmail.com

Victoria Chang
vchang7190@gmail.com

Marshini Chetty
marshini@umd.edu

Human–Computer Interaction Lab, College of Information Studies
University of Maryland, College Park
College Park, MD 20742

ABSTRACT

Users often do not install security-related software updates, leaving their devices open to exploitation by attackers. We are beginning to understand what factors affect this software updating behavior but the question of how to improve current software updating interfaces however remains unanswered. In this paper, we begin tackling this question by studying software updating behaviors, designing alternative updating interfaces, and evaluating these designs. We describe a formative study of 30 users' software updating practices, describe the low fidelity prototype we developed to address the issues identified in formative work, and the evaluation of our prototype with 22 users. Our findings suggest that updates interrupt users, users lack sufficient information to decide whether or not to update, and vary in terms of how they want to be notified and provide consent for updates. Based on our study, we make four recommendations to improve desktop updating interfaces and outline socio-technical considerations around software updating that will ultimately affect end-user security.

1. INTRODUCTION

Vulnerabilities in client-side applications that run on user devices are on the rise. Typically, software vendors roll out software updates or "patches" to protect users by fixing these vulnerabilities and making changes to the software—such as adding new features, enhanced performance, or bug fixes. For this reason, the United States (US) government, various security agencies, and security experts advise end-users to download and install updates in a timely fashion to keep their systems secure [43, 32, 12]. However, recent studies have shown that non-expert end-users report delaying updates because they lack awareness on the importance of installing these security patches [28] or possess incorrect mental models of how updating systems work [49]. Moreover, even users identified as "professionals", "software developers", and "security analysts" only install updates about half the time more than non-experts [34]. Despite this evidence that there are factors influencing updating behaviors, the majority of research on software updates focuses on the network and systems aspect of delivering updates to end users [17, 47, 24, 13, 23].

However, a growing number of studies have begun to explore the human side of software updates for Microsoft (MS) Windows users [46, 49] including the reasons why users avoid updates in the first place. While these studies uncover users issues with software updates, they do not answer the question of how to make practical improvements to current software updating interfaces that could enhance security or whether these findings hold for users of other operating systems. To answer these questions, we further examined what prevents users from applying software updates across different operating systems and used this evidence to explore how to improve desktop software updating interfaces.

To achieve this goal, we conducted a three-phased research study. First, we conducted a qualitative formative study of 30 US Internet users' desktop software updating practices to complement previous work that focused solely on MS Windows users [49, 46]. Second, we distilled our findings into the design of a *minimally-intrusive*, *information-rich*, and *user-centric*, low-fidelity prototype of an alternative desktop software updating interface. Third, we conducted a think-aloud study with 22 Mac OS X users to evaluate our designs and draw recommendations to improve desktop software updating interfaces.

We make the following contributions. First, we confirm the findings of previous studies [46] and show these findings also hold for desktop users of operating systems other than MS Windows. Specifically, our findings reveal that users avoid updates that interrupt them, they lack sufficient information to decide whether or not to perform an update, and that users vary on how they want to provide consent and be notified of updates. Second, our study newly identifies additional reasons that users avoid updates such as whether or not users trust the software vendor providing the update, obscure change logs, and unknown installation times. Third, we contribute the design of an alternative updating desktop interface for Mac OS X that addresses these issues.

Finally, based on the positive reaction to our design con-

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22–24, 2016, Denver, Colorado.

cepts in our think-aloud study, we contribute four validated recommendations for improving current updating interfaces for the desktop: personalizing update interfaces, minimizing update interruptions, improving update information, and centralizing update management across a device. We also discuss the socio-technical aspects of the updating process, namely around trust, consent, and control for making changes to in the wild software. We believe enhancing usability through improved desktop updating interfaces and further research to address the socio-technical aspects of software updating will ultimately lead to more secure systems.

2. BACKGROUND AND RELATED WORK

In this section, we explain the software updating process and touch upon the related work on software updates.

2.1 Software Updates and Automation

The usable security community has long recognized that human beings are the "weakest link" in the security chain [16, 39], attributing many security failures to human factors. Moreover, the community has recognized that most security decision making cannot be fully automated because humans often have to perform a part of the task—such as responding to security warnings (e.g., SSL [42], [3]) and identifying phishing emails [50]. Additionally, automation is often context dependent and limited by failure cases [18]. To compensate for the human element in security systems, Cranor [11] developed a framework to help designers fully consider all the factors to integrate users in the loop for security decisions in various systems. Software updates are no exception.

In fact, software updates typically involve users at various stages of the update process. An update process often varies based on the device type, operating system and application [31, 4], and the degree of automation and user involvement in each step can result in significantly different update experiences. Generally, a software update involves [13, 17]:

1. *Discovering the Update*: Users can either search for updates manually on websites or app stores, or set updating preferences for a specific application or the operating system to automatically notify them when updates become available.
2. *Downloading the Update*: Users can choose to either download available updates manually or set preferences for the system to automatically download them on their behalf.
3. *Installing the Update*: Users can manually install or have their system automatically install updates. Installation may involve closing applications affected by the update and often, an update is only applied after an application restart or machine reboot.
4. *Using the System Post-Update*: Once applied, updates may notify users that they have completed.

Depending on the degree to which the update system notifies and involves users, software update preferences are often referred to as [49, 17]:

- *Manual*: Users initiate and complete all the steps of the updating process, e.g., software drivers for input and output computer peripherals.
- *Automatic*: The update system automates one or more steps of the updating process such as downloading, installing, and notifying users. Users may have to briefly discontinue using the application to complete the installation or perform a restart of their machine or application, e.g., MS Windows patches and MS Office updates.
- *Silent*: The update system automates the entire update process and in addition, does not notify users explicitly at any step. Typically, in a silent update, the system installs the update without interrupting the user and applies it when users restart or re-open the application. Often, users fail to notice such updates, and lack the provision to disable or prevent them [34], e.g., Google Chrome updates [17].

In terms of reaching users' machines soonest after release, recent studies suggest that silent updating mechanisms may be the most effective in patching machines after an exploit is disclosed when compared to methods requiring a user's consent to download, install, or apply an update [34, 17]. Most software vendors and the US government recommend automatic updates for users to keep their systems secure instead of manual updates for this reason [43, 32, 12]. In our work, we examine user reactions to a low-fidelity prototype that conceptually makes all updates silent.

2.2 Deploying Software Updates

Numerous studies have explored ways to develop and deploy software updates, and compared the effectiveness of different mechanisms. For instance, Duebendorger and Frei studied the effectiveness of silent updates [17], Vojnovic *et al.* studied automatic patches [47], and Gkantsidis and Karagiannis studied the Windows patching system for distributing patches on a planet scale [24]. These studies comment on each patching mechanism's strengths and weaknesses and make suggestions for improving the creation and distribution of patches at scale but with no focus on how users will appropriate these updates. Another set of studies focuses on improving the deployment of patches in large organizations [13, 23]. For instance, Oberheide *et al.* help network administrators infer the impact of patches before deployment [36]. Others have investigated how to improve the deployment of patches via USB drives in regions with sporadic connectivity [9]. While these studies focus on improving the software patches themselves, they do not study the end-users who apply these patches, why they avoid patches, or how to improve patching interfaces as we do in our study.

2.3 User Experience with Software Updates

There is a growing body of work focused on understanding users' general online security behaviors and on user barriers to software updates. For example, Ion *et al.* [28] compared the capability of expert and non-expert users to process security advice. They found that non-experts updated their software less frequently compared to experts, lacked awareness about the effectiveness of software updates, and avoided updates that they felt introduced software bugs. Similarly, Wash and Rader surveyed almost 2000 US Internet users and found only 24% used protective security behaviors such as downloading patches [48]. Other studies show that users often disable or only perform updates on WiFi networks when

Phase	No of Users	Timeline
1. Formative Study	30	Jun '14–Sep '14
2. Prototype Design	–	Oct '14–Feb '15
3. Prototype Evaluation	22	Feb '15–May '15

Table 1: Research Timeline Overview.

they have limited and expensive Internet data plans [7, 29]. These studies provide evidence that users infrequently apply updates and touch on a few barriers in the process but they are not solely focused on users and software updates.

Several researchers have studied users and software updates in more depth. For instance, Fagan *et al.* [21] surveyed 250 users about attitudes toward software updating notifications. They found that users were reluctant to apply updates because they disliked being interrupted by notifications which were often perceived as obscure and unclear. In complementary work, Vaniea *et al.* [46] studied 37 non-expert MS Windows users and found that past negative updating experiences, such as dealing with user interface changes that required re-learning how to use an application, affect future updating behaviors. In another study of the same Windows users, Wash *et al.* [49] found that users' updating behaviors and intentions with their updating preferences are mismatched, often resulting in less secure systems. The authors conclude that there is a tension between automation and control in the updating process which may be difficult to resolve through improved usability alone. These studies focus on understanding users' software updating behaviors but not on how to improve updating interfaces as we do in our study.

Thus far, only two studies have sought to improve updating interfaces. First, Sankarpandian *et al.* [38] developed a desktop graffiti system TALC, which reminded users to install updates by painting their desktop with graffiti when their machines were left un-patched for a long amount of time. These researchers focused more on how to improve the process of gently notifying and nudging users to pay attention to install updates rather than with improving the overall experience of updating. Second, Tian *et al.* [45] developed a novel updating notification that used user generated reviews to help mobile users make privacy conscious decisions about which updates to apply based on what permissions were asked for by the updates. In contrast our study deals not only with notifications but updating as a whole on desktops, where privacy issues manifest differently because users do not explicitly grant permissions to applications.

3. PHASE ONE: FORMATIVE STUDY

We conducted a three-phased research process over the timeline shown in Table 1. In Phase One, we investigated users' current software updating behaviors and preferences through a qualitative interview-based study.

3.1 Method

3.1.1 Procedure

In mid to late 2014, we recruited 30 participants to take part in semi-structured interviews about their overall experience with software updates, including their likes and dislikes about software updates and their current software updating behaviors. We recruited participants through advertisements on university and affiliated mailing lists around the US, and social media (Facebook, Twitter) posts. We focused on finding adult Internet users that used Internet-enabled devices such as a desktop, laptop, tablet, or smartphone since they were likely to encounter software updates frequently. All interviews were conducted over the phone or Skype, audio-taped, and lasted between 45–60 minutes each. Participants were compensated with USD 15 gift cards for their time. The study was approved by the Institutional Review Board (IRB) of our institution.

The interview guide was developed after a survey of the existing literature on software updating and usable security at the time and informed by Cranor's human in the loop framework [17, 11] to ensure we covered all aspects of the software updating process. Cranor's model describes the human factors that affect secure systems, namely: *Communication* (informing the human that an action is necessary), *Communication impediments* (what might prevent the human from taking the action?), *Characteristics of the human receiver* (demographics, intentions, comprehension, and knowledge retention), and finally, *Behavior*. We used the framework to tease out the various human elements in the software updating process we discussed in Section 2. Concretely, we walked the participants through the specifics of the update process—discovering, downloading, and installing updates—and asked them the following questions for their applications and operating systems:

1. How do users learn about updates on their machines? Do they manually seek updates or wait for notifications?
2. Do users feel updates are important to security? What motivates them to either install or avoid updates?
3. How do users navigate the update process and how do they make decisions about security vs non-security related updates?
4. Do updates interrupt users' workflow? How does this affect their behavior?
5. Do users understand software update change logs and more generally, what action updates ask of them?

We also asked participants whether they ever changed, or sought help to change, the default update preferences for their operating systems and applications. In addition, we collected our participants' demographics (age, gender, education, income), their security management practices on their Internet-enabled devices such as installing anti-virus or enabling firewalls, their online security knowledge/actions when downloading software and dealing with suspicious emails, and past experiences with security incidents. The interview guide in its entirety is available in the Appendix.

3.1.2 Analysis

Once the interviews were transcribed, three researchers independently analyzed the transcripts. We inductively looked for patterns and threads in the data, marking them with labels, and then grouped and organized these labels into

Demographic	Phase One (N = 30)	Phase Three (N = 22)
Age		
18–34	66.7%	95.5%
35–54	26.7%	0%
>55	6.6%	4.5%
Gender		
Male	53.3%	36.4%
Female	46.7%	63.6%
Education		
College	6.7%	45.5%
Bachelor's	30.0%	40.9%
Master's	36.7%	9.1%
Other	26.6%	4.6%

Table 2: Demographic Information: Phase One and Phase Three.

themes [40]. The research team held regular meetings to discuss the initial results during this time, and arrived at the final set of themes shown in Table 3 after multiple rounds of discussions and consensus building. Example themes included "Updates interrupt users" and "Users need information for decision making". In the following section, we use the prefix *P* to indicate an interview participant.

3.1.3 Participants

Table 2 summarizes Phase One's demographics. Participants were predominantly between 18–34 years old, with a fairly even gender split. Most were educated, having college degrees, lived in the District of Columbia, Georgia, and Maryland, and earned a median annual income of USD 60,000. All participants owned desktops and 24/30 owned laptops as well. Two thirds of our participants used a single operating system: MS Windows (15/30), Mac OS X (3/30), and Linux (2/30). The remaining third used a combination of two operating systems: both Mac OS X and MS Windows (8/30) or Linux and MS Windows (2/30).

A large portion of our participants were aware of security breaches that were heavily publicized in the media. For instance, 18/30 were aware of the Heartbleed bug [8] and 11/30 were aware of the 2013 Target breach [44]. One-third of our participants had been victims of online breaches and malware including credit card frauds and computer viruses, and 8/30 participants stated they went above and beyond their e-mail providers' services to maintain their security. For instance, one participant reported using text-only mode for reading messages, and another reported scanning all downloaded attachments. Overall, while our participants were gender balanced, they represented a younger, more educated, and as a result, more technology savvy sample.

3.2 Findings

Our formative study showed that software updates interrupt users and their computing activities, supporting findings of previous studies [21, 46], for users of operating systems other than MS Windows . Our study also illuminates new evidence of information barriers to updates namely: trust in vendors, obscure change logs, and unknown installation times. Information barriers extend beyond the wording of unclear and obscure notifications [21] to the update's purpose, possible consequences of applying an update, and information to plan when to do an update. We also noted that unlike in constrained settings [7, 29], our participants were less concerned about updates using up Internet data. Finally, our participants varied on how they wanted to be notified or provide consent for updates based on the frequency of application use and the changes the update was going to perform. Additionally, they wished to manage and control all the updates on their devices centrally.

3.2.1 Interrupting Users

While our participants appreciated the importance of software updates for maintaining security, enhancing performance, and adding new features to their software, they felt that updates disrupted their computing activities in two ways: Interruptive update notifications and reminders diverted their attention while unwanted reboots and context switches lowered their productivity.

Notifications and Reminders: 22/30 participants reported that inopportune update notifications caused the largest disruption because they appeared during regular computing activities such as watching a video, doing a presentation, or during work times. These notifications were also hard to dismiss completely, so the same update could interrupt a user multiple times with reminder prompts. In a typical example, P17 remarked: *"I tend to let the update notifications go away but these days it looks like people keep forcing it so it comes back and back like a zombie."* Our participants prioritized dismissing update notifications and opted for reminders. Yet, these intrusive messages led to them ignoring many software updates because frequent interruptions were annoying and required active attention.

Rebooting and Context Switch: 19/30 participants reported that they delayed updates if they thought the update would require them to reboot machines, restart applications, and save their work. P9's example captures participants' feelings: *"I absolutely put them off until later, because the update requires me to stop what I'm doing, restart the program and computer, and then completely try to reconstruct where I left off."* Even if participants went through with updates, they became frustrated at having to recreate the context of their activities from which they were interrupted. This caused a negative perception of updates as a disruptive force. In another illustrative example, P12 expressed displeasure about restarts losing the context of open tabs in a browser: *"Usually when it tells me I have to shut down my browser, that's when I'm not happy. That and restarting, especially if I know I have a lot of windows or programs or something open, having to restart."*

3.2.2 Information for Decision Making

We asked participants what information they actively sought or wanted for informing decisions about applying software updates. They reported the following factors:

Update Categories: Vendor-specified update categories influenced our participants' decision making. 24/30 participants said they prioritized performing "major updates" including operating system and security related updates over others. P5's quote exemplifies the reasoning: *"I think if I saw the words security or something along those lines, I would be more apt to do the updates than if it said, you*

know, this improves usability."

Other participants revealed that existing vendor categories for updates inadequately captured an update's purpose and often led to them ignoring an update as P20's quote highlights: *"Just being told that it is critical does not really make me feel like it is critical. I need to feel the urgency and feel like there could be a consequence if I don't update it."* Concretely, they mentioned not knowing whether the update improved performance, fixed bugs, or enhanced security upfront and that these factors helped make decisions about going forward with an update or not.

Update Change Logs: Two thirds of our participants reported they glanced through the update change logs. In one telling example, a participant elaborated: *"I read almost all update notes and if it does not have any then I am going to disregard it. I take it pretty seriously. I have to know what the update is going to do on my software." (P17)* However when asked to reflect more deeply, 6 of those admitted the change log was unlikely to influence their decision to update. Close to half of the participants on the other hand, felt logs either presented too little or too much information, remarking that change logs could use less technical language or visual cues to better interpret the update information.

Trust in Vendors: Our participants' opinion of the software vendor influenced their decision to go forward with an update. 19/30 mentioned that they preferred updates from sources they trusted such as the app store or through the vendor's official website. This trust, they further explained, was amplified either through the reputation of the vendor (e.g., a large software company such as Microsoft or Apple), or through positive past experiences with applying updates from that vendor. P2's example quote captures this sentiment: *"I'm pretty good about—just when I see an update request, if it's from a source I know and trust—running it right then."* For our participants, trusting an update's creator was crucial for making a decision to go forward with a suggested update.

Even though trust was important to our participants, they often had trouble finding reputable sources for updates as P8 explains: *"Sometimes finding reputable sources for updates can be challenging and some software packages put their software out on all sorts of different sites. And, you know, it's like which one of these guys do I really trust to download from?"*. Participants tried to ensure that the updates they installed were legitimate but found it difficult to easily determine the authenticity of an update in current updating interfaces.

Compatibility Issues: 16/30 participants struggled with updates that caused unexpected consequences such as removing certain features they used or that led to compatibility issues with other software. Other participants felt that they were forced to install updates to ensure that software they used frequently would not stop working. In an example illustrating this theme, P13 said: *"Typically it's because I have no other choice. If the program that I want won't run on the version of Windows that I have, and I really want to run that program, I'll do the operating system update."* In another instance, P7 complained that their computer crashed after an update and they had to perform a system restore to get things working again. Overall, compatibility issues made participants reluctant to apply updates especially since they could not predict these interactions in advance.

User Interface Changes: 16/30 participants were dissatisfied with updates that changed the user interface because they had to re-acquaint themselves with the application. This is captured by P9's quote: *"For example, one of the updates on one of my frequently used programs switched around the confirm and cancel buttons."* This change caused him to inadvertently erase documents that he needed following the update. Participants thus became averse to updates with user interface changes but more importantly, they could not always predict which updates would have these changes.

Social Influences: Nearly a third of the participants discovered security flaws and updates through social influences such as online media blogs, the news, or through family and friends. While in some cases these social cues pushed participants to actively seek out updates, in other cases—especially with large and critical updates—they made users cautious about performing updates they had been warned against by the media or social networks. For example, P3 said: *"Usually when I hear about updates from things like PC Magazine or those people start talking about it, and then I would just read about it—just to get an idea before I even consider whether to do it."* In another instance, P6, a frequent Mac OS X user explained: *"There are some that I don't do at least until I go online and research it."* Participants therefore depended not only on vendor-specified information but also on social networks and the media to inform them about whether or not they should apply a particular update.

Results Post-Update: A little over one-third of our participants mentioned that they could not always discern the changes an update made post-update, because they received little, if any, feedback about the update's actions. This made them question the overall benefits of updates especially when they had invested considerable time and effort in applying the update (e.g., interrupting their primary activity and rebooting their machines).

Infrastructure Constraints: 8/30 participants told us they avoided updates because of infrastructure constraints such as insufficient disk space and slow Internet connections or because they believed updates slowed down their machines. In a typical example, P23 explained: *"Some software updates take a lot of additional space because it always comes with that extra storage amount that I need. So it is like all the updates that I am doing are only making my computer slower so it's an annoyance."* Participants also avoided updates to save data but to a lesser extent than suggested by findings in settings where Internet constraints are more prevalent [7, 29]. For this reason, participants told us they wanted the update size in advance to more easily weigh the costs of doing an update.

Installation Time: Because software updates disrupted our participants' workflow, they sought information to organize their activities such as the time required for updates to complete. 7/30 participants indicated a willingness to perform updates if they had access to this information in advance. P13 summed this up as: *"But if they actually tell me how long it'll take, that will make me more willing to start an update."* Knowing how long an update process would take, participants told us, would allow them to perform updates

at convenient times.

3.2.3 *Users in the Update Loop*

We asked our participants how they wanted to be notified and provide consent for downloading and installing updates.

Frequently Used Applications Matter: 17/30 participants mentioned that the frequency of use and their perception of an application's importance determined the degree of care they expressed towards keeping software and devices updated. In one example, P15 explained: *"An Evernote plug-in was not up to date and it asked me to update it. And I just deleted it because I don't want to deal with going through an update for a program that I don't use all that much."* Participants were most concerned about applications that mattered to them in some way; either by frequency of use or if it served some crucial function for them.

Tracking Updates: Just over a third of our participants found it difficult to track update downloads and installs because update settings and notifications were spread over multiple locations for the operating system and third party applications. 11/30 talked about needing a central update manager to review updates for all the software installed on their devices. Specifically, participants found it difficult to easily tell what needed to be updated and when or how to change the update settings for applications and the operating system. Overall, participants desired a central location on their devices to track all updates.

Phase One Themes	Participants (N = 30)
Interrupting Users	
Notifications & Reminders	22
Rebooting and Context Switch	19
Information for Decision Making	
Update Categories	24
Update Change Logs	20
Trust in Vendors	19
Compatibility Issues	16
User Interface Changes	16
Social Influences	12
Results Post-Update	12
Infrastructure Constraints	8
Installation Time	7
Users in the Update Loop	
Frequently Used Applications Matter	17
Tracking Updates	11

Table 3: Themes from Phase One.

At the end of Phase One, the research team decided to focus on addressing three main areas of concern stemming from the formative work as show in Table 3, namely updates interrupting users, the lack of adequate updating information, and finding ways to keep users in the update loop. We describe the part of our user-centered design process next.

4. PHASE TWO: PROTOTYPE DESIGN

In Phase Two, we created a low-fidelity, interactive prototype using MS PowerPoint to improve how users receive updates on their machines and how an update system could scaffold users' decision making. The prototype contains images of our mocked up updating interfaces linked together to create the illusion of a working system. Users can navigate through the mock system using the links to explore a limited set of predefined features with a system concept for each feature. Our goal with the prototype was to elicit user reactions to the different design concepts it embodies as described in this section. As such, we focused less on the implementation details such as how the information presented could be acquired ahead of time.

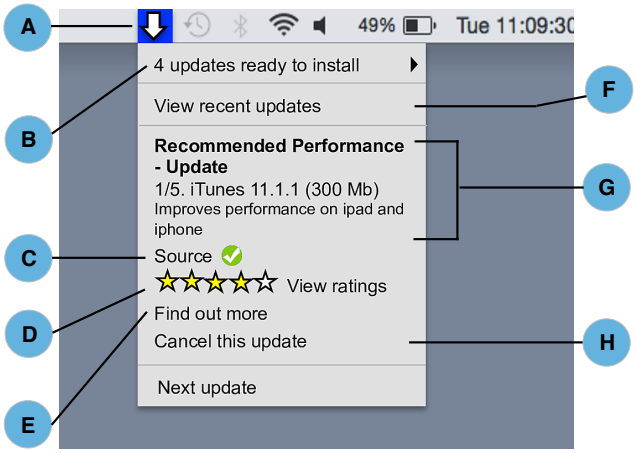


Figure 1: Update DropDown Menu. A: Update Icon. B: List of Remaining Updates. C: Source Verification. D: Update Ratings. E: Additional Information In Central Update Manager. F: List of Recent Updates In Central Update Manager. G: Summary Information For Current Update. H: Cancel Current Update.

Given that both desktop and mobile are heading towards an app-based model of software distribution [5, 27, 35], we chose to create the prototype for Apple's Mac OS X, a major operating system player [41], that has been using this model since 2010 [33]. For future work, we will extend this research to other operating systems and platforms.

4.1 Method

After identifying three areas of concern to users, the research team explored the design space for improved updating interfaces. We began with a lightweight sketching, brainstorming, and ideation phase over these themes. After several iterations of sketches, mockups, and designs were refined by feedback sessions with the research team, we settled on the following issues to alter in updating interfaces. First, we modified how users are interrupted about updates and the manner in which they provide consent to updates. Second, we augmented the information users need for making decisions about updates such as providing clear and concise update logs and the type of the update. Third, we consolidated the updates across a device into a single update manager for users to keep track of updates. We then sketched multiple designs of improved interfaces, discussing and validating the decisions we took at each point, and condensed these into a prototype we describe in the following paragraphs.

4.2 Altering Update Interruptions

Our *minimally-intrusive* low-fidelity interactive prototype alters how end-users are interrupted by either update or reboot notifications in two ways: first, in the concept behind our design, we assume that all update notifications are pushed through a single channel icon, and second, we assume we can piggyback updates requiring restarts to other times when users restart their applications or devices. We designed the following features in the interface mock-ups to reflect these concepts:

4.2.1 Single Update Notification Icon

All update notifications are reduced to a minimal visual cue, the subtle animation of a single system tray icon (Figure 1A). This inverted arrow icon animates only when an update is being downloaded or installed. Users can click on the icon to display a list of impending updates (Figure 1B).

4.2.2 Silent Updates

Conceptually, our design forces all updates to download and install automatically by default without a user's consent. When available, we envisioned that an update lives in the list of updates for a buffer period (e.g., 24 hours) to allow users to intervene. If needed, users can cancel it via the "Cancel this update" option (Figure 1H). When this buffer period expires, in our design concept the update is automatically downloaded and installed but users can continue using their applications without any reminders to restart. In cases where the restart is fundamental for an update to function, in our design concept, users' systems may remain unpatched until the next restart. Our design does not eliminate disruption from unwanted update changes but shifts the onus onto the user to decide whether or not to proceed with an impending update based on additional information. We made this design decision to be provocative to evaluate if users prefer a universally "silent" update mechanism across all their operating system.

4.3 Addressing Lack of Information

Our *information-rich* design adds to existing update information to help users accept or ignore updates via an update summary and post-update feedback.

4.3.1 Update Summary

Each impending update (Figure 1G) contains four important details we learned were lacking in the formative study and a "Find out more" (Figure 1E) link to more details in the centralized update manager:

1. *Source Verification:* Our design displays a green "tick-mark" (Figure 1C) next to the name of the software vendor to indicate a verified and trusted source.
2. *Update Type:* We tag each update with one of five categories (Figure 1G): UI fix (user interface changes), Bug fix (fixes software bugs), Security fix (fixes a major security flaw), Performance (performance enhancements), Compatibility (could cause compatibility issues with other software) to help users learn the update's purpose at a glance.
3. *Update Size:* To inform users about the data and disk space an update might consume, we display the size of the update (Figure 1G) in the summary.
4. *User Ratings:* Each update displays a five star rating (see Figure 1D) based on other users' experiences with it; with one star being poor and five stars being excellent.

4.3.2 Post Update Feedback

Participants in the formative study could not easily identify when an update was installed or what changes were made by the update. Our design shows a pop-up message when a user closes a newly updated application or the operating system for the first time post-update as shown in Figure 4. This message shows the changes made to the application and the date on which the most recent update was installed. The pop up also forces the user to rate the update before they can close the application. In our design concept, update ratings are mandatory to ensure they are eventually populated with information and to force participants to comment on this feature.

4.4 Centralizing Update Management

In the formative study, users desired a way to track *all* the updates across their device. Our *user-centric* prototype conceptually houses and controls the information and settings for all software updates through a central software update manager on the device. A "Pending" tab (see Figure 2) shows the updates that have been downloaded but not yet been installed, or that have been canceled by the user; the "Installed" tab shows recently installed updates to be viewed by last week, last 30 days, or all time; and finally, the "Ratings" tab shown in Figure 3 shows review comments, update ratings, and allow users to add their own reviews. Updates display a:

1. *Change Log:* Each update has a change log in bullet-point form clearly listing the changes the update makes as seen in Figure 2C.
2. *Time to Install:* We show the estimated time (Figure 2E) to install an update to help users plan when to do an update.
3. *Compatibility Report:* All the known possible disruptions an update might cause are listed in a report (Not shown).
4. *Update Settings:* Users can reconfigure updating settings—silent, automatic, or manual—for every application or the operating system from the central update manager (Figure 2D).

Comparison to current Mac OS X Updating System: The current Mac OS X operating system notifies users about an incoming update by means of two notifications (a pop-up, and an red call-out on the App Store icon) when updates are requested manually. No explicit notifications are provided when updates are downloaded and installed automatically. Our prototype switches all updates to silent, provides ambient notifications via an update notification icon, and does not seek users' consent to update by default.

We based our central manager design on the current Mac OS X App Store interface, which handles updates for all App Store applications and the operating system only but not third party applications. Our version of the manager adds

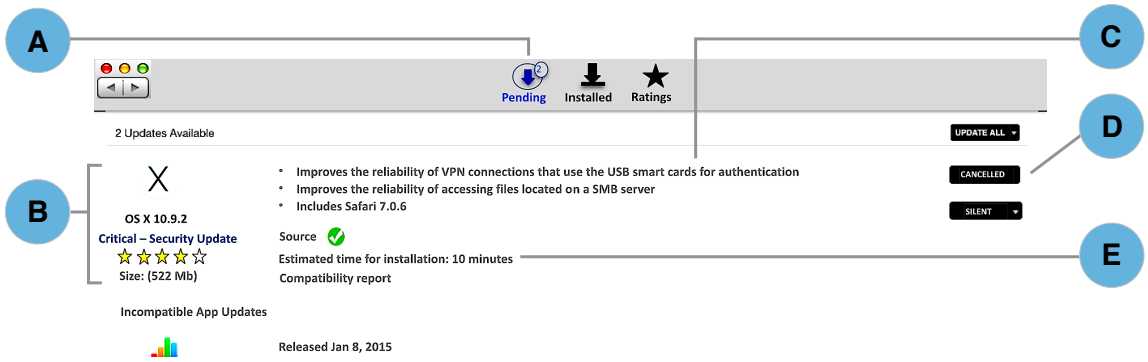


Figure 2: The Central Update Manager's "Pending Updates" Tab. A: Icon Showing No Of Pending Updates. B: Summary of the Update Information. C: Change Log. D: Update Preference Settings. E: Estimated Installation Time.

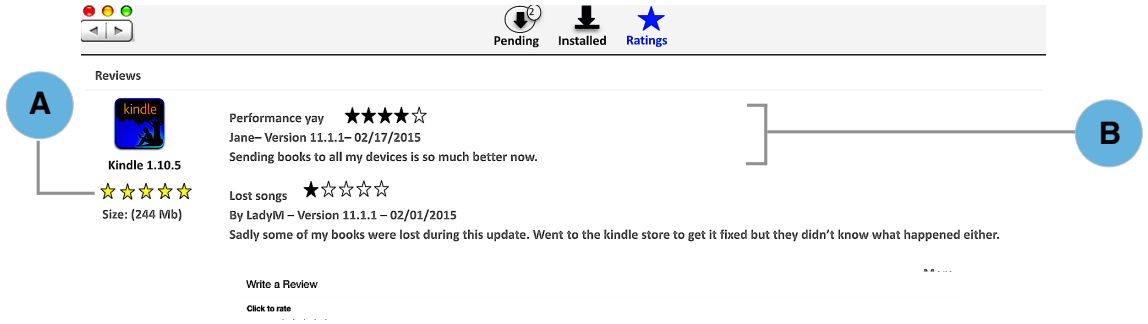


Figure 3: The Central Update Manager's "Ratings" Tab. A: Update Overall Rating. B: Update Rating and Reviews.

to the information the current App Store displays, with a clear and cohesive description of the change log in bulleted form, and the update's type, size and ratings, along with an estimate of the installation time and compatibility report. Unlike the current App Store, our version of the update manager includes an update configuration—manual, automatic or silent—for each application.

Post-update, the current Mac OS X operating system notifies users about an installed update by means of placing a tiny blue dot next to the application icon in the app "Launcher" menu. Our prototype, on the other hand, presents users with a dialog to notify them an update has taken place and to solicit a rating.

To sum up, we designed our proof of concept prototype to minimize interruptions, augment the update information available to users, and to centralize update management across a device. Our design was purposefully extreme in nature—in this case, having *all* updates as silent and having *all* applications solicit feedback post update, providing users with *all* the necessary information—much like a breaching experiment [10] to elicit user reactions and feedback.

5. PHASE THREE: EVALUATION

In Phase Three, we evaluated our low fidelity prototype. Although software updating, much like other security tasks, is a secondary task, our objective with evaluating the prototype was to elicit users' reactions and feedback on our design concepts and features.

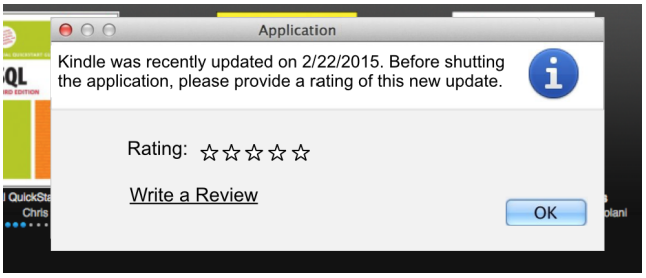


Figure 4: In-Application Post-Update Feedback Dialog.

5.1 Method

5.1.1 Procedure

To evaluate the prototype, we recruited 22 adult Mac OS X users via advertisements on our institutional mailing lists, social media (Facebook, Twitter), and from users who participated in Phase One of the study. Each participant completed a pre-study demographic survey before participating in the think-aloud session, a technique employed regularly in usability testing to gather feedback on proof of concept designs. Specifically, we employed the "speech communication" think-aloud protocol by Boren and Ramey [6], using verbal communication only as a means to acknowledge user response and keep the thought process alive [37].

Each participant performed a series of 11 tasks with our low-fidelity interactive prototype on a laptop provided by the research team. The tasks were designed to elicit user interaction with the single update icon, the central manager, and an application post-update. The first task was to search for updates on the machine to observe whether participants could find the update notification icon on their task bar. When a participant located and clicked on the icon, they were shown a list of updates. Next, participants were asked to cancel an update and following this task, they were asked to check for other pending updates on the machine and to verify whether the source of an update was genuine.

Participants were then introduced to the central update manager through a task to find out more about a particular update which linked them to the manager. In this view, they had tasks that asked them to view the specific additional information provided about updates such as the time to install and the compatibility report. Participants were also asked to do tasks that allowed them to see an empty list of pending updates and the installed and pending tabs in the central update manager. Finally, participants were tasked with rating and reviewing an update as well as changing the update setting for the operating system. The list of think-aloud tasks available as part of the Appendix.

Participants were instructed at the beginning of the session how to provide feedback, how the session would be recorded, and how they were to proceed through the tasks with the guidance of two facilitators from the research team. They were then presented with a paper list of tasks and two researchers observed the participants as they interacted with the prototype, recording their thoughts and actions as they completed the tasks. The researchers used probes such as "keep talking" and "um hmm" to remind participants to verbalize their thoughts when they stopped talking. When the participants failed to speak for more than 30 seconds, one of the researchers asked a stronger probing question relating to the task at hand. For instance, when participants failed to locate cancelled updates, the researcher asked "Where would you expect to see these updates and why?". For each participant, task responses that did not require these strong probing questions were marked "successfully completed".

Once the think-aloud session was complete, the researchers conducted an semi-structured exit interview with each participant. In this interview, we first explained the silent updating mechanism to our participants, i.e., how updates would install on their devices and how restarts would function. We then asked them about both their positive and negative reactions to the prototype, how they performed each task, and the design concepts embodied by the prototype. We also sought feedback for those tasks that the participants were unable to complete noting down how they expected the system to function. Each session lasted between 45 minutes to 1 hour and both the think-aloud session and exit interview were audio and video-taped. Participants were compensated with a USD 15 gift card for the entire session. The study was approved by the IRB of our institution.

5.1.2 Data Analysis

Once the think-aloud sessions and exit interviews were transcribed, two researchers—including one from Phase One—independently analyzed the data. We created profiles for each participant and noted their completion of the various think-aloud tasks. We also analyzed the transcripts using the same process as for Phase One—specifically seeking for differences in participants' reactions to the various design elements. In the following section, we use the prefix T to indicate a think-aloud participant.

5.1.3 Participants

Table 2 summarizes the demographics of the think-aloud session participants. Almost all of our participants were aged between 18 and 35. About 85% of them had completed their bachelor's degree or college, and either worked at, or near, or studied at our institution. Since many of our participants were students, the median annual income reported was <USD 25,000, Our participants owned a median number of 2 Internet-enabled devices and went online more frequently on laptops and smartphones than desktops, tablets, and gaming devices.

Similar to Phase One, our participants were aware of security breaches that were heavily publicized in the media—17/22 knew about the Heartbleed bug and 16/22 about the 2013 Target breach. Again, one-third of our participants had experienced either viruses or malware on their systems. Unlike Phase One, none of our participants reported extra measures to enhance their email security but claimed to ignore email attachments if the email was sent from an unknown source (15/22) or if it looked suspicious (18/22). Overall, our participants were younger, less educated, and less technical than our participants from Phase One.

5.2 Findings

Our prototype elicited varying reactions from our participants in the think-aloud session. Our participants wanted to be notified and actively consent to some but not all updates, they were positive about augmented update information as input for update related decision making, and appreciated the control of centralizing software update management.

5.2.1 Interruptions Sometimes Required For Control

Participants struggled to find the update icon—half failed to notice the system tray icon entirely. However, once participants discovered the icon, they were able to complete the remaining think-aloud tasks with ease. About half preferred our design to current software updating interfaces because they believed it would interrupt them less. T20 explained: *"It prompts me the least. I don't have to worry about it, I don't have to think about it."* The other half of our participants reacted negatively to the prototype's silent update mechanism and told us they wanted to be in the update process more actively. These participants' wanted updates with notifications at download and install time, particularly for applications they used frequently or depended on in some way, to prevent undesired changes. In an example quote, T8 said: *"I want to know how frequently the updates are, how frequently they're occurring and if there's something new or there's a bug. If there any changes, I want to know when and how they happened."*

When asked for further feedback, participants revealed they were willing to adopt a silent updating mechanism for a few of their applications. For instance, participants were amenable to silent updates from trusted vendors and for applications that did not impact their workflow significantly.

T9 said: *"I definitely prefer the silent so I wouldn't really have to do anything and to constantly be the best experience that I could have."* They believed a silent update process would keep their system secure and up to date since they often ignored updates otherwise.

5.2.2 Users Desire Improved Update Information

When completing tasks to read and interact with the additional update information we provided, participants were generally positive that the information would help them to make decisions about updates more so than current interfaces.

Update Type: 15/22 participants reacted positively to the update type, telling us these labels would clarify the purpose of various updates. However, about a third said the update type would not influence their decision to apply an update. Others preferred the binary classification of "critical" and "non-critical" to multiple labels, which they felt could be overwhelming and possibly confusing. Poor labels, participants told us, could backfire and cause a user to avoid an update if they obscured the update's purpose. In all, participants appreciated the concept of more informative labels for the update type to help in their updating decision making process.

Compatibility Report: 19/22 participants were able to find and interpret the compatibility report and 12/22 told us it would be helpful for singling out potentially problematic updates. In a typical example, T8 said: *"If I'm going to be using those apps, I wouldn't go ahead with the update because it will cause problems."* One suggested improvement was to only display compatible updates and only one participant worried how this report would be implemented. Overall, participants desired this predictive capability to help them prevent updates having an unwanted ripple effect.

Ratings: 15/22 reacted positively to the concept of update ratings with 12/22 saying that ratings would potentially influence their decision in going ahead with an update (especially with large updates). T9 explained: *"If there were mostly good reviews and there was nothing standing out as 'oh, this is a problem for my computer' it would help me make the decision to go ahead."* At least five participants wanted to see number of ratings for an update to better contextualize the information. Almost a third of participants felt that ratings would not add any value to their updating experience because they paid less attention to others' opinions. Most of our participants resonated with the idea of leveraging social networks for information to help them decide about performing updates.

Post-Update Feedback: All the participants were unanimous in disliking the concept of providing a mandatory rating post-update, seeing this as a nuisance in the long run. However, they were willing to provide feedback for updates that made visible changes (e.g. user interface modifications), and for applications that they frequently used or were important to them.

Time to Install: 13/22 found that having the information about how long was needed to install an update prior to beginning the process was useful for deciding when to do an update. T8 said: *"I think that's very important because if you're in a hurry or if you have some other work to do and sometimes you should know if you can finish the update by then or not."* 1 participant wanted aggregated view of the time required for all the pending updates and another wanted the time to install to be visible in the list of updates not just the central manager. For participants, having an estimate of the time involved for the update process was crucial for planning so as to minimize interruption to their activities.

Installation Size: Four participants reacted positively to the size of the update being displayed upfront. These participants desired some warning if the update would consume their remaining disk space. Two participants felt this information was less useful as they cared less about disk space on their devices.

Source Verification: Nine participants reacted positively to having the ability to easily identify an authentic update source. In an illustrative example, T17 remarked: *"If it's not from the verified source, then that will be the one thing that will stop me from installing the update."* However, several said they would probably not pay attention to this cue and a few felt that this cue was most important for third-party applications only. It was clear, however, that visual cues indicating update authenticity can build trust with users.

5.2.3 Users Prefer Centralized Update Management

Over half of the participants reacted positively to centralized software update management, especially for non-Mac applications that currently do not push updates via the app store. For example, T13 said: *"I like it: It seems more comprehensive because it has (for e.g.) the Microsoft stuff in it so you don't have to run the Microsoft updater as well as the app store updater mechanism."* Only a few participants thought that being able to control the update preference on a per-application basis in a central update manager was useful. However, it was unclear if participants reacted this way to the additional controls because they told us they generally preferred to keep default settings. Participants suggested the central update manager could also provide a history of updates so that changes could be rolled back to a state before the update occurred if something went wrong. In summary, participants felt managing updates in a single location would reduce information overload and make the updating process more consistent across a device.

6. DISCUSSION

Contrary to Wash *et al.*'s paper [49] which suggests that increasing usability may enable those users who wish to be less secure to apply fewer updates (i.e., to switch to manual updates), our findings suggest that users may want to be less secure in the first place because they suffer from a poor updating user experience. Improving usability therefore should still be a goal for encouraging users to apply updates that are security related. This is particularly important as the Internet of Things evolves and users are faced not only with updating their personal devices but devices in their homes, office, on their bodies, and elsewhere. Our findings suggest four primary directions to improve desktop software updating interfaces: personalizing update interfaces, minimizing update interruptions, improving update information, and centralizing update management. We outline considerations about the socio-technical aspects of the software updating process specifically around trust, control, and consent.

6.1 Personalizing Updating Interfaces

Our first recommendation is to personalize updating interfaces to minimize notifications about updates that can safely be made silent and find those applications or systems that a user is likely to want to monitor for changes. In this vein, we could effectively increase the number of updates applied silently and reduce the overwhelming amount of notifications that may not be of interest to the users. This recommendation stems directly from our findings that users desire some control over what changes are made to a machine or application that they depend on in any capacity.

We envision the personalization of updating interfaces could occur in a similar way as others have suggested [22] in the domain of requesting Android application permissions. In other words, users could be shown only update requests that demand their attention and decision making, such as for applications they use actively on a day to day basis. This would give users ample opportunity to cancel potentially disruptive updates and minimize unwanted consequences. We recommend empowering users with a "cancel" option to prevent an update from ever occurring at the risk of them never applying certain updates. This stands in contrast to current automated updating systems that delay updates but install them anyway if a user has not responded in a certain period of time. Giving users more power over their systems and applications may make them more likely to trust the updating process.

We also propose that any updates that users do not actively wish to monitor could be made silent. For example, security updates and updates for infrequently used apps (which if left unpatched can still be sources of vulnerabilities [34]) could be applied without the users consent, assuming disk and data constraints are not an issue. Personalizing updating interfaces in this manner of course depends on whether a system can learn which applications or updates the user cares about and how to apply silent updates selectively. Our study highlighted several factors a system could consider for this purpose include the frequency of use of applications over time, its importance and an update's characteristics (e.g., purpose, size, or, installation time) to determine if users should be notified and prompted for consent. Users could also be unobtrusively asked at installation time to designate whether a particular application should notify them of any changes. In such a system, update notifications could also better highlight why certain updates are recommended.

Update interfaces could also be personalized by profiling users' individual personalities traits to see whether these can be correlated with various updating behavior preferences. Already, the Security Behavior and Intentions (SeBIS) scale [20] has shown that users vary in their software updating behavior intentions and how differences in risk taking, decision making, impulsiveness are correlated with security decisions (and software updating in particular) [19]. Future research could identify how automation defaults or updating interfaces could be configured as a function of these characteristics and how to better involve users in decision making about when to apply updates.

6.2 Minimizing Update Interruptions

Our second recommendation for improving desktop updating interfaces is to minimize update interruptions where possible to increase the uptake of updates including those that are security related. Update interruptions can be improved at the interface level by making update notifications less intrusive. For example, similar to our proposed design, update notifications could sit somewhere between passive and active notifications using icons that subtly and visually morph to indicate to a user that an update is available and provide more information to only those that desire it.

Future work could also consider how we can leverage Dynamic Software Updates (DSU) [26] to avoid restarts caused by updates altogether to minimize update interruptions at the back-end. Updates restarts could also be piggybacked on times that a user restarts a system or application on their own. This would also create a need for designing visual cues and nudges to gently prompt a user to restart an application or their machine to enable an update to be applied. For example, Google Chrome colors the Chrome Menu icon from green to orange to red over time to nudge users to relaunch their browsers [25].

6.3 Enhancing Update Information

Our third recommendation to improve updating interfaces on the desktop revolves around better informing users about updates and their consequences specifically by providing more information that builds trust in the update process, e.g., via compatibility reports and update ratings. Further research into how to generate compatibility reports and how update ratings can be gathered and provided will help users make informed choices about which updates to apply. For instance, "social proof" has already been shown to improve security feature adoption in Facebook and update ratings could similarly help users assess if they should move forward with an update [15, 14]. Other visual enhancements to show that update sources are verified or vetted could also instill trust in the updating process and further motivate users to perform updates.

6.4 Centralize Desktop Update Management

Our fourth recommendation for the desktop updating interface is to centralize update management where possible. We envision a large scale change to current interfaces that would require operating system vendors to provide better ways for updates to be pushed through a single channel or for the information about updates to be gathered for a central update information repository across a device. Examples of applications that are already making strides in this vein include Metaquark's AppFresh [30] for centralizing Mac OS X updates and SparkleProject, a framework for third party application developers to push Mac OS X updates [1]. This model is already manifest on mobile phones where updates and their notifications already propagate more centrally than on the desktop through app stores.

We also propose that a central update manager could provide ways for users to preview the effects of an update on their system to mitigate the fact that users are averse to unwanted changes. For instance, users could be given the option to try out the new version of an application or operating system via an interface overlay or by using a parallel version of an application installed on a system without committing to the update process or applying the update. Providing an easier way to roll back unwanted changes may also make users more amenable to apply updates without fear of

breaking their workflow, and therefore can potentially enhance security if those updates are applied as well.

6.5 Socio-Technical Updating Aspects

Our study and recommendations highlight the complex socio-technical nature of updates and the stakeholders involved. Updates involve trust between users and those seeking to make changes to their systems, gathering consent from users to make those changes, and surrendering control to external parties to make those changes. This involves a complex interplay of actors such as application and operating system vendors, developers, and users.

The question of whether these stakeholders will want to make the updating interface improvements we recommend possible remains open. For example, our recommendations depend on application developers modifying the information they provide about updates, how updates are deployed to users, how users are notified about updates, and the potential consequences of applying any update. To centralize update management, operating system vendors will have to build supporting frameworks to enable developers to push updates centrally and to have a vetting process similar to mobile systems for checking all updates and whether they comply with established guidelines for best informing users of upcoming changes.

Vendors might have large incentives to improve updating interfaces or otherwise risk alienating and losing their user base. For instance, a recent study of Tinder and Tesla updates [2], showed the backlash of unhappy users when unwanted changes were made without their consent to these apps. Yet, it is unclear who should have ultimate control over software changes, and whether there needs to be some governmental oversight for security purposes and consumer protection, particularly as some user bases grow as large as the population of several countries.

Furthermore, if update interfaces are indeed personalized, the question of transparency and accountability for selectively applying silent updates is also open i.e., how would such systems explain their decision making to users in a comprehensible way and provide meaningful controls to users so that they can still provide consent for changes being made to their applications and systems. We will only know more about whether a personalized interface would make or break users' mental models of updates by testing these recommendations out in the wild.

In all, these socio-technical issues around software updating will require the usable security community to closely examine the practice of software updating from all viewpoints. We can certainly empower users through improved software updating interfaces but we need to better understand all the nuances of control, consent, and trust in updates. We should also be asking ourselves the question of who should be allowed to make changes to systems to properly address the issue of improving security through updates.

6.6 Limitations and Future Work

Our studies have several limitations. First, the use of self-reported data in Phase One is subject to recall bias, meaning there may exist errors and differences in the recollections recalled by our participants. Second, because of the low fidelity nature of our prototype, its evaluation in Phase Three is based on participants' opinions rather than their actual behaviors. The evaluation also required participants to work with updates as a primary task and thus, the results may not generalize to real-world settings where software updating is considered a secondary task. Third, our participant pools were dominated by a younger set of participants, with many participants in Phase One having advanced degrees, and many students in Phase Three. Therefore, the results from both our studies may not generalize to the entire population, and hold limited validity.

To address these limitations, future work could examine updating behaviors more deeply with a higher fidelity prototype that can be deployed in the field, and validated against actual user behaviors. We also recommend testing future updating interface designs against a more representative group of Mac OS X users and extending the work to other operating systems and devices. Future work could explore how to improve updating interfaces beyond the desktop space, i.e. to mobile and the Internet of Things. Finally, given that updating involves an ecology of stakeholders, future work could examine updating practices from other perspectives such as how network administrators manage updates for large groups of users or how developers create updates in the first place.

7. CONCLUSION

We used a three phased research process to investigate current user barriers to software updates and determine how to improve desktop software updating interfaces. We found that users avoid updates primarily because they interrupt their computing activities, they lack information that enables them to decide whether to apply an update or not, and they notify and involve users in ways that are undesirable. Users responded positively to our *minimally-intrusive*, *information-rich*, and *user-centric* low fidelity prototype designed to minimize how updates interrupt the user, to augment current updating information to help users make a decision to update or not, and to centralize update management across a device. Based on the evaluation, we suggest that updates can be improved by personalizing software updating interfaces, minimizing update interruptions, improving information for update decision making, and by centralizing update management across a device. We also believe that the socio-technical aspects of updating are complex and need to be explored in more depth for future work. Ultimately, improving update interfaces and addressing these open questions will enhance the security overall.

8. ACKNOWLEDGEMENTS

We would like to thank Susan Wyche, Michelle Mazurek, Katie Shilton, and members of the Human–Computer Interaction Lab at the University of Maryland for their feedback on drafts of this paper. Our research is based upon work supported by the Maryland Procurement Office under contract H98230-14-C-0137. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Maryland Procurement Office.

9. REFERENCES

[1] Sparkle Project. `http://sparkle-project.org`, 2015.

[2] A. Acker and B. Beaton. Software Update Unrest: The Recent Happenings Around Tinder and Tesla. In

Proceedings of the 49 Hawaii International Conference System Sciences, HICSS '16. IEEE, 2016.

[3] D. Akhawe and A. P. Felt. Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness. In *Presented as part of the 22nd USENIX Security Symposium (USENIX Security 13)*, pages 257–272, Washington, D.C., 2013. USENIX.

[4] Apple. Get Software Updates for Your Mac. `https://support.apple.com/en-us/HT201541`, 2015.

[5] D. Barrera and P. Van Oorschot. Secure Software Installation on Smartphones. *IEEE Security & Privacy*, 9(3):42–48, May 2011.

[6] T. Boren and J. Ramey. Thinking Aloud: Reconciling Theory And Practice. *Professional Communication, IEEE Transactions on*, 43(3):261–278, Sep 2000.

[7] M. Chetty, R. Banks, A. Brush, J. Donner, and R. Grinter. You're Capped: Understanding the Effects of Bandwidth Caps on Broadband Use in the Home. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 3021–3030, New York, NY, USA, 2012. ACM.

[8] Codenomicon. The Heartbleed Bug. `http://heartbleed.com`, April 2014.

[9] H. Corrigan-Gibbs and J. Chen. Flashpatch: Spreading Software Updates over Flash Drives in Under-connected Regions. In *Proceedings of the Fifth ACM Symposium on Computing for Development*, ACM DEV-5 '14, pages 1–10, New York, NY, USA, 2014. ACM.

[10] A. Crabtree. Design in the Absence of Practice: Breaching Experiments. In *Proceedings of the 5th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, DIS '04, pages 59–68, New York, NY, USA, 2004. ACM.

[11] L. F. Cranor. A Framework for Reasoning About the Human in the Loop. In *Proceedings of the 1st Conference on Usability, Psychology, and Security*, UPSEC'08, pages 1:1–1:15, Berkeley, CA, USA, 2008. USENIX Association.

[12] N. Cyber Security Alliance. Stay Safe Online. `https://www.staysafeonline.org/`, 2015.

[13] J. Dadzie. Understanding Software Patching. *Queue*, 3(2):24–30, Mar. 2005.

[14] S. Das, A. D. Kramer, L. A. Dabbish, and J. I. Hong. Increasing Security Sensitivity With Social Proof: A Large-Scale Experimental Confirmation. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 739–749, New York, NY, USA, 2014. ACM.

[15] S. Das, A. D. Kramer, L. A. Dabbish, and J. I. Hong. The Role of Social Influence in Security Feature Adoption. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 1416–1426, New York, NY, USA, 2015. ACM.

[16] P. Dourish, E. Grinter, J. Delgado de la Flor, and M. Joseph. Security in the Wild: User Strategies for Managing Security As an Everyday, Practical Problem. *Personal Ubiquitous Comput.*, 8(6):391–401, Nov. 2004.

[17] T. Duebendorfer and S. Frei. Why Silent Updates Boost Security. *TIK, ETH Zurich, Tech. Rep*, 302, 2009.

[18] W. K. Edwards, E. S. Poole, and J. Stoll. Security Automation Considered Harmful? In *Proceedings of the 2007 Workshop on New Security Paradigms*, NSPW '07, pages 33–42, New York, NY, USA, 2008. ACM.

[19] S. Egelman and E. Peer. The myth of the average user: Improving privacy and security systems through individualization. In *Proceedings of the 2015 New Security Paradigms Workshop*, NSPW '15, pages 16–28, New York, NY, USA, 2015. ACM.

[20] S. Egelman and E. Peer. Scaling the Security Wall: Developing a Security Behavior Intentions Scale (SeBIS). In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 2873–2882, New York, NY, USA, 2015. ACM.

[21] M. Fagan, M. M. H. Khan, and R. Buck. A Study of Users' Experiences and Beliefs about Software Update Messages. *Computers in Human Behavior*, 51, Part A:504 – 519, 2015.

[22] A. P. Felt, E. Ha, S. Egelman, A. Haney, E. Chin, and D. Wagner. Android Permissions: User Attention, Comprehension, and Behavior. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, SOUPS '12, pages 3:1–3:14, New York, NY, USA, 2012. ACM.

[23] T. Gerace and H. Cavusoglu. The Critical Elements of the Patch Management Process. *Commun. ACM*, 52(8):117–121, Aug. 2009.

[24] C. Gkantsidis, T. Karagiannis, and M. VojnoviC. Planet Scale Software Updates. In *Proceedings of the 2006 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, SIGCOMM '06, pages 423–434, New York, NY, USA, 2006. ACM.

[25] Google. Update Google Chrome. `https://support.google.com/chrome/answer/95414?hl=en-GB`, 2015.

[26] M. Hicks and S. Nettles. Dynamic Software Updating. *ACM Trans. Program. Lang. Syst.*, 27(6):1049–1096, Nov. 2005.

[27] D. Hinchcliffe. The App Store: The New "Must-Have" Digital Business Model. `http://www.zdnet.com/article/the-app-store-the-new-must-have-digital-\newlinebusiness-model`, January 2010.

[28] I. Ion, R. Reeder, and S. Consolvo. "...No one Can Hack My Mind": Comparing Expert and Non-Expert Security Practices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 327–346, Ottawa, July 2015. USENIX Association.

[29] A. Mathur, B. Schlotfeldt, and M. Chetty. A Mixed-methods Study of Mobile Users' Data Usage Practices in South Africa. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 1209–1220, New York, NY, USA, 2015. ACM.

[30] Metaquark. Appfresh. `http://metaquark.de/appfresh/mac`, 2015.

[31] Microsoft. Install Windows Updates. `http://windows.microsoft.com/en-us/windows-vista/install-windows-updates`, 2015.

[32] Microsoft. Microsoft Security Intelligence Report Volume 18: July through December 2014. http://download.microsoft.com/download/7/1/A/71ABB4EC-E255-4DAF-9496-A46D67D875CD/Microsoft_Security_Intelligence_Report_Volume_18_English.pdf, 2015.

[33] D. Murph. Apple mac app store: open for business starting January 6th. http://www.engadget.com/2010/12/16/apple-mac-app-store-open-for-business-\newlinestarting-january-6th/, 2010.

[34] A. Nappa, R. Johnson, L. Bilge, J. Caballero, and T. Dumitras. The Attack of the Clones: A Study of the Impact of Shared Code on Vulnerability Patching. In *Security and Privacy (SP), 2015 IEEE Symposium on*, pages 692–708, May 2015.

[35] J. Newman. How Mobile Apps are Changing Desktop Software. http://www.pcworld.com/article/259110/app_invasion_coming_soon_to_your_pc.html, July 2012.

[36] J. Oberheide, E. Cooke, and F. Jahanian. If It Ain't Broke, Don't Fix It: Challenges and New Directions for Inferring the Impact of Software Patches. In *Proceedings of the 12th Conference on Hot Topics in Operating Systems*, HotOS'09, pages 17–17, Berkeley, CA, USA, 2009. USENIX Association.

[37] E. L. Olmsted-Hawala, E. D. Murphy, S. Hawala, and K. T. Ashenfelter. Think-aloud Protocols: A Comparison of Three Think-aloud Protocols for Use in Testing Data-dissemination Web Sites for Usability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 2381–2390, New York, NY, USA, 2010. ACM.

[38] K. Sankarpandian, T. Little, and W. K. Edwards. Talc: Using Desktop Graffiti to Fight Software Vulnerability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1055–1064, New York, NY, USA, 2008. ACM.

[39] M. A. Sasse, S. Brostoff, and D. Weirich. Transforming the "Weakest Link"–a Human/Computer Interaction Approach to Usable and Effective Security. *BT Technology Journal*, 19(3):122–131, July 2001.

[40] I. Seidman. *Interviewing As Qualitative Research: A Guide for Researchers in Education and the Social Sciences*. Teachers college press, 2013.

[41] Statista. Global market share held by operating systems Desktop PCs from January 2012 to June 2015. http://www.statista.com/statistics/218089/global-market-share-of-windows-7, 2015.

[42] J. Sunshine, S. Egelman, H. Almuhimedi, N. Atri, and L. F. Cranor. Crying Wolf: An Empirical Study of SSL Warning Effectiveness. In *Proceedings of the 18th Conference on USENIX Security Symposium*, SSYM'09, pages 399–416, Berkeley, CA, USA, 2009. USENIX Association.

[43] Symantec. 2015 Internet Security Threat Report, Volume 20. https://know.elq.symantec.com/LP=1542, 2015.

[44] Target Corporation. Data Breach FAQ. https://corporate.target.com/about/shopping-experience/payment-card-issue-faq, April 2014.

[45] Y. Tian, B. Liu, W. Dai, B. Ur, P. Tague, and L. F. Cranor. Supporting Privacy-Conscious App Update Decisions with User Reviews. In *Proceedings of the 5th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices*, SPSM '15, pages 51–61, New York, NY, USA, 2015. ACM.

[46] K. E. Vaniea, E. Rader, and R. Wash. Betrayed by Updates: How Negative Experiences Affect Future Security. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 2671–2674, New York, NY, USA, 2014. ACM.

[47] M. Vojnovic and A. Ganesh. On the Effectiveness of Automatic Patching. In *Proceedings of the 2005 ACM Workshop on Rapid Malcode*, WORM '05, pages 41–50, New York, NY, USA, 2005. ACM.

[48] R. Wash and E. Rader. Too Much Knowledge? Security Beliefs and Protective Behaviors Among United States Internet Users. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 309–325, Ottawa, July 2015. USENIX Association.

[49] R. Wash, E. Rader, K. Vaniea, and M. Rizor. Out of the Loop: How Automated Software Updates Cause Unintended Security Consequences. pages 89–104. USENIX Association, 2014.

[50] M. Wu, R. C. Miller, and S. L. Garfinkel. Do Security Toolbars Actually Prevent Phishing Attacks? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 601–610, New York, NY, USA, 2006. ACM.

APPENDIX

A. PHASE ONE: INTERVIEW GUIDE

Thank you for participating in today's study. As you read in the consent form, we will be recording the session so we can review it to make sure that we don't miss any part of our conversation. Your name will not be associated with the recording or with any data I collect. Your comments and opinions will only be used in combination with the feedback gathered from other people participating in this study. Your individual comments will remain confidential. Do you have any questions regarding the consent form? Do I have your permission to start the recording?

Session Introduction

Today, I'm going to be talking with you about security issues and discussing your software updating habits. The interview should last around 30 to 45 minutes. Before we begin, there are a few things I would like to mention:

1. During our discussion, I will ask you to share with me your thoughts and opinions on the different topics that we cover. I would like you to be honest and straightforward about your knowledge and habits regarding software security. I might ask you to expand on anything you mention if the information could be useful to the study. Please keep in mind that there are no right or wrong answers.
2. We are not here to test you or your knowledge about security issues or software technology. We want to get

a better understanding of how people really think and act when it comes to these issues.

3. Please feel free to comment on any thoughts or ideas you have as we talk. Your feedback is important in that it helps us get a better picture of real user behavior. Do you have any questions before we begin? Okay, let's get started.

Security Awareness

1. Are you aware of any major breaches to personal online security?
2. Have you heard about the following security breaches: Heartbleed, Adobe 2013, Target 2013
3. How do you find about online security issues? From which sources?
4. How do you determine integrity of the source of software update? Have you ever avoided installing an update because it was not from an authentic source?
5. Have you ever been a victim of an online security breach?

Security Habits

1. Do you secure your physical electronics (like laptop, tablet, phone)? How?
2. Are you concerned about email security? Do you take any measures to protect your security on email?
3. Are you concerned about software security? Do you take any measures to protect your security in regards to software?

Software Updates

1. What comes to mind when you think of software updates? How do you feel about them? Do you usually have a positive or negative feeling? What is your motivation for acting on software updates? How do you decide whether to go through the update or not?
2. How comfortable are you in installing software updates? If not, what would make you more comfortable?
3. Do you treat all software updates the same? Or do you think or act differently depending on the type, say whether they are app updates, or operating system updates?
4. Do you feel that it's necessary to update software every time an update is available? Are there some updates that you always do and some that you usually ignore?
5. Does the process of software updating feel like a necessity or more of an interruption? Why?
6. Do you typically understand what the update is going to do to the software before you download and install it? Do you read the information presented in the update notice? Do you understand the information? Does this information have an impact on your decision to go through with the update or not?

Software Update Process

Discovering the update

1. How do you find out about a software update? Do you usually wait to be notified or if you hear about an update from an external source, do you seek it out?
2. Are there any barriers that get in the way of you discovering updates?
3. What makes it easy to discover updates?

Downloading the update

1. How do you download updates? Where do you go or what do you do? Do you seek them out?
2. Are there any barriers to downloading updates? For example, connectivity issues like data usage, Wi-Fi availability?
3. What makes it easy to download updates?

Installing the update

1. How do you typically install updates? Automatically or manually?
2. When do you typically install updates? After download or later (why if later)?
3. Are there any barriers to installing updates (restarting, downloading an installer, interrupting current activity)?
4. What makes it easy to install updates?

Applying the Update

1. What do you typically expect to have happen after you install an update and begin using the software again?
2. Does your interaction with the software typically match your expectations after
3. Do you usually have a positive or negative experience after installing an update?
4. Do any of these factors have an effect on how you feel about future updates?

Software Updating Preferences

1. Do you configure any of your systems to do updates in any of silent, automatic, manual ways? Do you prefer any of these mechanisms? Why or why not? Do you know where to change these settings in your operating system? And the settings for each piece of software.
2. What are the reasons you go through with software updates? Does it depend on the device? Piece of software? Operating system? How do you determine if an update is critical or not? How do you feel about whether an update might address a security issue vs. a feature change?
3. What are the reasons why you might avoid software updates? Frequency of patches (Do you avoid if they're released more frequently?), cost of updates, connection speed, incompatibility with other software, fear of breaking existing software/changing interface.

Current Software Updating Interfaces

1. How frequently do you update your software?
2. Which device do you most frequently update and why? (Does the update behavior depend on the device being used?)
3. What software do you update more often?
4. Have you installed updates both on Windows PC and Mac OS? Which one did you prefer?
5. What browsers do you use to access Internet?
6. Which one do you prefer most in terms of update?
7. Did you ever avoid installing an update to avoid incurring additional charges when your Internet was not free? (wifi/3G/broadband)
8. How would you want to improve the updating process?

B. PHASE THREE: THINK-ALOUD TASKS

Locating updates

1. Find out if there are software updates for your machine
2. Find the list of available updates
3. Cancel the OS X update from taking place

Authenticating the Source

1. Verify that the source of iTunes 11.1.1. update is authentic
2. How would you get additional information about what the iTunes 11.1.1 update does?

Installed Updates

1. Close the update manager and find out if there are any more updates
2. Find out which updates were recently installed

Information About the Update

1. How much time will the next update will take to download and install?
2. Change the update preferences for iTunes updates
3. Check for applications the OS X 10.9.2 update is incompatible with
4. Read update ratings and comments from users

Why Do They Do What They Do?

A Study of What Motivates Users to (Not) Follow Computer Security Advice

Michael Fagan
University of Connecticut
Storrs, Connecticut, USA 06269
michael.fagan@uconn.edu

Mohammad Maifi Hasan Khan
University of Connecticut
Storrs, Connecticut, USA 06269
maifi.khan@uconn.edu

ABSTRACT

Usable security researchers have long been interested in what users do to keep their devices and data safe and how that compares to recommendations. Additionally, experts have long debated and studied the psychological underpinnings and motivations for users to do what they do, especially when such behavior is seen as risky, at least to experts. This study investigates user motivations through a survey conducted on Mechanical Turk, which resulted in responses from 290 participants. We use a rational decision model to guide our design, as well as current thought on human motivation in general and in the realm of computer security. Through quantitative and qualitative analysis, we identify key gaps in perception between those who follow common security advice (i.e., update software, use a password manager, use 2FA, change passwords) and those who do not and help explain participants' motivations behind their decisions. Additionally, we find that social considerations are trumped by individualized rationales.

1. INTRODUCTION

Academics have widely accepted that privacy is not only valued by individuals, but also helps aspects of our society function [24]. Computer/data privacy is no different: many report putting a high value on the ability to control who can access their data and information [17, 22]. Since security of computers is the first step towards computer privacy, it is imperative that we not only create new, stronger cryptographic and security tools, but that we also understand how to best motivate users to adopt new tools and techniques.

The facts of what people can do to stay safe and how they use that advice have been well studied, with many finding divergence between recommended and actual protections [15, 8]. The failure of current and past motivational and/or security approaches [5, 1], lack of information about many facets of the problem, including adaptability of many security advices [12, 13], and issue specific (e.g., updating) concerns [29, 28, 11] have all been noted as part of the explanation for the gap. That said, to the best of the authors' knowledge, no one has broadly approached the question of **"why do some follow security advice, while others do not,"** using empirical data collected and analyzed for that purpose. Though some work has looked at the concerns of users in many specific scenarios of user security, we seek to sift through the context-specifics and overall economics of some security decisions, with the hopes of gaining insight into the overall problem. By investigating these kinds of trends, we can better understand motivation in this area, and evaluate/improve current approaches towards increasing the security of users.

With this study, we investigate the motivations of users to follow or not follow common computer security advice. We model decision-making with a rational, cost/benefit framework, expanding it to include both the concept of risk, which is expected to be key to security decisions, as well as social motivations. Using this grounding in interdisciplinary prior work, we design a web-based survey distributed to those 18 and over living in the U.S. via the service Mechanical Turk. We use 4 common security recommendations (i.e., updating software, use of a password manager, use of 2FA, changing passwords) as a foundation for our surveys. With each advice, we form two groups: one of those who follow the advice (Yes groups), and one that does not (No groups). In all, we collect 290 survey responses constituting both qualitative and quantitative data. Through analysis of this data, we extract the following key findings related to the question "why do some follow security advice, while others do not?":

- Benefits of following are rated higher by those who follow each advice compared to those who do not. Those who do not follow rate the benefits of doing so as higher than the groups that practiced each advice.
- Risks of not following are rated higher by those who follow each advice compared to those who do not.
- Costs of not following are also seen as higher by those who follow each advice compared to those who do not for all but one case (using 2FA).
- Security and convenience are common themes in the qualitative comments. For all tools, Yes groups in many cases report following because they think doing so is more secure. In some cases (i.e., updating and using a password manager), those who follow are also drawn by added convenience.
- Individual concerns are rated higher than social concerns for all variables, indicating low social motivations around computer security.

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22–24, 2016, Denver, Colorado.

These highlighted findings and the full results presented in this paper help towards understanding why those who follow computer security advice, do and those who do not follow, don't. This study continues a long running track approaching this problem and brings new information to the debate on how to best address it. We find evidence to support prior suggestions that users know the costs/benefits involved, but also see gaps between those who do and do not follow each advice on benefits and risks. This implies that at least one or both of the groups for each advice are miscalculating in their considerations, since both positions cannot be simultaneously "true." Knowing who is truly "wrong" is imperative, and so the authors echo the calls of prior work [12] on the need for more data about what users think and experience, as well as measures of actual risk to best interpret ours and other's results.

2. RELATED WORK

This study is influenced by several arms of usable security research. First, our work supposes that for all security related decisions, users are making a rational choice by weighing the costs against the benefits. Second, we add perception of risk to our considerations since this is integral to motivation around secure behavior. Third, we argue there is or should be a social component to users' motivations, so that aspect is also incorporated into this study. Fourth, we choose to look at the dichotomy between those who adhere to good security behavior and those who do not. All four of these tenets are grounded in the literature.

2.1 Security Decisions as a Rational Choice

Though complex, human decision-making can be viewed as a consideration of costs and benefits, where humans are rational actors who choose to minimize cost and/or maximize benefit. This view of computer security decision-making has been prominent. Herely in 2009 was one of the first to suggest that users' failure to adhere to good security behavior could be attributed to them finding the costs too high and/or benefits too low [12]. He supports this supposition by citing the low chance of an actual security breach for any given user and the high cost of daily security maintenance. Herley goes on to suggest that more data is needed to determine the actual costs and benefits of these decisions to better inform the advice experts give. By 2014, Herley found that the approach of researchers had not changed much, leading him to say in a follow-up work:

> It is easy to fall into the trap of thinking that if we can find the right words of slogan we can convince people to spend more time on security. . . . We argue that this view is profoundly in error. It presupposes that users are wrong about the cost-benefit tradeoff of security measures, when the bulk of the evidence suggests the opposite.

What Herley suggests is that rather than users being ill-informed about security, they could be making a perfectly rational decision, at least in their eyes. This view is echoed by Geordie Stewart's 2012 work "Death by a Thousand Facts." Here, Stewart and David Lacey argue that "security-awareness" based approaches to increasing user security have and will continue to fail because, unlike how some researchers assume, users are not ignorant of good security behavior [5].

On the other hand, many studies have generated results that suggest users are miscalculating the costs and benefits. "Out of the Loop" by Rick Wash et al. found that a significant portion of sampled Windows 7 users did not understand what updates were changing in their system and could not execute their intentions for computer management [29]. Vaniea et al.'s "Betrayed by Updates" has similar findings that suggest prior negative past experience could play a large role in users deciding not to apply updates [28].

This divide in the literature on user motivations around computer security could be related to differences in perceptions between people about computer security. Specifically, it is possible that experts and others who follow advice do see the costs and/or benefits of adhering to good security behavior differently. Our study hopes to investigate this view of the issue to shed light on the motivations of everyone around these decisions, but we also extend the simple cost-benefit decision model for the context.

2.2 The Significance of Risk Perception

For security decisions, the literature shows us that risk perception, specifically a user's idea about the possible negative outcomes resulting from their decsions is key towards understanding security related behavior. Howe's 2012 survey of work about human psychology in computer security identified that security risks and risk perceptions were central considerations for many researchers [14]. Studies that have investigated mental models, such as Camp's 2006, Asgharpour et al.'s 2007, and Kang et al.'s 2015 works as well as other studies that looked directly at risk perceptions in different contexts, all focus on the importance of risk in the very design of their studies [4, 3, 16, 9, 11].

Some researchers have gone further and have tried to alter risk perceptions to improve communication and/or motivations. Harbach et al.'s 2014 work that appeared in CHI leveraged personal information to highlight the effect of Android permissions on user's data [10]. This was meant to alter their perception of the risks associated with each permission they are asked to grant, hopefully making them realize what exactly is at stake. The study found that users made more privacy-conscious decisions when presented with such information during app installation.

Since the perception of risk in particular has been repeatedly highlighted in work investigating security motivations, our study separates "cost" into explicit cost/inconvenience (e.g., time, money) and risk to provide a fuller picture of participants' perceptions and motivations.

2.3 Social Motivation

Though risk perception is intrinsically linked with security decisions, we also add another component absent from many other studies on this subject. Social motivations are integral towards voluntary compliance. Tyler's 2010 book "Why We Cooperate" details his theory on human motivation and cooperation [27]. In short, he argues that social motivations (i.e., motivations driven by values or wanting to help/please others) are much stronger and longer lasting than instrumental motivations (i.e., motivations related towards gaining material reward or avoiding material cost). Tyler presents his theory in contrast to the view of social motivations as simply a kind of instrumental motivation. Rather than trying to gain a future material benefit from

someone, Tyler says people who act in a socially positive way do so because they're motivated by their existing social connections.

The importance of social motivation is not new. Prosocial behaviors, as they are sometimes called have been studied for decades, being investigated all levels (i.e., individual, small, and large groups) and in many contexts [20].

Though Tyler and many others have theorized on the source of prosocial behavior and by extension social motivations, that debate is beyond the scope of this work. We simply accept that social motivations, regardless of the biological or psychological source, are important to human decision-making. Thus, our study considers participants' motivations with regards to other users, if any.

Ours is not the first work to acknowledge the importance of social considerations in technology decision-making. In SOUPS 2014, Das et al. found that social motivations play a role in cyber-security behavior [6]. Specifically, they found that observability of a secure behavior was a "key enabler of socially triggered behavior change," showing that social motivations could be important to technology decisions. The authors showed that users could be better motivated to act securely online if their peers would know the decisions they were making.

2.4 "Good" Actors and "Bad" Actors

Though Herley may be right and users may be properly assessing the computer security situation when they make what seems to be poor decisions, there is evidence in the literature suggesting that experts and average users do think and act differently when it comes to computer security. Two recent reports that support this statement appeared in SOUPS 2015. One, Ion et al.'s "No one can hack my mind..." showed that experts and regular users reported different behaviors when asked which they think are the best for staying safe, showing a divide in thinking [15]. The study also found that experts reported different security behaviors than non-experts. Additionally, another SOUPS 2015 work, "My Data Just Goes Everywhere" by Kang et al. found that mental models of computer security and privacy were different, specifically that average users had simpler models than expert users, again showing a difference in thinking between experts and everyone else [16]. The authors further found that more detailed models enabled experts to articulate more privacy threats, the first step towards avoiding them. That said, Kang also found that there was no direct correlation between participants' technical background and the actions they took to control their privacy, indicating that even those who should know better sometimes behave insecurely.

Though there are many documented differences between experts and average users, there is also substantial evidence that an expert is not necessarily a "good" actor. Our study wants to examine the difference in motivation between "good" and "bad" actors, which in this context are those who adhere to secure behavior and those who do not, respectively. As such, rather than compare experts with non-experts, our study compares those who report following common security advice with those who report not following such advice.

Combing all these concepts, the authors developed a web-study to gain insight into many aspects of why some users may follow computer security recommendations while others ignore them. The results of this study transcend prior work on this topic by collecting and analyzing a large dataset containing a sample of users' self-reported motivations for following or not following a broad range of security advice.

3. METHODS

Our study design incorporates both quantitative and qualitative methods to help outline differences between users on the topic of four instances of security advice, which are as follows:

1. Keeping your software up to date
2. Using a password manager
3. Using two-factor authentication
4. Changing passwords frequently

1-3 are commonly recommended by computer security experts to help users stay safe. Advice 4 is a common folk advice that isn't necessarily recommended by experts. All are extracted from Ion et al.'s 2015 work [15]. For each, we formulate two groups of users, one that follows (who uses the tool or does the action) and one that does not follow (does not use the tool or do the action). We are interested in comparing these groups because we want to understand why there is a decision gap between otherwise similar users, which would help towards identifying ways to encourage better online behavior among more of the Internet-using population. To help in describing the study, we will refer to samples of users who follow each advice as "Yes" groups for those respective advices, while we will refer to samples of users that do not follow each as "No" groups.

As explained in the prior section, we use a rational choice perspective to frame our study. Using cues from multiple recent works [4, 3, 16], we extend the traditional cost/benefit analysis to include perception of risk. Finally, we consider the social aspect of each decision as well, also inspired by recent literature [27, 6]. This study has 12 variables we investigate, named as follows:

1. *Individual Benefit of Following*
2. *Social Benefit of Following*
3. *Individual Cost/Inconvenience of Following*
4. *Social Cost/Inconvenience of Following*
5. *Individual Risk of Following*
6. *Social Risk of Following*
7. *Individual Benefit of Not Following*
8. *Social Benefit of Not Following*
9. *Individual Cost/Inconvenience of Not Following*
10. *Social Cost/Inconvenience of Not Following*
11. *Individual Risk of Not Following*
12. *Social Risk of Not Following*

Since Yes groups assumedly follow the advice and No groups report not following the advice, the same survey question phrasing could not be used to define each variable for both groups. Thus, we must compare responses to a slightly different question from each in our analysis. In other words, we must contrast what those who follow say their experience is to what those who do not follow expect their experience to be if they *did* follow. Specifically, variables 1-6 are defined using the following phrasings:

Yes How much would you say [you | users of other computers] are [benefited | cost or inconvenienced | put at risk] by you [following the advice]?

No How much would you say [you | users of other computers] would be [benefited | cost or inconvenienced | put at risk] if you did [follow the advice]?

Note that variables 1-6 are defined for Yes groups using the phrase highlighted above while the same variables are defined with the other phrase for No groups. The portion of the phrasings above that are separated by vertical bars and/or in brackets are the wordings used to form the question for each of the variables 1-6 we test, as appropriate. For example, "you" is used to replace the first bracket for *Individual* variables, while "users of other computers" is used for *Social* variables. The second brackets are likewise replaced for variables that ask about benefits, costs/inconveniences, and risks, respectively. Finally, "follow(ing) the advice" is replaced as appropriate for each advice that we test in the surveys (e.g., the 2FA Follow groups' was replaced with "using/use two-factor authentication," etc.).

Similarly, variables 7-12 are defined using the following phrasings:

Yes How much would you say [you | users of other computers] would be [benefited | cost or inconvenienced | put at risk] if you did not [follow the advice]?

No How much would you say [you | users of other computers] are [benefited | cost or inconvenienced | put at risk] by you not [following the advice]?

Instruments for these variables are created in the same fashion as described above.

As mentioned, each variable is defined in a slightly different format for Yes and No groups, meaning our analysis will compare ratings that are more/less hypothetical depending on the group. For example, considering the *Individual Benefit of Following*, Yes groups' reported benefits will be compared with the benefits No groups report they *would* get from following the advice. Though there may seem to be an issue with comparing hypothetical ratings to more grounded reports, the goal of this work is to identify the possible gaps in perceptions between those who follow security advice and those who do not. For many decisions, but particularly in the contexts examined in this study, a user must imagine at least some hypotheticals when considering whether to follow an advice or not, since the user may be pondering a behavior they have not practiced in the past. We hope to identify the skewed or biased perceptions users may have about the possible outcomes, thus it is valuable to compare the reported effects from followers of an advice with the projected effects from those who do not currently follow. This requires comparing some more hypothetical ratings with those that are more grounded.

Surveys containing the instruments described were created for each group (Yes/No) for each of the 4 tested pieces of advice. A qualitative question asking survey-takers *why* they chose to follow or not follow the target advice (i.e., "Please explain in a few sentences why you choose to (not) [follow the advice].") was also included. The qualitative question was shown first, alone on a separate page in all surveys to avoid biasing the open-ended responses towards our overall study framework as seen in the structure of other survey instruments (i.e., the focus on benefits/costs/risks). On the next survey page participants were asked about benefits/costs/risks of the actual target decision they reportedly made, followed by the benefits/costs/risks they *would* expect if they made the opposite of their decision on the final page. Survey templates for both groups, showing the order of questions, can be seen in the Appendix.

3.1 Sampling Methodology

Participants were recruited with a single Mechanical Turk posting that showed the information sheet for the study and directed interested users to a University-hosted Qualtrics survey that asked basic demographic questions and 4 screening questions to be used to assemble the Yes and No groups. Participants who responded to the screening survey were compensated $0.25 for their time and effort. The full screening survey can be seen in the Appendix.

After collecting the screening data, samples of 50 participants were assembled into 8 groups (one for the Yes and one for the No groups for each advice). Unique and independent Yes and No groups were formed by randomly selecting participants who reported "Yes" or "No," respectively to the screening question of whether or not they follow each advice.

Each group of participants was contacted with their corresponding group survey. Participants were contacted with a link through Mechanical Turk's messaging system that directed them to a new posting with the same information sheet as before, but this time a link to the appropriate survey. Participants were informed that they could take this survey if they wanted, but were under no obligation to reply. If they chose to answer, they were compensated another $4 for their time and effort on the longer survey.

3.2 Coding Methodology for Qualitative Data

To facilitate useful analysis of the qualitative data collected, we adopted a Grounded Theory approach to developing our codebook and coding our data [25]. The codebook was developed by the lead researcher, with the addition of some codes generated during analysis. Deductive codes, based on the study design and pertinent literature, along with the structure of the codebook were developed before data collection began. The focus here was on broad concepts like "avoid risk" or "increase security" since context specific codes could be best developed inductively, while looking at the data. There were seven deductive codes developed.

When data was collected, a random sample of one third of all comments from each group was selected and used to develop inductive codes by the lead researcher that focused on more specific concerns extracted from user comments. Some examples of codes developed though inductive coding are "I don't want to" and "increase financial security," showing the range of reasons given by participants. Since reasons between groups and advices varied, most of these codes were not broadly applicable, but some were. For example, "Low/no risk/Don't care if hacked" is an inductive code that was applied many times for several instances of advice. A total of 32 inductive codes were created for all groups. These codes (deductive + inductive) were used as

the codebook by another researcher who was less involved in the study and its design than the lead investigator. The same codebook was used to analyze all qualitative data. In the process of coding, several more tags were created from patterns found in the full samples, which were added to the codebook. Twenty-four of these codes were created.

Using the methodology described in this section, we collected a sample of active Internet users' motivations to follow or not follow computer security advice, which we analyze in the next section.

4. EVALUATION

To drive our analysis, we formulate the following hypotheses related the question "why do some follow security advice, while others do not?"

- H-1a For all decisions, the *Benefits of Following* will be seen as higher by the Yes groups compared to the No groups.
- H-1b For all decisions, the *Benefits of Not Following* will be seen as higher by the No groups compared to the Yes groups.
- H-2a For all decisions, the *Risks of Not Following* will be seen as higher by the Yes groups compared to the No groups.
- H-2b For some decisions, the *Risks of Following* will be seen as higher by the No groups compared to the Yes groups.
- H-3a For all decisions, the *Costs of Not Following* will be seen as higher by the Yes groups compared to the No groups.
- H-3b For all decisions, the *Costs of Following* will be seen as higher by the No groups compared to the Yes groups.
- H-4a Those who follow each advice will do so, generally, to increase their security and/or for convenience purposes.
- H-4b Those who do not follow each advice will do so, generally, to avoid a cost/inconvenience or due to confidence in current behavior (i.e., they might know they should change, but don't want to).
- H-5 *Social* considerations will be lower than *Individual* concerns for all decisions.

These predictions are based on prior findings and intuition. With hypotheses 1-3, we contend that participants will rate the benefits/costs/risks of their decision in a way that justifies their decision. For example, it is likely that each group will see the benefits gained and risks avoided by their decision as higher than if they had made the opposite. We expect the reasons given for these decisions will center on security and convenience, as these are the two core things at stake in many of these cases. Finally, the magnitude of social concerns is expected to be lower than individual concerns due to the nature of computing, which physically separates individuals, possibly obscuring how one's decisions affect others online. Please note, Hypothesis 2b is expected to only apply to some of the tested advice. This divergence compared to the other hypotheses is based on clues from data collected by the authors for prior studies about user decisions in the contexts of software updates, using 2FA, and using a password manager.

Before we demonstrate how these hypotheses are supported by our data, we first describe the overall sample collected and each group in terms of demographics. Once the make-up of our participants is established, we use our hypotheses to guide the rest of our evaluation.

4.1 Sample Details

As explained in Section 3, we collected an initial sample from Mechanical Turk using a short screening survey. A total of 805 participants enrolled in this step, but not all were considered for inclusion in groups to be contacted with follow-up surveys. We removed participants with incomplete answers on the screening survey and those who did not own a computer of their own from the eligible list, which reduced the pool to 764.

59% of the 764 are male, while 41% report female as their gender. The overall average age of participants is 34 years old. When asked how often they use the computer, 96% report "Often" or "All the time." The average general computer expertise rating is 4.15, while the average rating for computer security is 3.6. In both cases participants were simply asked "How would you rate your [general computer | computer security] expertise?" Both are measured on a 5-point Likert scale, anchored at 1-Very Poor and 5-Very Good. Though the instruments for measuring expertise are broadly defined, which could result in some level of error from a "true" measure, our approach was deliberate in order to ascertain the participants' general confidence in their computer and security knowledge and proficiency. The statistics are merely used to describe the sample collected and did not influence group forming or analysis other than attempting to control the variables between groups, where possible.

These statistics are not representative of the general population due to the nature of Mechanical Turk and the voluntary recruitment method used. That said, responses to the screening questions used for grouping show similar statistics as reported in prior studies [15]. Adoption rates for some advice were seemingly higher than would be expected for the general population, an effect that could also be attributed to the nature of the Mechanical Turk population or self-selection in the recruitment methods. Summaries of responses to grouping questions from the full sample of 764 can be seen in Table 1. In all, our sample represents a group of active computer users who generally rate their computer and security proficiency as higher than average, but are not all followers of the tested advice.

We formed 8 randomly selected, independent, unique groups of 50 participants each from the full sample initially collected. A participant is considered eligible for a group if they are not already in another group and exhibit the group's target behavior. For example, only participant who answered "Yes" to the question "Do you keep your computer's software up to date?" were considered eligible for the Yes group for the updating advice. The groups of 50 were gender-balanced so that 25 eligible males and 25 eligible females were contacted for each group. One group, those who do not keep their software up to date, only had 47 eligible participants out of the total pool of 764.

Not all participants contacted for each group responded. All groups ended up with 30-40 participants, which are used for this analysis. Details about the profile of each group sample can be seen in Table 2.

	Yes	No	I Don't Know
Do you keep your computer's software up to date?	701 (92%)	47 (6%)	15 (2%)
Do you use a password manager (e.g., LastPass, OnePass, KeePass) to manage your online account passwords?	157 (21%)	599 (78%)	8 (1%)
Do you use two-factor authentication (e.g., 2-Step Verification) for at least one of your online accounts?	471 (62%)	210 (28%)	81 (10%)
Do you change your passwords frequently?	311 (41%)	446 (58%)	5 (1%)

Table 1: Response frequencies (and rates) from all initial participants who own their own computer and completed the full screening survey (n=764) for each question used to form groups.

All group samples are similar on all demographic questions except self-rated computer security expertise, which had significant differences between groups when tested using a Kruskal-Wallis test [19]. Self-rated security expertise (i.e., "How would you rate your computer security expertise?") is lower for some No groups (i.e., update, changing passwords). Tests of the correlation between participants' rating for security expertise and their responses to survey instruments for all our variables using Spearman's correlation coefficient [2] resulted in no strongly significant values ($\forall$, $p > 0.05$, except *Individual Benefit of Following* where $p = 0.045$). This suggests that though there are slight differences between some Yes and No groups for self-rated security expertise, security expertise itself is not a good predictor of most perceptions. Essentially, as best as we can measure, the groups we compare are similar in most respects, security expertise being a notable exception, but even this difference is only apparent between some groups. Despite overall demographic similarity, we find differences in perceptions about these decisions in follow-up data.

4.2 Differences in Perception

Prior work and the intuition of the authors led to this study's focus on the cost/benefit analysis around these security advices. Regardless of which group's perceptions are more in line with reality, something that is mostly out of the view of this study, it is very likely that each group views the benefits, costs, and risks involved in the decision as different, which could at least in part be leading to the divergence in behavior.

Our first three hypotheses each focus on one of the three tenets of our study framework: benefit, cost, or risk. For all three, the guiding principle is that those who follow each advice are expected to have perceptions that are more supportive of adhering to the advice than the No groups. Please note that though only significant statistical results are detailed in this section due to space constraints. The results of all tests performed for this section can be found in the Appendix.

4.2.1 Benefits

As a core component of most rational decision models, it is natural to look at the benefits of a decision as perceived by those who are asked to make the decision. Specifically, for one to convince a person to do something, one must convince them that it is in their interests to do it. Through our design, we look at two kinds of benefits: the *Benefits of Following* (the security advice) and the *Benefits of Not Following* (the advice).

As explained in Section 3, each variable is defined using a single survey instrument that measures the variable on a 4-point Likert scale. Summaries of ratings for *Individual Benefit of Following* from each group, along with the results of a Mann-Whitney U-Test [18] comparing the response distributions of each Yes and No group can be seen in Table 3. Mann-Whitney U-Tests are appropriate for our data because the responses are independent and in the form of an ordinal scale. The test is non-parametric and measures if one distribution has a significantly higher median than the other, which would indicate, in our case, that one group rated the variable significantly higher than the other group. To analyze effect size, we use Cohen's d defined using the U-Test's Z score, divided by the square root of the number of samples compared [23].

	Yes	No	U-Test		
	Avg.(Med.)	Avg.(Med.)	U	Sig.	d
Upd.	3.77(4)	2.97(3)	274.5	<0.001	0.51
P.M.	3.78(4)	2.50(2.5)	154.5	<0.001	0.73
2FA	3.71(4)	2.90(3)	243.5	<0.001	0.49
Chg.P.	3.47(4)	2.53(3)	256	<0.001	0.57

Table 3: Rating summaries for *Individual Benefit of Following* for each group with U-Tests comparing the distribution between each Yes (those who follow the advice) and No (those who do not follow) groups. Effect size is measured with Cohen's *d*.

As can be seen in Table 3, for all advices, the Yes group rate their perceived benefit of following the advice as significantly higher than the No group, with most effects measuring "medium" (0.5) and one approaching "large" (0.8). This is unsurprising for the Yes groups since we expect that they are making a decision that they at least think benefits them. What's interesting here is the significantly lower ratings given by the No groups when asked to project the benefit they expected to receive from making the opposite decision of what they reported. As prior work has suggested [12, 13], these results support the idea that, at least in the eyes of some computer users, following security advice may just not be beneficial. This finding also supports our Hypothesis 1a, "For all decisions, the *Benefits of Following* will be seen as higher by the Yes groups compared to the No groups." Interestingly, ratings for *Social Benefit of Following* are not significantly different between groups for any advice, indicating that both see the benefits to "users of other computers" from each secure behavior as about the same. Of course, it could be that our samples are too small to show a significant effect and/or participants had a hard time conceptualizing the social benefits.

If one is interested in motivating more adherence to these and similar advices, this result suggests a gap between some users in how much benefit they see in adhering to these elements of advice. Addressing this gap through informational campaigns or other interventions may help, but providing

Advice	Group	n	Gender		Age		Comp. Expertise		Sec. Expertise		How Often Use Comp.	
			Male	*Female*	*Avg.*	*St.D.*	*Avg.*	*St.D.*	*Avg.*	*St.D.*	*Avg.*	*St.D.*
Update	Yes	39	20	19	38.4	14	4.15	0.7	3.56	0.8	4.79	0.4
	No	30	12	18	35.8	11	3.77	0.8	2.93	0.6	4.4	0.9
Password Manager	Yes	41	19	22	33.2	8.7	4.24	0.6	3.63	0.9	4.61	0.4
	No	38	16	22	34.0	9.7	4.3	0.7	3.50	0.7	4.79	0.4
2FA	Yes	36	20	16	36.6	13	4.31	0.7	3.86	0.9	4.69	0.5
	No	31	19	12	32.9	9	4.26	0.7	3.77	0.7	4.58	0.6
Change Passwords	Yes	37	20	17	36.0	10	4.22	0.6	3.78	0.8	4.73	0.6
	No	38	19	19	34.1	9.6	4.05	0.7	3.39	0.8	4.68	0.5

Table 2: Sample demographics for all groups used in this paper's analysis. "Comp[uter] Expertise", "Sec[urity] Expertise", and "How Often [Do You] Use [the] Comp[uter]?" are all measured on a 5-point Likert scale. The expertise questions are anchored from 1 = Very Poor to 5 = Very Good. The final question is anchored 1 = Never to 5 = All the Time, with the most common responses being "Often" or "All the Time."

better security tools, options, and education could go further towards increasing the adoption of secure behavior.

	Yes	No	U-Test		
	Avg.(Med.)	Avg.(Med.)	*U*	Sig.	*d*
Upd.	1.51(1)	2.13(2)	347.5	0.002	0.38
P.M.	1.68(1)	2.70(3)	302	<0.001	0.49
2FA	1.59(1.5)	2.62(3)	161.5	<0.001	0.61
Chg.P.	1.70(2)	3.03(3)	176	<0.001	0.66

Table 4: Rating summaries for *Individual Benefit of Not Following* for each group with U-Tests comparing the distribution between each Yes (those who follow the advice) and No (those who do not follow) groups. Effect size is measured with Cohen's *d*.

On a similar note, we do find significant differences between Yes and No groups' ratings on the variable *Individual Benefit of Not Following*, which supports our Hypothesis 1b. For updating, the effect is somewhat low, though still well above the "small" threshold (0.2), but other advice has solidly "medium" effects. Like before, there are no significant differences for *Social Benefits of Not Following*. As seen in Table 4, No groups consistently self-rate the benefits they receive from not following as significantly higher than the benefits the Yes groups' participants project they would receive from altering their behavior (i.e., to no longer following the advice). Like the ratings for *Individual Benefits of Following*, it should not be all too surprising that participants rate the benefits of their decision highly. If they thought the benefits were low, they likely would not be making the decision they claim they are. Still, for benefits, there is a perceptions gap when it comes to not following, as much as there is a perceptions gap for following. If those who do not behave securely see a lot of benefit in doing so, that must be addressed to alter their actions if so desired.

4.2.2 Risks

In addition to benefits, we also look at ratings of risk for more fine-grained insight into participants' considerations with respect to the tested advice. In the realm of security behavior, risk perception is a particularly important component to individuals' decisions as many behaviors are explicitly done to protect against a risk.

First, we analyze the perceptions of *Risks of Not Following*, covered by Hypothesis 2a. Table 5 shows the summaries for ratings to both *Individual Risk of Not Following* and *Social Risk of Not Following* along with U-Tests comparing each Yes groups' distribution with its corresponding No groups' distribution. Cohen's d is used to interpret effect size.

		Yes	No	U-Test		
		Avg.(Med.)	Avg.(Med.)	*U*	Sig.	*d*
Individ.	**Upd.**	3.42(4)	2.77(3)	336.5	0.002	0.37
	P.M.	2.88(3)	1.80(2)	302.5	<0.001	0.52
	2FA	3.42(3)	2.61(3)	243.5	<0.001	0.53
	Chg.P.	3.14(3)	2.63(3)	440.5	0.003	0.34
Social	**Upd.**	2.67(3)	1.76(1)	262.5	<0.001	0.44
	P.M.	1.92(2)	1.29(1)	409	0.002	0.37
	2FA	2.48(3)	1.79(2)	289	0.013	0.32
	Chg.P.	1.70(1)	1.29(1)	483	0.044	0.24

Table 5: Rating summaries for *Individ[ual] Risk of Not Following* and *Social Risk of Not Following* for each group with U-Tests comparing the distribution between each Yes (those who follow the advice) and No (those who do not follow) groups. Effect size is measured with Cohen's *d*.

Like with benefits, it is natural that those who follow each advice would see the risks of stopping that behavior as high since they are likely following to protect themselves from risks. In all cases, across both individual and social concerns, the Yes groups consistently rate the risks of no longer following the group's target advice as higher than the risks reported by those who already do not follow the target advice. This supports our hypothesis 2a, which states "For all decisions, the *Risks of Not Following* will be seen as higher by the Yes groups compared to the No groups." Effect sizes were sometimes low in these comparisons, but generally "medium."

As we stated before, we are not attempting to test the correctness of either group's perceptions, which would require data different than what was collected for this study. With that in mind, there is still much to learn from this result. There are many ways of interpreting the gap in risk perception between groups. On one hand, the risks could be low and those who follow the advices are exaggerating, as shown by the ratings from individuals who are *actually* at risk (No groups). In this view, one must assume that those who do not follow each advice are correctly experiencing the threat. This is where the alternative view comes in: it's possible, some may say, that those who do not follow have just not yet been affected, causing them to underestimate the risk of their behavior.

The existence of this perception gap calls for more research. In particular, as noted in prior work [12], identifying the

reality of risk faced by users is key to identifying which group is "correct." Knowing this can help further inform how (or where) to motivate behaviors that will increase security for users in a real way. Regardless, the risk perception gap can still be approached using the current state of knowledge. The authors also contend that these results could suggest that the targets of interventions (i.e., users who do not follow security advice) do not view their behavior as risky, despite the large amount of information and advice available online that should convince them otherwise. Thus, if the goal is altering these individual's decision, doing so may require new or alternative approaches. That said, the lower effect sizes compared to the other results presented up until this point could mean less of a gap here than for benefits.

It is also important to consider the perceived *Risks of Following* each advice since some may consider the tool or behavior risky. There are only strongly significant differences between groups on *Individual Risk of Following* for one advice: using a password managers ($U = 342.5$, $p < 0.001$. $d = 0.49$). The distribution of responses for the password manager Yes and No groups can be seen in Figure 1, which highlights the divergence in responses between the two groups. One other advice, changing passwords frequently shows a weaker, but significant difference for *Individual Risk of Following*($U = 498.5$, $p = 0.014$, $d = 0.28$). No other advice shows any significance in differences between groups on this variable. *Social Risk of Following* shows no significant differences for any advice.

Thus, our Hypothesis 2b is only partially supported by the data, particularly in the case of using password managers. Thinking about the function of a password manager in particular brings some insight. Password managers centralize passwords, an action some participants may view as risky, therefore increasing perceptions of risk of using the tool, especially among those who don't use it. Section 4.3.2 provides more information about possible reasons to explain this divergence.

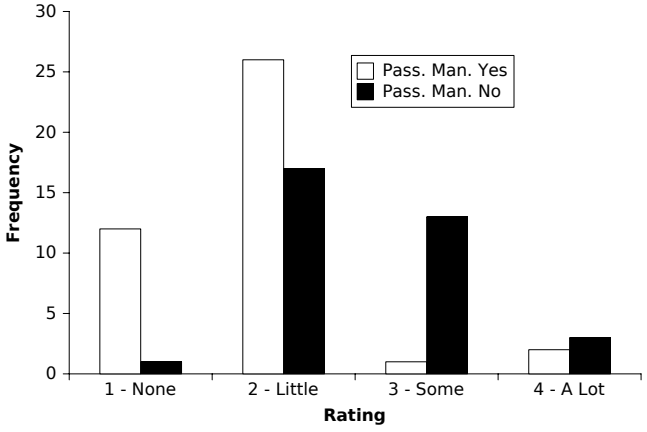


Figure 1: Response distributions representing the variable *Individual Risk of Following* for the password manager Yes (use) and No (don't use) groups.

4.2.3 Costs

Finally, besides benefits and risk, many decisions have some kind of cost associated with them. For example, updating one's system may take time in the form of a restart, or using 2FA on your phone may cost money in the form of charges for text messages. These costs will certainly play a role in the decision being made, thus we examine the cost ratings along with ratings of benefit and risk. Table 6 shows the summaries for the variable *Cost of Not Following* (both *Individual* and *Social*) along with U-Test results comparing the distributions of responses from the Yes group of each advice with the distribution of its corresponding No counterpart.

		Yes	No	U-Test		
		Avg.(Med.)	Avg.(Med.)	U	Sig.	d
Individ.	**Upd.**	2.95(3)	2.00(2)	247.5	<0.001	0.48
	P.M.	3.15(3)	1.75(1)	244.5	<0.001	0.60
	2FA	1.76(1)	1.57(1)	446.5	0.451	0.09
	Chg.P.	2.28(3)	1.61(1)	425.5	0.003	0.35
Social	**Upd.**	2.32(2)	1.59(1)	248	0.001	0.41
	P.M.	1.84(1)	1.03(1)	354	<0.001	0.49
	2FA	1.69(1)	1.41(1)	343	0.356	0.12
	Chg.P.	1.50(1)	1.24(1)	525.5	0.174	0.16

Table 6: Rating summaries for *Individ[ual] Cost of Not Following* and *Social Cost of Not Following* for each group with U-Tests comparing the distribution between each Yes (those who follow the advice) and No (those who do not follow) groups. Effect size is measured with Cohen's *d*.

As can be seen in Table 6, some advice has significant differences in how participants in each group rate the costs of not following the advice. Updating and using a password manager shows the strongest differences, with there being divergence on both the individual and social forms of the variable. Effect sizes are "medium" in each case. Changing passwords also shows significant differences, but only for the *Individual Cost of Not Following*. Here the effect size is smaller than for other differences in cost. For updating, many users may see a benefit in terms of performance when updating and so see not updating as incurring them a cost (i.e., in performance). Similarly, password managers help with things like account creation and log in, so they provide a convenience benefit in addition to security benefit. Thus, it is likely that those who stopped using a password manager would feel a cost in terms of time and/or effort. What's interesting is that there are differences between the groups on costs for some of the advice, which could mean that there is a real benefit incurred by updating and/or using a password manager that is not known until trying. Overall, these results somewhat support our Hypothesis 3a, just not for all cases as predicted.

There is only one strongly significant result when comparing ratings from the Yes and No groups for the variable *Cost of Following*, which is *Individual Cost of Following* for changing passwords. The No group rates this significantly higher than the Yes group, with averages of 2.97 and 2.35 respectively (U=449.5, p=0.005, d=0.33). Two other elements of advice also have weaker, but signifigant differences on this variable: using a password manage (U=533, p=0.011, d=0.28) and using 2FA (U=405.5, p=0.036, d=0.26). For the individual cost of updating and social phrasings of the *Cost of Following* for all pieces of advice, differences are not significant. Thus, we only have limited data to directly support the hypothesis "For all decisions, the *Costs of Following* will be seen as higher by the No groups compared to the Yes

groups."

Though the ratings from participants provide a window into their minds, the quantitative data is limited in richness due to its nature. As such, we supplement our numerical data with open-ended responses as to *why* participants made the decision they did, as explained in Section 3. Analysis of these comments helps answer some of the questions presented by quantitative analysis.

4.3 Why Do They Do What They Do?

In addition to finding perception differences, this study hopes to shed some light on the reasons people have for their decisions, which can help us explain some of the gaps. Hypotheses 4a and 4b deal with this aspect of the study and are supported using analysis of the qualitative data. Responses are coded using the process described in Section 3.2. To examine how the reasons provided by each group differ, and how well our hypotheses predict our results, in this section, we present the most prominent and noteworthy codes assigned to comments from each group.

4.3.1 Updating

Keeping one's software up to date is one of the most commonly recommended practices from security experts to keep data and machines protected. When comparing the reasons for each decision (i.e., to update or not), we find stark differences, but also some interesting similarities.

First, a large number of those who update said they do so for security purposes. Approximately 49% of all 39 comments received from Yes group participants mention increasing some kind of security. Additionally, good performance, specifically avoiding bugs and software issues is a chief concern for the group of participants who update. Twenty-two of 39 comments mention avoiding bugs and/or issues, making up 56% of the comments from this group. Ten (26%) comments mention wanting to get the most recent changes, while 7 (18%) indicate a desire to avoid malware specifically.

Unsurprisingly, these codes were not assigned to any comments from the No group. Instead, common concerns for that group are getting a convenience and/or avoiding an inconvenience, not finding a need [to update], or not being willing to put in the effort involved. Five of 30 comments (15%) mention not needing to update, while another five comments mention being too lazy to update and/or updates being too much work to apply. 23% of the comments allude to or mention avoiding an inconvenience and/or getting a convenience by not applying updates. Interestingly, 13% of the comments from those who do update also bring up avoiding an inconvenience/getting a convenience. It would seem that both groups see some convenience in their decision, be it through avoiding undue effort, as in the case for the No group, or through getting the latest features, as for the Yes group.

Those who do not update have many other specific reasons for their decision. Avoiding harm (3 comments), avoiding change (5 comments), and finding updates too frequent (4 comments) are also common reasons from the No group, showing the spread of concern among these individuals. By contrast, most of those who update report similar reasons (i.e., security, best features, avoid software faults) for their decision.

Looking at updating, both our hypotheses for this aspect of the study hold up. Hypothesis 4a states "Those who follow each advice will do so, generally, to increase their security and/or for convenience purposes," which is supported by the large number of comments from the Yes group saying they update to increase their security or to avoid software issues. Of those who do not update, many choose that route to get a convenience/avoid an inconvenience, supporting Hypothesis 4b. Many others also mention a confidence in their current approach by saying or suggesting they have no need to update. It should be noted that 3 comments bring up a specific bad update in the past as a reason for their update avoidance, so it's possible some participants' skepticism is warranted or, at least understandable from a rational decision standpoint, as suggested in prior work [29, 28].

4.3.2 Using a Password Manager

Password managers help create and manage passwords for online accounts by allowing automatic form filling, which alleviates the need for users to remember many, long, complex (and therefore secure) passwords. "Secure" password managers help increase overall privacy by affording users the ability to auto-generate and auto-fill hard-to-crack passwords on all their accounts. Recommended password managers encrypt the stored data to reduce the obvious security risk introduced by storing all passwords in a single, noticeable, predictable place. Password managers that do not encrypt passwords are generally considered insecure, but our study specifically asks participants if they use more secure password managers (e.g., LastPass).

Those who use a password manager report the convenience added by the tool (i.e., automatic form-filling) as a reason for using in an overwhelming majority of their comments. Thirty-seven of 40 comments (93%) from those who use a password manager mention the added convenience of the software. 55% of comments from the same group indicate the added security they get from using their password manager as a reason for their decision to use.

By contrast, 45% of those who do not use a password manager say they avoid them to avoid a security risk, showing that many in the No group feel that password managers are not worth the added benefit at log-in because they think the tool opens them up to attack. Twelve (32%) of 38 comments from the No group specifically mention avoiding centralizing their passwords as a reason not to use a password manager. Calling back to prior results from this study, these comments can shed some light on the significantly higher ratings for the *Individual Risk of Following* from those who do not use a password manager compared to those who do use one. It seems that a large proportion of those who do not use a password manager explicitly do not because they view the tool as a security risk.

Additionally, half of the comments mentions a confidence in the participant's current security/password mechanism. These approaches include remembering passwords (which could lead to insecure passwords used on websites due to cognitive limitations for individuals to remember log-ins) and writing passwords down in a "secure" place, which may seem satisfactory, but ignores risks from local threats and could also lead to bad passwords due to complacency.

Thinking to our Hypotheses 4a and 4b, these results sup-

port those predictions. Users of password managers report becoming users in a majority of cases examined because of the security and convenience they feel they get from their action, thus supporting 4a. On the other side, those who do not use a password manager in many cases do so because they feel their current method of password management is sufficient (i.e., confidence in current behavior), partially supporting 4b. That all said, password managers were different from the other advices we tested in that many non-users reported their impression of password managers as a fundamental risk as a reason for not using them. This is reflected in the qualitative data presented in Section 4.2.2.

4.3.3 Using 2FA

Two-factor authentication (2FA) is another common technique for increasing account security. In addition to a username and password, users of 2FA are sent a one-time password through email, SMS, etc. that is used in the specific instance of that log in. The addition of the one-time password, which is only good for the single log in attempt, increases security by adding another factor (of authentication) that must be stolen by a would-be attacker. If a hacker, for example, gets access to your user-name and password, by using 2FA, they would also need access to the account and/or device you use to receive your one-time passwords to be able to access your account.

A large proportion of the 36 comments from participants in the group who report using 2FA say they do so to increase their security (86%) and/or because it's safer than the alternatives they're aware of (61%). Additionally, 25% say they use 2FA because it "feels better" than not using 2FA. Overall, these comments suggest that 2FA users are strongly motivated by the security benefits they see in the technique. This should not be surprising as 2FA is less commonly used and is known for increasing security, so those who *do* use it are likely to be drawn by that prominent benefit.

On the flip side, 48% of the 31 comments from the 2FA No group say they do not use 2FA to avoid an inconvenience and 23% mention avoiding a cost. In both cases, the most common cost and/or inconvenience is the need for a second factor, which slows log-in. Additionally, 26% of the comments mention that the participants' current approach is good enough, 19% say they do not see the risks of not using 2FA and/or don't care if they're hacked, and 13% allude to or say there is no need for using 2FA.

Like before, these findings broadly support Hypotheses 4a and 4b. The Yes group for 2FA greatly values the security they get from using 2FA, but unlike updating and using a password manager, none think 2FA offers them a convenience. Convenience or more specifically the avoidance of the inconvenience of 2FA is a chief concern among those who don't use 2FA. Not seeing a need to use 2FA and the idea that their current approach is good enough (compared to 2FA) also influence the No group.

4.3.4 Changing Passwords Frequently

Frequently changing passwords, though not a common advice from experts, is seen as a secure behavior in the eyes of many users [15], likely due to password changes being recommended in corporate environments and/or after a security breach. Changing passwords frequently is not likely to help protect an individual account, assuming all passwords used are of sufficient security. The security benefits come in when the attacker may have access to your *current* password, but by changing it, you thwart their attack.

Like the use of 2FA, those who frequently change their passwords commonly cite the added security they get from doing so, as was the case for 26 (72%) of 36 comments from the Yes group. 19% of the comments from this group specifically mention increased account security, and 22% mention avoiding theft and/or unauthorized access of their account. None mention a convenience increase as a reason for their decision to use.

For those who do not change their passwords frequently, also like 2FA, many (53% of 38 comments) say they do so to avoid an inconvenience. Other concerns, like confidence in their current approach (13%) and seeing a low risk of attack (18%) are also common reasons for not changing passwords. Unlike 2FA, though, many (39%) comments say they do not change passwords often because doing so is too hard to remember and/or their passwords would be too hard to remember if they did. Also, interestingly, 32% of comments from this group mention not having problems before as a reason not to start changing passwords (and therefore continue **not** changing passwords), while only 6% of those who don't use 2FA mentioned such a theme in their comment. It could be that, due to changing passwords being less "work" than using 2FA, participants who do not follow the advice feel more reason to justify their decision in another way, in this case using an argument of "if it ain't broke, don't fix it," while those who do not use 2FA feel justified in avoiding the somewhat substantial extra cost of enabling the feature.

The Yes group's focus on the perceived security benefits of changing passwords frequently supports, at least in part Hypothesis 4a. By worrying about the inconvenience of changing passwords frequently and not seeing much risk in their behavior, the comments from the No group also supports Hypothesis 4b.

4.3.5 Social Content in Comments

One very strong theme across all comments is the focus on the individual in the reasons given. Only 13 of 290 (4.5%) of all comments mention some social motivation behind the decision, all from Yes groups. This is in line with prior work showing the positive effect of social motivation around computer security [6], since the few comments that did mention a social motivation were all from participants that followed security advice. Examples of social motivations in comments include the desire to protect family/other users (5 comments), trust in developers (2 comments), acting on a friend's/family member's recommendation (4 comments), and concern for their place in the Internet/network in general (2 comments). With this lead and hints from prior work, we further investigate the individual/social motivation divide.

4.4 Individual vs. Social Concerns

As described in the Methods section, each component of our decision model is toned in both an individual and social context. All participants were asked for an individual and social rating for each component (i.e., benefit/cost/risk) related to following and not following the advice. Figure 2 shows the average overall rating (across all groups) for each variable in our study plotted together to contrast the difference between

averages for individual and social phrasings.

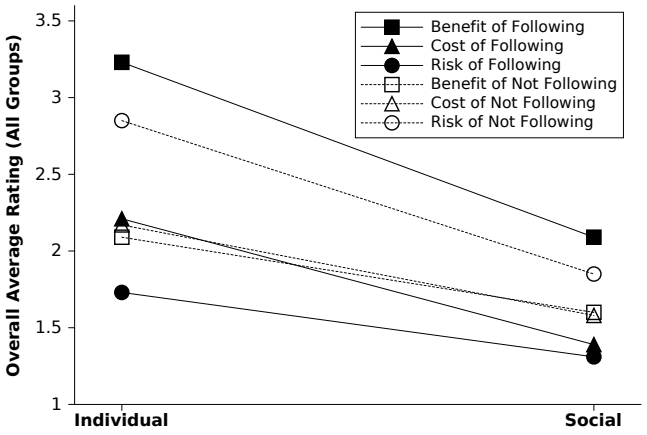


Figure 2: Plot of average overall ratings for each variable, arranged to show the consistently lower social ratings compared to individual ratings. A sign test of each variable pair (ind. vs. soc.) found significant ($p < 0.001$) differences for all variables.

As can be seen in Figure 2, for each variable, the individual phrasing average is higher than the corresponding social phrasing's average. To statistically test these differences, we use a sign test [7]. Put simply, the sign test determines if one variable from the pair tested is rated consistently higher than the other. A low p value for the sign test indicates that participants in the sample consistently rated one variable from each pair as higher than the other variable in the pair. We meet the assumptions for this test since our data is ordinal and observations from different participants are independent, thus the differences in their individual and social scores are also independent.

In the context of our study, we use the sign tests to determine if the differences demonstrated by the averages plotted in Figure 2 are representative of statistically significant and consistent differences between individual and social ratings, regardless of advice, aspect of decision, or context (i.e., following vs. not following the advice). Data was aggregated for each variable across all 8 groups, and the sign test compares *Individual* with *Social* ratings. For all pairs tested, we find strongly significant differences ($\forall$, $Z <$ -5, $p < 0.001$), indicating that ratings for individually phrased variables are consistently higher than the socially phrased version's ratings. Effect sizes measured using Cohen's d were greater than 0.5 for all tests except *Benefit of Not Following*, which was 0.36, indicating that the differences between groups could be considered "medium." Full results of these tests can be found in the Appendix.

Lower social ratings than individual indicate that most participants may give more consideration to how the option of following each tested security advice affects them than how it affects others. As prior work has indicated, social motivation are stronger regulators of behavior than individual motivations [27, 20]. Computer security is ripe for social considerations as one's security behavior can have an effect on other's security, especially if your behavior causes a breach of some kind. For example, if by not updating your operating system, your machine is infected and becomes a member of a malicious botnet, your decisions will have affected others when the botnet is used against websites or other web-services utilized by other computer users. Thus, increasing the strength of social considerations around computer security is not only possible, but preferable to focusing on individual considerations when trying to motivate good security behavior.

Though there are strong differences when aggregating, we also use sign tests to compare the *Individual* and *Social* ratings for each variable separated by group to see how the overall results hold up when looking at specific contexts. In most cases, the difference between individual and social ratings holds in significance. Only 23 of 48 tests have significant values greater than 0.001. Sixteen of those tests show weaker, but nonetheless significant differences, including 11 tests resulting in $p \leq 0.008$ and 5 additional tests returning $p < 0.04$. The remaining 7 cases do not show significant differences, but these are mostly *Benefit* and/or *Cost of Not Following*, which could be hard concepts for some participants to wrap their heads around. Additionally, larger samples may show stronger differences for these variables. These findings support our final hypothesis, "Social considerations will be lower than individual concerns for all decisions." As before, the full results of these tests can be found in the Appendix.

5. DISCUSSION AND LIMITATIONS

Overall, our findings enlighten in solving the problem of motivating secure behavior. Even as the science of security improves and new, better tools are released, the task of getting users to take up these tools and techniques will always exist. Discovering trends in perceptions around these decisions and more importantly using them to help develop strategies of persuasion are both key towards a more secure ecosystem.

Central is the propensity of each group to see the benefits of their decision as higher than their counterparts predict their benefits would be. Though unsurprising, it is important for those giving security advice to keep in mind that even though you, as an adherent to a behavior see certain benefits, others, particularly those who do not adhere are likely to not see the same benefits. Though this may suggest a simple solution is to better inform users about benefits (which assumes the No groups are wrong in their perceptions), prior work argues that such an approach is likely to fail [12, 13, 5], indicating that the users (those who do not follow the advices included) are at least aware of the benefits involved and do not need to be simply informed. Besides simple ignorance, there may be many other reasons for these perception gaps. It could be that some do not realize the value of the benefits, or the benefits are actually not as high as the Yes groups seem to think. These and other explanations for the differences in benefits require a different solution than simply disseminating information. Instead, the task calls for a nuanced, issue-tailored approach that addresses what users are likely thinking and what they actually experience to help them overcome the barriers to desired behavior.

Risk perception is important too, as to be expected in the realm of security decision-making. Like with benefits, we found that Yes and No groups felt differently about the risks they were protected from by following each behavior. It is very hard to know which group is more accurately estimating the risks involved as there is limited data on the

costs and risks experienced by an average, individual user. Though there is much data on the macro-level (e.g., number of attacks per year, accounts compromised every day, etc.), there has been no large-scale, regular data collection to give empirical and scientific power to statements about the danger of general online security risks to a particular user. As researchers, we try our best to estimate these risks, but without hard, consistent data, any advice we give is on some level speculative and based on our incomplete picture of what users face. Calls for this kind of data are not new, but have thus far gone unanswered.

The convergence of most participants' justifications for their decisions around the topics of security and/or (in)convenience is also notable as the convenience/security trade-off is a commonly discussed concept around computer security [26, 30, 21]. In general, many note that security requires some kind of inconvenience while taking a more convenient route will likely prove less secure. For example, it is much easier to make and manage a single account for a shared machine, but such a set-up makes activity and data from different users visible to others, resulting in less security than individual accounts. In some cases of our study, such as changing passwords frequently and using 2FA, most who followed the advice say they do so for a security benefit, while most who do not follow say their decision is to avoid an inconvenience, suggesting participants making these decisions are considering a security/convenience trade-off. Time was a very common theme, with participants citing a lack of time to follow the tested advice. As one non-updating participant put it: "I'm busy, dang it!"

Many No group comments express similar sentiments. Use of a password manager also plays into this paradigm, but shuffles it due to the specific functionality of password managers. Many of those who do not use password managers report avoiding the security risk of centralization as their concern with the tool, while many users cite the convenience benefit afforded by auto-login features. Updating is also reported to come with benefits (e.g., in the form of better software performance) that are appreciated by participants. These findings suggest that motivating more secure behavior could be done with better management of the convenience/security trade-off considerations being made for particular context.

Finally, our results show that individual rather than social concerns are rated higher in quantitative data and are more prominent in the qualitative data. Though the lack of social comments could be due to question wording (i.e., the open-ended question's phrasing may encourage responses biased towards individual concerns), the existence of several comments that **do** mention a social motivation and the quantitative results related to social vs. individual concerns both show that many participants are thinking predominately about themselves when making these decisions.

Psychology has long studied the occurrence of prosocial behavior [20], in no small part because such behavior is very beneficial to society as a whole and so society is inclined to encourage it in individuals where possible. Newer research has pointed to the power of social motivations [27]. If the social consciousness of these decisions could be increased, it is likely that some users will be motivated to follow despite the costs they may incur. Like before, more data on the real risks and ramifications of security threats and efficacy of various behaviors in protecting adherents is important here because knowing the social effects is key to properly adjusting user's perceptions, when necessary.

Our approach is not without its limitations. Though we were able to find statistically significant differences in many places, additional data could generate new findings or provide insight in existing results. In particular, larger and more varied samples could garner larger effect sizes than those reported in this study, which were generally "medium." In addition, examination of more advices and contexts (e.g., perceptions of benefits/risks/costs for specific kinds of devices) could also expand the picture. An expanded decision-making framework may provide more insight, but would likely require a larger study from the design presented and used here, introducing different limitations. Finally, as Mechanical Turk's user-base may not be representative of the general population, replication of this study with more samples would help generalize the findings.

6. CONCLUSION

Our results show differences in the perceptions of benefits, risks, and costs associated with decisions to adhere to a variety of security behaviors that are commonly recommended by experts. Both those who do and do not follow each advice report that their current decision gets them more benefit than if they changed. Those who follow rate the risks of changing their decision as much higher than the risks reported by those who do not follow. Costs of not following are also seen as higher by most that follow compared to those who do not. When looking into the reasons participants gave for their decisions, we find strong trends highlighting the convenience/security trade-off. The value of convenience in particular may be used to help motivate the use of security tools and techniques. Finally, we found that individual concerns are rated consistently higher than social concerns. Increasing social motivations could motivate more secure decision-making, according to theory from prior work [27].

Additional data regarding the real benefits/risks/costs of these and related contexts, not just perceptions of them are needed to help better paint the complete picture of what is happening in users' minds and address the gaps identified. Nonetheless, this study has provided insight into user motivation to guide future efforts towards the broader goals of usable security.

7. ACKNOWLEDGMENTS

The authors thank the reviewers and the paper's shepherd, Lujo Bauer for their helpful guidance with this paper. This work is supported by the National Science Foundation under Grant no. CNS-1343766 and by GAAAN Fellowship no. P200A130153. Any opinions, findings, or recommendations expressed are those of the authors and do not necessarily reflect the views of the funding agencies.

8. REFERENCES

[1] A. Adams and M. A. Sasse. Users are not the enemy. *Communications of the ACM*, 42(12):40–46, 1999.

[2] R. A. Armstrong and A. C. Hilton. Nonparametric correlation coefficients. *Statistical Analysis in Microbiology: Statnotes*, pages 91–94.

[3] F. Asgharpour, D. Liu, and L. J. Camp. Mental models of security risks. In *Financial Cryptography*

and Data Security, pages 367–377. Springer, 2007.

[4] L. J. Camp. Mental models of privacy and security. *Available at SSRN 922735*, 2006.

[5] N. Clarke, S. Furnell, G. Stewart, and D. Lacey. Death by a thousand facts: Criticising the technocratic approach to information security awareness. *Information Management & Computer Security*, 20(1):29–38, 2012.

[6] S. Das, T. H.-J. Kim, L. A. Dabbish, and J. I. Hong. The effect of social influence on security sensitivity. In *SOUPS*, pages 143–157, 2014.

[7] W. J. Dixon and A. M. Mood. The statistical sign test. *Journal of the American Statistical Association*, 41(236):557–566, 1946.

[8] D. Florencio and C. Herley. A large-scale study of web password habits. In *Proceedings of the 16th international conference on World Wide Web*, pages 657–666. ACM, 2007.

[9] M. Harbach, S. Fahl, and M. Smith. Who's afraid of which bad wolf? a survey of it security risk awareness. In *Computer Security Foundations Symposium (CSF), 2014 IEEE 27th*, pages 97–110. IEEE, 2014.

[10] M. Harbach, M. Hettig, S. Weber, and M. Smith. Using personal examples to improve risk communication for security and privacy decisions. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 2647–2656, New York, NY, USA, 2014. ACM.

[11] M. Harbach, E. von Zezschwitz, A. Fichter, A. De Luca, and M. Smith. It's a hard lock life: A field study of smartphone (un)locking behavior and risk perception. In *Symposium on Usable Privacy and Security (SOUPS 2014)*, pages 213–230, 2014.

[12] C. Herley. So long, and no thanks for the externalities: the rational rejection of security advice by users. In *Proceedings of the 2009 workshop on New security paradigms workshop*, pages 133–144. ACM, 2009.

[13] C. Herley. More is not the answer. *IEEE Security & Privacy*, (1):14–19, 2014.

[14] A. E. Howe, I. Ray, M. Roberts, M. Urbanska, and Z. Byrne. The psychology of security for the home computer user. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 209–223. IEEE, 2012.

[15] I. Ion, R. Reeder, and S. Consolvo. "... no one can hack my mind": Comparing expert and non-expert security practices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 327–346, 2015.

[16] R. Kang, L. Dabbish, N. Fruchter, and S. Kiesler. "my data just goes everywhere:" user mental models of the internet and implications for privacy and security. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 39–52, 2015.

[17] M. Madden and L. Rainie. Americans' attitudes about privacy, security and surveillance. Online at: `http://www.pewinternet.org/`, May 2015.

[18] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[19] P. E. McKight and J. Najab. Kruskal-wallis test. *Corsini Encyclopedia of Psychology*, 2010.

[20] L. A. Penner, J. F. Dovidio, J. A. Piliavin, and D. A. Schroeder. Prosocial behavior: Multilevel perspectives. *Annu. Rev. Psychol.*, 56:365–392, 2005.

[21] S. Prabhakar, S. Pankanti, and A. K. Jain. Biometric recognition: Security and privacy concerns. *IEEE Security & Privacy*, (2):33–42, 2003.

[22] L. Rainie, S. Kielsler, R. Kang, and M. Madden. Anonymity, privacy, and security online. Online at: `http://www.pewinternet.org/`, September 2013.

[23] R. Rosenthal, H. Cooper, and L. Hedges. Parametric measures of effect size. *The handbook of research synthesis*, pages 231–244, 1994.

[24] B. Rossler and R. D. Glasgow. *The value of privacy*. Polity, 2005.

[25] A. Strauss and J. Corbin. Grounded theory methodology. *Handbook of qualitative research*, pages 273–285, 1994.

[26] L. Tam, M. Glassman, and M. Vandenwauver. The psychology of password management: a tradeoff between security and convenience. *Behaviour & Information Technology*, 29(3):233–244, 2010.

[27] T. R. Tyler. *Why people cooperate: The role of social motivations*. Princeton University Press, 2010.

[28] K. E. Vaniea, E. Rader, and R. Wash. Betrayed by updates: How negative experiences affect future security. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 2671–2674, New York, NY, USA, 2014. ACM.

[29] R. Wash, E. Rader, K. Vaniea, and M. Rizor. Out of the loop: How automated software updates cause unintended security consequences. pages 89–104. USENIX Association, 2014.

[30] C. S. Weir, G. Douglas, M. Carruthers, and M. Jack. User perceptions of security, convenience and usability for ebanking authentication tokens. *Computers & Security*, 28(1):47–62, 2009.

APPENDIX

A. SURVEY INSTRUMENTS

Templates for all instruments used in this study are listed in the subsections below.

A.1 Intial Survey

The questions derived from thew following template were shown to 805 participants who initially enrolled in the study from Mechanical Turk.

1. What is your age?
2. What is your gender?
   - Male
   - Female
   - Other
3. Do you use a laptop or desktop computer that you or your family owns (i.e., not provided by school or work)?
   - Yes
   - No
4. How would you rate your general computer expertise?
   - Very Poor
   - Poor
   - Fair

_ Good
_ Very Good

5. How would you rate your computer security expertise?
   _ Very Poor
   _ Poor
   _ Fair
   _ Good
   _ Very Good
6. How often would you say you use the computer?
   _ Never
   _ Rarely
   _ Sometimes
   _ Often
   _ All the Time
7. Do you keep your computer's software up to date?
   _ Yes
   _ No
   _ I Don't Know
8. Do you use two-factor authentication (e.g., 2-Step Verification) for at least one of your online accounts?
   _ Yes
   _ No
   _ I Don't Know
9. Do you use a password manager (e.g., LastPass, OnePass, KeePass) to manage your online account passwords?
   _ Yes
   _ No
   _ I Don't Know
10. Do you change your passwords frequently?
    _ Yes
    _ No
    _ I Don't Know

A.2 Follow-Up Surveys

After groups were formed, the following templates were used to create surveys for each advice Yes and No group. To form each survey, replace [follow(ing) the advice] in the templates with each of the following phrases for the corresponding advice:

- **Update** - "keep(ing) your computer's software up to date"
- **Pass. Man.** - "us(e/ing) a password manager"
- **2FA** - "us(e/ing) two-factor authentication"
- **Change Pass.** - "chang(e/ing) your passwords frequently"

A.2.1 "Yes" Group Template

1. Please explain in a few sentences why you choose to [follow the advice].
2. How much would you say you are benefited by you [following the advice]?
   _ None
   _ Little
   _ Some
   _ A Lot
   _ Not Sure
3. How much would you users of other computers are benefited by you [following the advice]?
   _ None
   _ Little
   _ Some
   _ A Lot
   _ Not Sure
4. How much would you say you are cost or inconvenienced by you [following the advice]?
   _ None
   _ Little
   _ Some
   _ A Lot
   _ Not Sure
5. How much would you users of other computers are cost or inconvenienced by you [following the advice]?
   _ None
   _ Little
   _ Some
   _ A Lot
   _ Not Sure
6. How much would you say you are put at risk by you [following the advice]?
   _ None
   _ Little
   _ Some
   _ A Lot
   _ Not Sure
7. How much would you users of other computers are put at risk by you [following the advice]?
   _ None
   _ Little
   _ Some
   _ A Lot
   _ Not Sure
8. How much would you say you would be benefited if you did not [follow the advice]?
   _ None
   _ Little
   _ Some
   _ A Lot
   _ Not Sure
9. How much would you say users of other computers would be benefited if you did not [follow the advice]?
   _ None
   _ Little
   _ Some
   _ A Lot
   _ Not Sure
10. How much would you say you would be cost or inconvenienced if you did not [follow the advice]?
    _ None
    _ Little
    _ Some
    _ A Lot
    _ Not Sure
11. How much would you say users of other computers would be cost or inconvenienced if you did not [follow the advice]?
    _ None
    _ Little
    _ Some
    _ A Lot
    _ Not Sure

12. How much would you say you would be put at risk if you did not [follow the advice]?
    - None
    - Little
    - Some
    - A Lot
    - Not Sure
13. How much would you say users of other computers would be out at risk if you did not [follow the advice]?
    - None
    - Little
    - Some
    - A Lot
    - Not Sure

A.2.2 "No" Group Template

1. Please explain in a few sentences why you choose not to [follow the advice].
2. How much would you say you are benefited by you [following the advice]?
    - None
    - Little
    - Some
    - A Lot
    - Not Sure
3. How much would you users of other computers are benefited by you not [following the advice]?
    - None
    - Little
    - Some
    - A Lot
    - Not Sure
4. How much would you say you are cost or inconvenienced by you not [following the advice]?
    - None
    - Little
    - Some
    - A Lot
    - Not Sure
5. How much would you users of other computers are cost or inconvenienced by you not [following the advice]?
    - None
    - Little
    - Some
    - A Lot
    - Not Sure
6. How much would you say you are put at risk by you not [following the advice]?
    - None
    - Little
    - Some
    - A Lot
    - Not Sure
7. How much would you users of other computers are put at risk by you not [following the advice]?
    - None
    - Little
    - Some
    - A Lot
    - Not Sure
8. How much would you say you would be benefited if you did [follow the advice]?
    - None
    - Little
    - Some
    - A Lot
    - Not Sure
9. How much would you say users of other computers would be benefited if you did [follow the advice]?
    - None
    - Little
    - Some
    - A Lot
    - Not Sure
10. How much would you say you would be cost or inconvenienced if you did [follow the advice]?
    - None
    - Little
    - Some
    - A Lot
    - Not Sure
11. How much would you say users of other computers would be cost or inconvenienced if you did [follow the advice]?
    - None
    - Little
    - Some
    - A Lot
    - Not Sure
12. How much would you say you would be put at risk if you did [follow the advice]?
    - None
    - Little
    - Some
    - A Lot
    - Not Sure
13. How much would you say users of other computers would be out at risk if you did [follow the advice]?
    - None
    - Little
    - Some
    - A Lot
    - Not Sure

B. STATISTICAL RESULTS

This section contains statistics generated and tests performed for this study, including those not included in the paper's main text. Tables 7- 9 on the following pages contain the details for Mann-Whiteny U-Tests and sign tests used in this paper's analysis.

			. . . of Following					. . . of Not Following				
			Yes	*No*	U-Test			*Yes*	*No*	U-Test		
			Avg.(Med.)	Avg.(Med.)	*U*	Sig.	*d*	Avg.(Med.)	Avg.(Med.)	*U*	Sig.	*d*
Benefit	*Individ.*	Upd.	3.77(4)	2.97(3)	274.5	<0.001	0.51	1.51(1)	2.13(2)	347.5	0.002	0.38
		P.M.	3.78(4)	2.50(2.5)	154.5	<0.001	0.73	1.68(1)	2.70(3)	302	<0.001	0.49
		2FA	3.71(4)	2.90(3)	243.5	<0.001	0.49	1.59(1.5)	2.62(3)	161.5	<0.001	0.61
		Chg.P.	3.47(4)	2.53(3)	256	<0.001	0.57	1.70(2)	3.03(3)	176	<0.001	0.66
	Social	Upd.	2.71(3)	2.39(3)	338	0.286	0.14	1.40(1)	1.58(1)	371	0.371	0.12
		P.M.	2.08(2)	1.70(1)	498.5	0.155	0.17	1.39(1)	1.68(1)	511	0.142	0.18
		2FA	2.48(2)	2.29(2)	390	0.489	0.09	1.59(1)	1.92(1.5)	313.5	0.237	0.16
		Chg.P.	1.73(1)	1.48(1)	463.5	0.235	0.15	1.74(1)	1.58(1)	511	0.467	0.09
Risk	*Individ.*	Upd.	1.56(2)	1.72(2)	496.5	0.335	0.12	3.42(4)	2.77(3)	336.5	0.002	0.37
		P.M.	1.83(2)	2.53(2)	342.5	<0.001	0.49	2.88(3)	1.80(2)	302.5	<0.001	0.52
		2FA	1.56(1)	1.62(1)	498.5	0.729	0.04	3.42(3)	2.61(3)	243.5	<0.001	0.53
		Chg.P.	1.35(1)	1.71(2)	498.5	0.014	0.28	3.14(3)	2.63(3)	440.5	0.003	0.34
	Social	Upd.	1.13(1)	1.38(1)	369.5	0.047	0.25	2.67(3)	1.76(1)	262.5	<0.001	0.44
		P.M.	1.41(1)	1.53(1)	628	0.707	0.04	1.92(2)	1.29(1)	409	0.002	0.37
		2FA	1.31(1)	1.48(1)	433.5	0.47	0.09	2.48(3)	1.79(2)	289	0.013	0.32
		Chg.P.	1.19(1)	1.17(1)	628.5	0.709	0.04	1.70(1)	1.29(1)	483	0.044	0.24
Cost	*Individ.*	Upd.	2.03(2)	2.1(2)	527.5	0.444	0.09	2.95(3)	2.00(2)	247.5	<0.001	0.48
		P.M.	1.73(2)	2.18(2)	533	0.011	0.28	3.15(3)	1.75(1)	244.5	<0.001	0.60
		2FA	2.00(2)	2.39(2)	405.5	0.036	0.26	1.76(1)	1.57(1)	446.5	0.451	0.09
		Chg.P.	2.35(2)	2.97(3)	449.5	0.005	0.33	2.28(3)	1.61(1)	425.5	0.003	0.35
	Social	Upd.	1.22(1)	1.29(1)	431	0.781	0.04	2.32(2)	1.59(1)	248	0.001	0.41
		P.M.	1.28(1)	1.52(1)	565.5	0.213	0.15	1.84(1)	1.03(1)	354	<0.001	0.49
		2FA	1.52(1)	1.44(1)	403.5	0.786	0.04	1.69(1)	1.41(1)	343	0.356	0.12
		Chg.P.	1.28(1)	1.65(1)	491	0.073	0.21	1.50(1)	1.24(1)	525.5	0.174	0.16

Table 7: Rating summaries for all variables with U-Tests comparing the distribution between each Yes (those who follow the advice) and No (those who do not follow) groups. Effect size is measured with Cohen's *d*.

	. . . of Following						. . . of Not Following					
	Ind.>	Soc.>	Tie	*Z*	Sig.	*d*	Ind.>	Soc.>	Tie	*Z*	Sig.	*d*
Benefit	176	10	62	-12.10	<0.001	0.77	108	38	99	-5.71	<0.001	0.36
Risk	112	8	148	-9.40	<0.001	0.57	174	6	85	-12.45	<0.001	0.76
Cost	165	21	75	-10.49	<0.001	0.65	102	11	140	-8.47	<0.001	0.53

Table 8: Sign test results comparing *Individual* and *Social* ratings for each variable from all participants aggregated across both groups and all advice tested. Along with the *Z* and *p* values, we also show difference frequences to show how often participants' *Individual* ratings were higher, lower, or tied with their *Social* rating. Effect size is measured with Cohen's *d*.

			...of Following					...of Not Following				
			Ind.>	Soc.>	Tie	*Z*	Sig.	Ind.>	Soc.>	Tie	*Z*	Sig.
Yes Groups	Update	**Benefit**	25	0	10	-	<0.001	8	2	25	-	0.109
		Risk	17	1	21	-	<0.001	21	0	14	-	<0.001
		Cost	27	0	10	-5.004	<0.001	17	2	14	-	0.001
	P.M.	**Benefit**	30	1	6	-5.029	<0.001	11	2	23	-	0.022
		Risk	19	2	20	-	<0.001	24	2	11	-4.118	<0.001
		Cost	17	2	20	-	0.001	26	2	9	-4.341	<0.001
	2FA	**Benefit**	21	1	8	-	<0.001	7	6	15	-	1
		Risk	9	1	25	-	0.021	19	0	12	-	<0.001
		Cost	14	4	13	-	0.031	3	3	22	-	1
	Chg.P.	**Benefit**	29	1	2	-4.930	<0.001	13	12	9	-	1
		Risk	10	2	25	-	0.039	27	1	5	-4.725	<0.001
		Cost	25	1	10	-4.511	<0.001	16	0	16	-	<0.001
No Groups	Update	**Benefit**	12	1	10	-	0.003	12	4	8	-	0.077
		Risk	10	1	13	-	0.012	22	1	6	-	<0.001
		Cost	17	4	3	-	0.007	12	1	13	-	0.003
	P.M.	**Benefit**	19	4	9	-	0.003	20	3	8	-	<0.001
		Risk	21	0	10	-	<0.001	13	1	19	-	0.002
		Cost	21	5	7	-2.942	0.003	13	0	20	-	<0.001
	2FA	**Benefit**	15	2	9	-	0.002	14	6	5	-	0.115
		Risk	8	0	18	-	0.008	16	1	12	-	<0.001
		Cost	18	1	8	-	<0.001	4	3	19	-	1
	Chg.P.	**Benefit**	25	0	8	-	<0.001	23	3	6	-3.726	<0.001
		Risk	18	1	16	-	<0.001	32	0	6	-5.480	<0.001
		Cost	26	4	4	-3.834	<0.001	11	0	27	-	0.001

Table 9: Sign test results comparing *Individual* and *Social* ratings for each variable from tests performed on response sets separated by elements of advice and participants' group in the study (i.e., Yes or No group). For tests where there are fewer than 26 non-ties, the exact p is listed. In other cases, an asymptotic signifigance value is listed. Since Z statistics were not calculated for exact signifigance tests, this table only lists such a value where applicable.

Expecting the Unexpected: Understanding Mismatched Privacy Expectations Online

Ashwini Rao
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA
arao@cmu.edu

Florian Schaub
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA
fschaub@cs.cmu.edu

Norman Sadeh
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA
sadeh@cs.cmu.edu

Alessandro Acquisti
Heinz College
Carnegie Mellon University
Pittsburgh, PA, USA
acquisti@andrew.cmu.edu

Ruogu Kang
Facebook
Menlo Park, CA, USA
ruoguk@fb.com

ABSTRACT

Online privacy policies are the primary mechanism for informing users about data practices of online services. In practice, users ignore privacy policies as policies are long and complex to read. Since users do not read privacy policies, their expectations regarding data practices of online services may not match a service's actual data practices. Mismatches may result in users exposing themselves to unanticipated privacy risks such as unknowingly sharing personal information with online services. One approach for mitigating privacy risks is to provide simplified privacy notices, in addition to privacy policies, that highlight unexpected data practices. However, identifying mismatches between user expectations and services' practices is challenging. We propose and validate a practical approach for studying Web users' privacy expectations and identifying mismatches with practices stated in privacy policies. We conducted a user study with 240 participants and 16 websites, and identified mismatches in collection, sharing and deletion data practices. We discuss the implications of our results for the design of usable privacy notices, service providers, as well as public policy.

1. INTRODUCTION

Privacy policies serve as the primary mechanism for notifying users about a website's data practices, such as collection and sharing of personal information. However, website privacy policies, written in natural language, can be long, time consuming to read [18, 30], and difficult to understand for users [42, 46]. They are therefore often ignored by users [9, 43]. One approach for helping users is to provide additional privacy notices that are based on privacy policies, but are shorter, easier to understand and more usable [10,22,49,55]. Prior work on privacy notices has focused on summary notices that display data practices in an easy to understand visual format [10, 22, 49, 55]. Even with simplified privacy notices, much of the information may not be relevant to users. Many data practices are expected and obvious, may not create concern, or do not apply to the user's current interaction with a service. For instance, it is likely obvious to users that when they explicitly provide their contact and payment details to an online store that that information will be collected and used to fulfill the purchase. However, data practices that are unexpected may result in a loss of trust and a sense that one's privacy has been violated, even if the practices in question were disclosed in the service's privacy policy [47]. More importantly, expectations influence decision making [17] and mismatches between users' expectations and website data practices may lead to incorrect privacy-related decisions.

The framework of contextual integrity highlights the impact of social context and information type on flow of information [34,35]. Expectations regarding flow of information may vary by social context and information type. For instance, collection of financial information on a banking website may be more expected than collection of health information. Privacy expectations are further influenced by an individual's personal, social and cultural background, as well as expectations in social roles and other "borders" that delineate spheres of privacy [29, 39]. For instance, depending on their technical knowledge, some users may expect that websites they visit can infer their rough location based on their IP address. For others, inference of their location may be completely unexpected.

Although unexpected data practices may be described in a privacy policy, they are likely to be overlooked among descriptions of practices that are expected or irrelevant to the user's current transactional context. The verbosity of privacy policies may be necessary to comply with legal and regulatory requirements, but it also means that privacy policies are not helpful to users in making informed privacy decisions [9]. In order to provide transparency to users, compliance-oriented privacy policies should be complemented with short form notices tailored to the user's transactional context [49] that should warn users about unexpected prac-

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22–24, 2016, Denver, Colorado.

tices in particular [14]. The challenge, however, lies in identifying unexpected practices. Users' privacy preferences have been studied in different contexts [23, 38, 51]. However, privacy behavior differs from stated preferences [36], and preferences are not reliable for identifying mismatches between privacy expectations and a company's actual practices.

1.1 Contributions

To advance toward more practical solutions that can impact privacy notice design, we propose a practical approach for determining mismatches between users' expectations and services' data practices, as stated in their privacy policies. Research in other fields, such as marketing, has highlighted that the term "expectations" can mean at least four different things in consumers' context [32], but in the privacy context most work has focused on expectations in the desired sense or preferences [23,33], or has not clarified the meaning of expectation [13, 16, 26]. We propose to elicit privacy expectations, in the sense of "expected occurrence likelihood," rather than aspirational privacy preferences, and use the elicited expectations to identify mismatches with stated data practices. By focusing on expectations of what is happening, we avoid problems with unreliable subjective preferences of what should happen.

We compared expectations elicited from users with website data practices extracted from website privacy policies with manual annotations. Our analysis shows that characteristics of a website, such as its type, as well as user characteristics, such as privacy knowledge and concern, are strong predictors of data practices that are likely to be unexpected.

From our results, we derive guidelines on what data practices are likely unexpected and should therefore be emphasized in privacy notices. Knowledge about which characteristics affect expectations can be used to contextualize notices to the type of website and transactional context, as well as personalize notices to specific audiences in order to make unexpected practices more salient compared to expected practices, and thus make it easier for users to obtain information relevant for making informed privacy decisions. Our insights can benefit third-parties that generate simplified privacy notices, for example via browser extensions, as well as service providers. Both can use our approach to identify data practices that users will likely not expect and may cause privacy concern. Service providers could assuage user concerns by explaining the rationale behind such data practices.

While we manually extracted data practices from privacy policies to ensure reliable ground truth data, recent advances in the semi-automated analysis of privacy policies [6, 25, 48, 53, 55] show promise that our approach can be automated and scaled up to a large number of websites once such techniques are sufficiently robust.

2. BACKGROUND & RELATED WORK

In the United States, website privacy policies serve as the dominant mechanism for informing Internet users about website data practices such as collection, sharing and retention of personal information [47]. A website privacy policy, written in natural language, contains statements about the website's data practices. Policies may be long and time consuming to read [30] and require high language proficiency skills [18], which can lead to differences in how the general public and legal scholars interpret policy statements [46].

One approach to help users understand website data practices is to provide more concise privacy notices in addition to privacy policies [49]. Such privacy notices may be based on privacy policies, but are generally shorter and more usable. They could be provided either by website operators or by third parties. Research on privacy notices has focused on display formats that are easier for users to understand [10, 22, 31, 48, 55].

Users' privacy preferences and willingness to share information have been studied in many contexts [2, 23, 38, 51]. Acquisti et al. [2] note that privacy preferences and privacy decision making are prone to uncertainty, context-dependent, shaped by heuristics and cognitive biases, malleable and easily influenced by framing. Elicited privacy preferences can therefore be difficult to generalize, and actual behavior often deviates from stated preferences [36]. Observing privacy behavior is preferable, but behavioral studies can be challenging and resource-intensive to conduct at scale.

Privacy research has also explored the concept of expectations of privacy, including seminal work by Altman [3, 28], Marx [29] and Nissenbaum [34, 35]. For instance, Altman showed that individuals continuously modify their behavior to achieve an expected level of privacy [3]. Nissenbaum discusses how expectations of privacy are shaped by context [34]. However, to the best of our knowledge, privacy research has not focused on the potential for multiple levels or types of expectations. For example, in Altman and Nissenbaum's work, there is a single notion of expectation that may change based on different factors such as context. Privacy research typically differentiates between expected privacy and actual privacy, for example, Altman differentiates between desired and achieved levels of privacy [3].

However, research in other domains indicates that individuals have multiple levels or types of expectations [15,32,50,54] and these types of expectations can impact constructs such as consumer satisfaction [50] and performance [15]. For instance, Miller distinguishes four expectation types: *Ideal, Expected, Minimum Tolerable,* and *Deserved* [32]. The Ideal represents what users think performance "can be." The Expected is objective, without an affective dimension, and represents what users think performance "will be." The Deserved has an affective dimension and represents what users feel performance "should be." Lastly, the Minimum Tolerable is what users think the lowest performance "must be."

Based on Miller's work [32], we argue that people likely also have multiple levels of privacy expectations beyond desired and achieved privacy. Therefore, we conceptually distinguish between Expected ("will be") and Deserved ("should be") types of expectation in measuring user expectations for website data practices, and focus on eliciting the Expected ("will be") type to identify mismatches.

We identify mismatches in user expectations regarding website data practices. We study if users expect that a website *will* collect, share or delete data. Prior work has studied mismatches in other types of expectations [13, 16, 26, 33]. To measure expectation, these studies either used an expectation type in the sense of desired preferences (*should*) [33], or they did not clarify the type of expectation [13, 16, 26]. Earp et al. studied Internet users' privacy values and analyzed privacy policies for respective statements [13]. They find

that Internet users' concerns and values are not adequately reflected in privacy policies. Gomez et al. also compared websites' data practices with practices users find concerning [16]. Milne and Bahl examined differences between consumers' and marketers' expectations regarding use of eight information technologies [33]. Liu et al. measured disparity between expected and actual Facebook privacy settings. In contrast to our study on website data practices, Lin et al. studied expectations regarding data practices of mobile apps [24]. Further, their work did not differentiate between different types of expectations, and, while eliciting expectations, did not clarify the type of expectation being elicited.

In contrast to prior work, we propose an approach that facilitates direct comparison of individuals' expectations of what a website's data practices are to the website's actual claims of what they do as stated in their privacy policy.

3. METHODOLOGY

Our goal is to identify mismatches between user privacy expectations regarding website data practices and the practices websites disclose in their privacy policy. We define privacy expectation as what users think a website "will" do or is doing as opposed to what they prefer a website "should" do, which corresponds to Miller's distinction of the Expected and Deserved expectation types [32]. We elicited user expectations for different online scenarios that varied in terms of data practices, website type, and other website characteristics, in order to understand the impact of contextual factors on privacy expectations. We also studied how user characteristics influence expectations. To identify mismatched expectations and unexpected practices, we compared elicited expectations with the data practices described in websites' privacy policies. In the rest of this section, we describe the study design, studied parameters, and the procedure we used to identify and classify mismatched expectations.

3.1 Study Design

To assess the impact of different website scenarios on privacy expectations, we conducted an online study involving 16 websites and 240 participants. We opted for a between-subjects design to prevent fatigue and learning effects, in which we asked participants to answer questions about one website randomly assigned to them. Website type (health, finance, dictionary) and popularity (low, high) were the main independent variables in our study, resulting in a 3x2 design with six conditions. We based website type and popularity on website categories and traffic rankings respectively obtained from Alexa.com [4]. In total, we studied 16 websites, which are listed in Table 1, across three website types (7 Health, 7 Finance, 2 Dictionary). Fifteen participants were assigned to each website, resulting in the following number of participants per condition: 60 in Health-Low, 45 in Health-High, 60 in Finance-Low, 45 in Finance-High, 15 in Dictionary-Low, and 15 in Dictionary-High.

Website	Type	Subtype	Context	Rank
Webmd.com	Health	Reference	Private	107
Medhelp.org	Health	Reference	Private	2,135
Medlineplus.gov	Health	Reference	Government	558,671
Walgreens.com	Health	Pharmacy	Private	315
Bartelldrugs.com	Health	Pharmacy	Private	54,737
Mayoclinic.org	Health	Clinic	Private	297
Clevelandclinic.org	Health	Clinic	Private	2,629
Americanexpress.com	Finance	Credit	Private	76
Discover.com	Finance	Credit	Private	324
Bankofamerica.com	Finance	Bank	Private	33
Woodlandbank.com	Finance	Bank	Private	915,921
Banknd.nd.gov	Finance	Bank	Government	5,267
Paypal.com	Finance	Payment	Private	21
V.me	Finance	Payment	Private	27,289
Merriam-webster.com	Dictionary	–	Private	266
Wordnik.com	Dictionary	–	Private	8,412

Table 1: Websites used in the study (Alexa website rank as of March 10, 2015).

3.1.1 Survey Questionnaire

We designed a questionnaire to measure user expectations for eight collection data practices (4 information types collected with or without account), eight sharing data practices (4 information types shared for core or other purposes), and one deletion data practice. These website practices, listed in Table 2, were treated as 17 dependent variables.

The survey questionnaire consisted of three sections: introduction, main questionnaire and post-questionnaire. Privacy-related questions, which could bias participant responses, were asked in the post-questionnaire. While designing the questionnaire, we used think-aloud and verbal-probing cognitive interviewing techniques [52] in pilot tests with six participants. We tested whether participants understood the questions. We iteratively refined the questionnaire based on participant feedback. We summarize the questionnaire below. The full questionnaire is provided in Appendix B.

At the beginning of the questionnaire, we explained the purpose of the study. We framed the purpose of the study as understanding user opinions about websites rather than their knowledge of data practices, to avoid self-presentation issues associated with knowledge questions [7]. We also did not mention privacy or data practices to avoid biasing participants. After explaining the purpose, we asked whether participants had visited or used the assigned website before.

We instructed the participants to familiarize themselves with the website assigned to them. Since participants may explore websites in different ways, we wanted them to look at what they considered important and did not want to bias their thinking by providing too specific instructions. Based on participant feedback from our in-lab pilot tests, we asked participants to look at the website for 2–3 minutes. Initially, we had instructed the participants to take their time familiarizing themselves with the website. However, after about three minutes of interaction, our in-lab participants were either ready to provide their opinions or were not sure what else to look at. Two participants specifically told us that it would be helpful if we told them how much time they should spend looking at a website. Because the website was opened in a separate browser window, participants could go back to the website at any point during the study.

After participants interacted with the website, we provided definitions of contact, financial, health and current location information. For example, we described contact information as "Examples include (but are not limited to) email address, postal address, phone number, home phone number, etc." Definitions for all information types are provided in Appendix A.

In the main part of the questionnaire, we asked participants about their expectations regarding different website data practices, listed in Table 2. First, we asked them questions about data collection practices in two scenarios: collection without account and collection with account. Before asking questions related to a scenario, we showed scenario descriptions. For instance, for the collection without account scenario, we showed the description "*Imagine that you are browsing [website name] website. You do not have a user account on [website name], that is, you have not registered or created an account on [website name].*" We then asked them about their expectations concerning whether and how the website collects different types of data. These questions were framed as likelihood questions: "*What is the likelihood that [website name] would collect your information in this scenario?*" Note that we framed the questions as "would collect" in order to capture participants' objective expectations, and not what they would prefer. We provided a 4-point scale {Likely, Somewhat likely, Somewhat unlikely, Unlikely} as the response option. We wanted respondents' "best guess" and thus did not provide a neutral or not sure option. We did so because users often do not read privacy policies and decide about data practices of a website based on incomplete information, that is, their best guess. We asked an open-ended question to understand how they thought the website collected their information without having an account on the website. After answering questions about the without account scenario, participants read the scenario description for collection with an account and answered the same questions regarding this scenario.

After collection-related questions, we asked participants questions regarding data sharing practices. We first asked them questions about a scenario where data is shared for core purposes, which we defined as sharing only for the purpose of providing a service that the user requested. We them asked them questions regarding a scenario where data is shared for other purposes, which we defined as a purpose unrelated to providing a service that the user requested. To answer the questions, participants had to understand three concepts. First, what are core purposes for the given website? Second, what are other purposes for the given website? Lastly, with whom could the website possibly share information? To encourage them to think about these concepts, we asked them three open-ended questions before asking questions related to sharing. Concerning the data deletion practice, we asked participants whether they expected that the website would allow them to delete all, some or none of their data.

In the post-questionnaire, we captured different user characteristics in order to study their impact on the participants' privacy expectations. We list these characteristics in Table 3. We ordered the questions based on ease of answering, level of threat, and effect on subsequent answers [7]. First, we asked questions about their *past experiences* with the assigned website including if they had an account on the website, how much they had used the website, familiarity with the website and the website's perceived trustworthiness. Users' past experience may influence their expectations, for example, having an account may expose them to additional parts of a website that may improve their awareness of the website's data practices. Participants then provided demographic information (gender, age, education, occupation) and whether they had a background in computer-related fields, which may indicate an enhanced understanding of online data practices. We also asked for their U.S. state of residence, to assess whether privacy regulation on the state level, e.g., in California, impacts privacy expectations. We further included questions about privacy-protective behavior [37] and their familiarity and knowledge of privacy concepts and privacy-enhancing technologies [21]. We also asked whether participants had negative online experiences [44], as they may expect data practices to be more privacy invasive. Lastly, we included the 10-item IUIPC scale [27] to assess online privacy concerns.

3.1.2 Study Deployment & Demographics

Our study received approval from Carnegie Mellon University's Institutional Review Board. To recruit participants efficiently and rapidly, we used the Amazon Mechanical Turk crowdsourcing platform [5]. Research has shown that the Mechanical Turk sample pool is more diverse than traditional sample pools [40], and that data quality is typically good [8, 40, 41]. In February 2015, we recruited 240 participants. We restricted participation to individuals located in the United States, with at least a 95% approval rate and at least 500 completed tasks on Amazon Mechanical Turk. Participants received $3.50 for completing the study. Each participant was randomly assigned to one of the 16 websites. We implemented our survey on SurveyGizmo. Participants were redirected from Amazon Mechanical Turk to SurveyGizmo to complete the survey. We used a combination of SurveyGizmo and Mechanical Turk features to ensure that participants took the survey only once. We implemented timers to measure how long participants interacted with a website and to measure time spent on survey questions. As instructed, participants, spent on average 1.99 min (SD=2.41, median=1.56) interacting with a website. Statistical analysis did not show a significant impact of the amount of time spent on a website or on the survey questions.

To ensure data quality, we screened for participants that completed the study in less than 10 minutes (pilot tests suggested a 30-minute completion time), and checked whether participants answered two questions about prior experience with the assigned website at the beginning and the end of the survey consistently. All participants passed at least two of three quality criteria.

The 240 participants completed our online survey in 22.5 minutes on average (SD=12.8, median=18.6). The sample was 42% female and 58% male. The average age was 34.4 years (SD=10.3, median=32). The majority (85.3%) had at least some college education and 61.6% reported an Associates, Bachelors or Graduate degree. A fifth of the participants (19.5%) had a college degree or work experience in a computer-related field. The top primary occupations were administrative staff (17.5%), service (14.1%), and business/management/financial (12%).

3.2 Scenario Parameters

We defined multiple scenarios that varied in key parameters, namely data practices and website characteristics. We hypothesized that these parameters may influence privacy expectations and mismatches.

Action	Scenario	Information type
Collection	With account	Contact
		Financial
		Health
		Current location
	Without account	Contact
		Financial
		Health
		Current location
Sharing	For core purpose	Contact
		Financial
		Health
		Current location
	For other purpose	Contact
		Financial
		Health
		Current location
Deletion	–	Personal data

Table 2: Studied data practices.

Website characteristic	
Type	Finance
	Health
	Dictionary
Popularity	More
	Less
Context	Private
	Government
User characteristic	
Demographic: age, gender, education, occupation computer background, state of residence	
Privacy protective behavior	
Familiarity with privacy concepts and tools	
Knowledge of privacy concepts and tools	
Negative online experience	
Online privacy concern	
Experience with website: amount of recent use, has account, familiarity, trust	

Table 3: Studied website and user characteristics.

3.2.1 Data Practices of Interest

We decided to focus on data practices concerning *collection, sharing and deletion of personal information* as prior research has shown that users are especially concerned about surreptitious collection, unauthorized disclosure and wrongful retention of personal information [47]. We considered the collection and sharing of four categories of privacy-sensitive information [1, 19, 23]: *contact information* (e.g., email or postal address), *financial information* (e.g., bank account information, credit card details, or credit history), *health information* (e.g., medical history or health insurance information), and *current location* (e.g., from where a user is accessing the website). The definitions are provided in Appendix A.

We further distinguished between scenarios in which users have or do not have an *account with the website.* Websites typically collect data when users create an account, often explicitly provided by the user. Hence, users may have different expectations depending on whether they have an account or not. In general, users may not be aware of implicit or automated data collection, e.g., of IP addresses and cookies. Websites may use IPs, email addresses and other information to acquire additional data about individuals, such as purchase history or interests, from social media services and data brokers [45].

Similarly, information sharing with third parties, while abundant, is less visible to users. Websites assume to have the users' permission because they are using the website and therefore implicitly consent to its privacy policy. We distinguish between third party sharing for *core purposes*, such as sharing a user's information to provide the requested service (e.g., payment processing or providing contact information to a delivery service), and sharing for unrelated *other purposes*, such as advertising or marketing. In all, we studied 17 data practices summarized in Table 2.

3.2.2 Website Characteristics

To understand whether mismatched privacy expectations vary based on context, we considered three website characteristics: website type, popularity and ownership. *Website type* may influence what information users expect a website to collect [34]. We selected three website categories: finance, health and dictionary. Users may expect finance and health websites to collect sensitive information (health or financial data, respectively). In contrast, users may not expect dictionary websites to collect sensitive information. In the financial category, we included banking, credit card and online payment websites. In the health category, we included pharmacy, health clinic and health reference websites. Website categories were determined using Alexa website categories [4].

Users' expectations may be influenced by their offline interactions with entities affiliated with a website, such as visiting a bank branch or a clinic. Hence, we included websites with *offline interactions* as well as online-only websites in the health and financial categories; dictionary websites were online-only.

Interestingly, popular financial websites have been shown to have more privacy-invasive data practices than less popular ones [12]. Therefore, we studied websites of comparable utility but varying in *popularity*, as determined by their traffic rankings [4].

For a given website type, *government or private ownership* may influence user expectations. Our sample population was limited to the United States, and in the post-Snowden era, people may expect government websites to be more privacy invasive than private websites. Hence, we studied whether user expectations varied between government and privately-owned health and financial websites. Table 3 summarizes the website characteristics that we considered in our model.

3.3 Identifying Mismatched Expectations

To identify mismatched expectations and, thus, unexpected data practices, we compare participants' expectations concerning a specific data practice with the results of our privacy policy analysis with regard to that practice. The infor-

mation about a given website data practice extracted from the website's privacy policy, may be Yes, No, Unclear or Not addressed. We elicited an objective "will" expectation from study participants. They rated their expectation of whether a website *will* engage in a specific data practice on a 4-point scale (Unlikely–1, Somewhat unlikely–2, Somewhat likely–3, Likely–4). These ratings can be interpreted as indications of a positive (Yes) or a negative (No) expectation that can be compared to the policy analysis results. Comparing a website's data practices and users' expectations this way, results in eight potential combinations, as shown in Table 4. For Yes–Yes and No–No, users' expectations match the websites' practices. Yes–No and No–Yes combinations constitute explicit mismatches. For Unclear–Yes, Unclear–No, Not addressed–Yes and Not addressed–No, it is not clear whether expectations are mismatched because the website's policy is unclear or silent on the particular data practice.

It is worth taking a closer look at the implications of the different types of mismatches. Although, both Yes–No and No–Yes are mismatches, they may impact users' perception of privacy violations differently. In the case of Yes–No, the website will collect or share information, but users optimistically expect it not to. Due to lack of awareness that the website shares information, users may decide to use the website. By doing so, they give up data that they do not want to share, resulting in a violation of their privacy. Although the website discloses its data practice in its policy, from a user viewpoint, the practice could be considered surreptitious unless users are appropriately and explicitly made aware of it. When found out, such data practices may damage a company's reputation.

In contrast, in the case of No–Yes, a website will not engage in a collection or sharing practice, but users pessimistically expect it to. As a result, users may have reservations to use the website or some features, which may affect their utility but not their privacy. In such cases, websites should aim to make users aware of the privacy-protective practices to assuage pessimistic expectations.

The number of unclear website data practices can be high, for example, ∼40% of collection data practices in this study are unclear. Hence, it is important to analyze the impact of unclear data practices. Consider the Unclear–Yes case. If the website is really collecting information, then it would be a Yes–Yes match. If the website is not collecting information, then it would be a No–Yes mismatch. The same applies to Unclear–No. As discussed, a Yes–No mismatch, could potentially violate user privacy. Hence, for analysis purposes, we could treat Unclear as a likely Yes. We use a similar approach for Not addressed–Yes and Not addressed–No.

We can similarly analyze mismatches in case of the data deletion practice by considering two types of Yes values, Yes–Full and Yes–Partial, separately. We could also simplify the analysis by combining the two Yes values. In case of deletion, users may use a website if they think that the website allows deletion, whereas for collection and sharing they may not use the website. Hence, in case of deletion, the implications of No–Yes and Yes–No mismatches are reversed.

4. STUDY RESULTS

To identify unexpected practices – those that did not match participants' privacy expectations – we first analyzed the privacy policies of the websites used in our study and then compared them to participants' expectations.

		User:	Yes	No
Website:	Yes		✓	X
	No		X	✓
	Unclear		?	?
	Not addressed		?	?

Table 4: Overview of matched and mismatched expectations. Match (✓) or mismatch (X) between a website's data practice and a user's expectation. If the website's policy is unclear or silent on a practice, it cannot be determined if it matches user expectations (?).

4.1 Website Privacy Policy Analysis

Two annotators, one with legal and another with privacy expertise, independently read each of the 16 privacy policies (cf. Table 1) and extracted the relevant collection, sharing and deletion data practices described earlier. Agreement was generally high, for instance, among the 17 data practices, the highest inter-annotator agreement was κ=1 and lowest agreement was κ=0.718. All disagreements were resolved jointly after initial independent coding. Following an annotation approach similar to Reidenberg et al. [46], annotators coded collection and sharing practices as *yes, no, unclear* or *not addressed*, in order to take ambiguity in the policy language (*unclear*) or silence on a specific practice (*not addressed*) into account. For example, the statement "When you use our Websites, we collect your location using IP address." makes it clear that the website collects location information. However, the statement "We collect the IP address from which you access our Website." mentions collecting IP address but is unclear whether the website collects location information. Collection and sharing practices were analyzed with regard to contact, financial, health and current location information, as well as for two collection contexts (with/without user account) and for two sharing purposes (core/other). Deletion practices were annotated as *full deletion* (websites allows deletion of all user data), *partial deletion* (deletion of only some data), *no deletion*, *unclear*, or *not addressed*. Table 5 shows a sample annotation for Bank of America's privacy policy. Annotating privacy policies is an active area of research, and recent results [6, 53] show the possibility of achieving acceptable level of agreement with semi-automated techniques and non-expert crowdworkers. Such techniques can enable scaling up our approach to large number of websites.

Figure 1 gives an overview of data practices extracted from the privacy policies of the 16 websites (7 financial, 7 health, 2 dictionary) used in our study. It shows the percentage of collection and sharing data practices that are clear, unclear or not addressed in the privacy policies. We find that policies in all three website categories are mostly clear about practices concerning the collection or sharing of contact information, i.e., they make explicit statements about whether they collect or not collect contact information and make clear statements about sharing (dominantly yes for core purposes; no for other purposes).

Data practice	Answer
Collect contact – with account	Yes
Collect contact – without account	Unclear
Collect financial – with account	Yes
Collect financial – without account	No
Collect health – with account	Yes
Collect health – without account	No
Collect location – with account	Unclear
Collect location – without account	Unclear
Share contact – core purpose	Unclear
Share contact – other purpose	Unclear
Share financial – core purpose	Yes
Share financial – other purpose	Yes
Share health – core purpose	Yes
Share health – other purpose	No
Share location – core purpose	Unclear
Share location – other purpose	Unclear
Deletion	No

Table 5: Annotations for the 17 data practices of BankofAmerica.com's privacy policy.

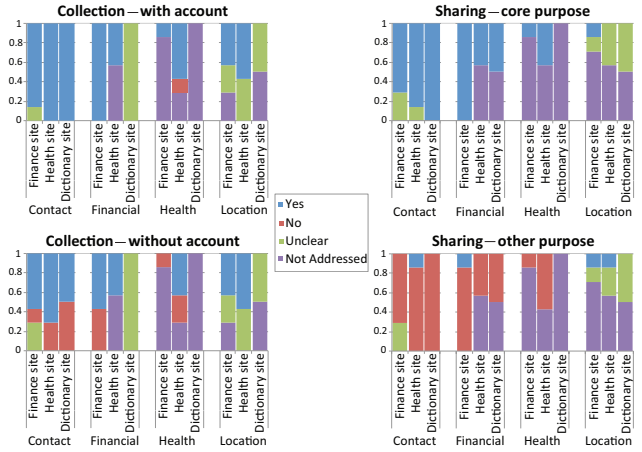


Figure 1: Collection and sharing data practices of the 16 websites used in our study, based on the analysis of the websites' privacy policies.

Not surprisingly, finance websites make explicit statements about collection and sharing of financial information. Note that credit card and online payment finance websites collect financial information even from non-registered users, e.g., when users buy products, but banking websites do not. About half of the health websites' privacy policies also make explicit statements concerning financial information, however, the other half is silent on whether they collect or share financial information. Interestingly, the dictionary websites make statements that leave it unclear if they may collect financial information, but are either explicit or silent on sharing of financial information. Dictionary sites mention processing payments or posting transactions but not explicit collection of financial information.

All dictionary websites and all but one of the financial websites do not address collection or sharing of health information. One of the finance websites, BankofAmerica.com is explicit about collecting health information from registered users and sharing it with third parties for core purposes. It does so via its insurance-related affiliates, which may not be obvious to users. However, all but two of the health websites are explicit about whether they collect health information. Both health clinic websites do not address collection of health information in their website privacy policy, but contain links to additional policies, which may disclose their collection practices. Health websites are less explicit about sharing of health information compared to collection of health information.

About half of the financial and health websites are clear about collection of current location information, but none of the dictionary sites are clear on this aspect. Almost all website privacy policies are unclear or silent on whether they share location information with third parties. Only one finance website explicitly states that it shares user location for core and other purposes. Only one health website explicitly states that it shares user location for other purposes, but it is unclear whether it shares it for core purposes.

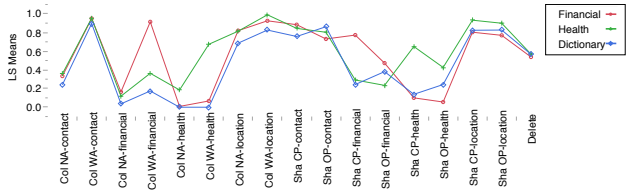
Financial websites are more explicit about deletion data practices compared to health and dictionary websites. Nearly 71% (5) of the financial websites clearly disclose their practice in contrast to 50% (1) of the dictionary websites and 28% (2) of the health websites. However, nearly half of the financial websites (3) do not allow any deletion of data and two only allow partial deletion. In contrast, when clear about the practice, health websites (2) and dictionary websites (1) allow full deletion.

The privacy policy analysis shows that some data practices are common across different website types, whereas others are category-specific or even vary within a category. This suggests that if users would rely on website characteristics to anchor their privacy expectations, these heuristics may lead to mismatches between their expectations and a website's stated data practices.

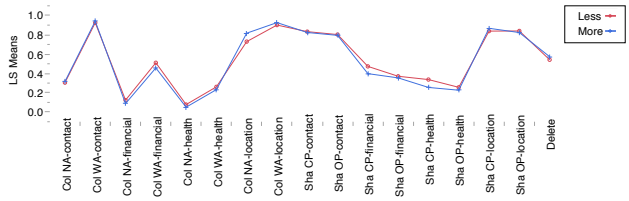
4.2 Impact of Website Characteristics

We find that a website's type has a significant impact on user expectations. This implies that what data practices users expect a website to engage in is influenced by the type of website. We did not find significant differences for popularity or ownership, suggesting they play no or a lesser role in shaping privacy expectations. For example, users expect data practices of BankofAmerica.com, a finance website to be different than those of WebMD.com, a health website. However, they have similar expectations for two finance websites even if one of them is more popular than the other (e.g., in our dataset BankofAmerica.com's popularity rank is 33 and WoodlandBank.com's is 915,921). Similarly, expectations do not differ between privately-owned and government-operated websites. We describe our analysis in more detail in the following.

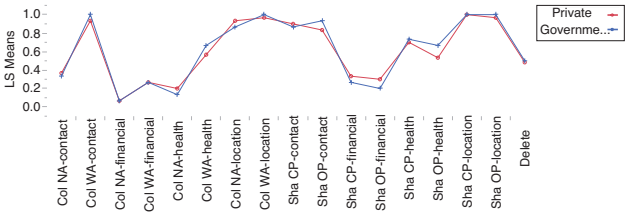
We used a mixed-model ANOVA to analyze the impact of website type and popularity on user expectations. We considered website type (health, finance, dictionary) and popularity (high, low) as nominal between-subjects independent variables. We considered participant expectations concerning the 17 data practices as continuous repeated mea-



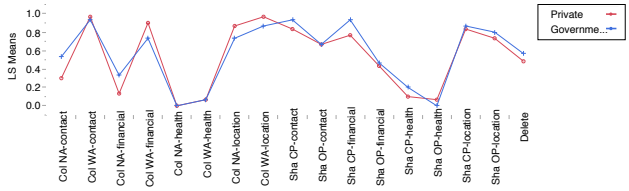
(a) Website type (*sig.*)



(b) Popularity (*n.s.*)



(c) Ownership for health websites (*n.s.*)

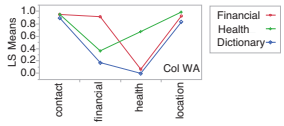


(d) Ownership for finance websites (*n.s.*)

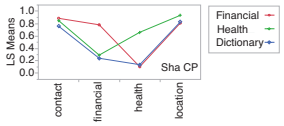
Figure 2: Interaction of website characteristics and user expectations for the 17 data practices. Higher Least Square Means value implies users expect data practice to be more likely (Col: Collection, Sha: Sharing, WA: With Account, NA: No Account, CP: Core Purpose, OP: Other Purpose).

sures dependent variables (DV), which, as a group, measured users' overall expectation. We verified that the group of DVs has an approximate normal distribution with a normal-quantile plot of a linear combination of the individual DV scores. A Shapiro-Wilk W test showed only moderate departure from normality (W=.988, p=.041).

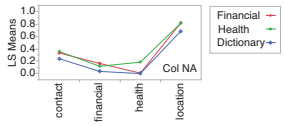
Results showed that interaction of website type and data practices was significant ($F(32.438)$=12.819, p<.0001), see Figure 2a for an interaction plot. This interaction effect suggests that website type impacts what data practices users expect. Compare, for instance, the impact of financial website type on users' expectations concerning collection of financial and health information from registered users (*COL WA-financial*), *COL WA-health*). Higher Least Square Means value implies that users are more likely to expect a data practice. Users expect financial websites to collect financial (high $LSMeans$), but not health data (low $LSMeans$). Figures 2b–2d further show interactions of website popularity



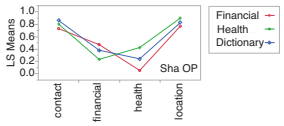
(a) Collection (w. account)



(b) Sharing (core purpose)



(c) Collection (no account)



(d) Sharing (other purpose)

Figure 3: Interaction of website type and expectations for specific data practices. Website type significantly interacts with user expectations for financial and health information. Higher Least Square Means value implies users are more likely to expect a data practice.

and ownership, which were not significant. Note that only the health and finance categories contained government-operated websites, dictionary websites are therefore not shown in Figures 2c and 2d.

We also studied the impact of website type on individual data practices. The distribution of values of individual data practices was non-normal. We treated them as two-level nominal variables and used a χ^2 statistical test. Figure 3a shows what information types participants expect websites to collect from registered users. If $LS\ Means$>0.5, users are likely to expect the data practice. Type of website has a significant impact for expectations of collection of financial ($\chi^2(2,240)$=87.7, p<.0001, R^2=.302) and health information ($\chi^2(2,240)$=105.826, p<.0001, R^2=.3935), but not for collection of contact and current location information. Users expect all types of websites to collect contact and location information when they have an account. However, they expect only financial websites to collect financial data and health websites to collect health data. A financial website collecting health data would lead to a mismatch in expectations. Most financial websites we studied do not collect health data. However, one financial website in our study, BankofAmerica.com, collects health information when users have an account, which violates user expectations.

As shown in Figure 3c, in the without account scenario, participants expect only collection of location information, but for all types of websites. Participants are unlikely to expect websites to collect contact, financial and health data from users without an account. As we will discuss shortly, websites can collect contact and financial data without an account, leading to a mismatch with expectations.

Concerning expectations of data sharing, Figure 3b shows that participants likely expect all types of websites to share contact and current location information for core purposes. Website type has a significant interaction effect for expectations of sharing financial information ($\chi^2(2,240)$=59.175, p<.0001, R^2=.1868) and expectations of sharing health information ($\chi^2(2,240)$ =77.935, p<.0001, R^2=.2642). Participants expect only financial websites to share financial data and health websites to share health data. One financial web-

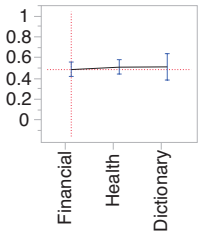


Figure 4: Website type does not impact deletion data practice. LS Means (least square mean) higher value implies users expect data practice to be more likely.

site, BankofAmerica.com, shares health information for core purposes, which violates user expectations.

Figure 3d shows expectations of websites sharing for other purposes. In this case, users expect all types of websites to share contact and location information for other purposes. They do not expect any type of website to share financial or health information for other purposes. Users expecting websites to share contact information for other purposes is interesting because, as we discuss later, most websites do not do so. Lastly, we did not find significant interactions of website type with participants expectations concerning websites' data deletion practices. Participants expected all website types to permit deletion of data, as shown in Figure 4, but this expectation does not match reality.

Further analysis shows that user expectations can vary for individual data types within a larger data type category. For example, for collection of contact information in the with account scenario, participants expected that websites were more likely to collect email address (93.3% participants) than postal address (75%) or phone number (70.8%). Expectations for specific data types can also vary within website sub-categories. For instance, for collection of health information in the with account scenario, participants expected that pharmacy websites were more likely to collect health insurance information than medical history (66.6% vs. 53.3%), but health clinic websites were more likely to collect medical history than health insurance (67.7% vs. 54.8%). Although we could analyze expectations at a finer granularity, identifying mismatches in expectations at finer granularity is problematic because website privacy policies do not typically disclose data practices at such fine granularity. Privacy policies generally discuss data practices at the level of coarse grained categories such as contact information rather than email address or postal address.

4.3 Impact of User Characteristics

We analyzed the effect of multiple user characteristics on participants' data practice expectations. We find that privacy knowledge, privacy concept familiarity, privacy concern, privacy-protective behavior, negative online experience, age, trust in website, website familiarity, whether participant has an account, and recent use have a significant impact on participants' expectations for certain data practices. Other user characteristics elicited in the survey had no statistically significant impact.

For analysis, we considered user characteristics as naturally-occurring, continuous IVs. The DVs were the user expectations for the 17 data practices. Distributions of the individual DVs were non-normal. Therefore, we considered them as two-level nominal variables (Yes, No) and built a nominal logistic regression model for each DV. We assessed internal consistency of summated scale responses using Cronbach's α. For responses to online privacy concern, privacy concept familiarity, privacy knowledge, privacy protective behavior and negative online experience scales, reliability estimates were 0.88, 0.91, 0.63, 0.78, 0.68 respectively. For building regression models, we standardized IV values. To avoid biasing the model due to collinearity of IVs, we computed bivariate non-parametric Spearman rank correlations between IVs and subsequently excluded IVs that had moderate or higher correlation (>0.5). Privacy concept familiarity and privacy-protective behavior were removed from regression models as they correlated with privacy knowledge. Website familiarity and whether the participant has an account were removed because they correlated with the amount of recent use. Our analysis of initial regression models showed that, among demographic variables, only age accounted for a significant amount of variance. Therefore other demographics were removed to improve reliability of the regression models.

As a result, each of the 17 final regression models contained six IVs: privacy knowledge, privacy concern, negative online experience, age, trust in website and recent use. Table 6 lists the user characteristics (IV) and regression models in which the IV was statistically significant in predicting user expectation (DV). Below, we explain the user characteristics (IVs) that can significantly predict user expectations (DVs).

Privacy Knowledge: An individual's privacy knowledge impacts user expectations. Specifically, privacy knowledge can impact if a user expects the collection of health information from unregistered users. An individual with a one unit increase on the privacy knowledge scale is two times more likely to expect that a website will not collect health information without an account. Privacy familiarity and privacy protective behavior correlated with privacy knowledge, and are likely to impact users' expectations in a similar way. Recall that users expect websites, especially non-health websites, to collect health information only when they have an account. If a website did collect health information without an account, there would be a mismatch in expectations.

Privacy Concern: Individuals with higher online privacy concern (IUIPC [27]) expect data practices to be more privacy invasive. Specifically, individuals with one unit increase in online privacy concern are twice as likely to expect that a website will collect current location information when users have an account. They are ∼1.6 times more likely to expect that a website will share contact and current location information for core purposes. Although, most users in our study expect such collection and sharing practices, the segment of users with higher privacy concern are even more likely to expect such practices.

Age: Individuals' age impacts expectations regarding deletion; with one unit increase in age, they are ∼1.8 times more likely to expect that a website will not allow deletion of user data. Older users correctly expect websites not to permit deletion of user data. Hence, the likelihood of mismatch is higher in case of younger users.

Trust in Website: User perception of a website's trustworthiness impacts expectations regarding sharing and deletion data practices. With a one unit increase in trust, individuals are ∼1.7 times more likely to expect that a website will not share health and financial information for other purposes. They are 1.5 times more likely to expect that a website will share location information for core purposes. Lastly, individuals are twice as likely to expect the website to allow deletion of user data. Although, users' expectations based on trust hold for sharing practices, their expectations for deletion does not match reality.

Recent Use: Participants self-reported use of the website in the last 30 days impacts expectations regarding three data practices. With one unit increase in usage, individuals are 1.6 times more likely to expect that a website will not collect current location information from registered users. Individuals are 1.5 times more likely to expect that the website will not share contact information for core purposes. Lastly, individuals are 1.6 times more likely to expect that website will not allow deletion. User expectations are likely to vary similarly based on website familiarity and whether the participant has an account, because both correlated with the amount of recent use. These results confirm our hypothesis that users who have more access to a website have different expectations. However, it is not always true that their expectations are more accurate. For instance, their expectations regarding deletion are more accurate, but expectations regarding sharing are not.

4.4 Matched and Mismatched Expectations

As shown in Figure 5, overall, expected and unexpected data practices varied for different information types, and collection and sharing scenarios. We analyzed mismatches when websites explicitly disclosed their data practices, as well as when websites were unclear or did not address the data practices. When data practices were explicit, we observed three important mismatches. Collection of contact information without an account was mainly a Yes–No mismatch, that is, participants did not expect websites to collect information, but websites did. Similarly, collection of financial information without an account was a Yes–No mismatch. Sharing of contact information for other purposes was also a mismatch, but a No–Yes mismatch, that is, participants pessimistically and incorrectly thought that websites would share their contact information. For the remaining data practices, participants' expectations either predominately matched website practices or the level of match was equal to the level of mismatch.

For the data deletion practice, 32% of participants expected websites to allow full deletion, but only 19% of the analyzed websites allow it. Similarly, 48% expected partial deletion, but only 12% of websites permit it. However, about 20% of the participants thought that websites would not allow deletion of any data and 19% of the websites do not allow deletion of any data. Participants' expectations were similar across the three website types. There is a mismatch in expectations regarding deletion – participants seem to expect websites to allow deletion more than websites actually do.

As we discussed earlier, the number of data practices that are unclear or not addressed in a privacy policy can be high. As shown in Figure 5, websites mostly do not address data

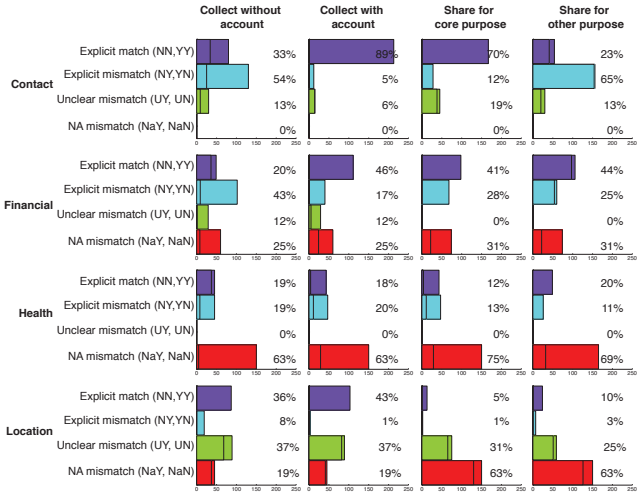


Figure 5: Matches and mismatches in user expectations. Explicit match or mismatch occurs when websites are clear about their data practice. When practice is unclear or not addressed, mismatch is not evident.

practices regarding health information. In contrast, they are mostly unclear or do not address data practices regarding location information. Considering Yes–No mismatches to be more privacy invasive, let us assume that a website engages in a data practice when its disclosure is unclear or not addressed. For health information practices, this results in mainly Yes–No mismatches for all scenarios. However, for location information practices, it results in No–Yes mismatches.

5. DISCUSSION

We identified data practices that do not match user expectations. Our results show that the number of mismatches can be substantial depending on the data practice, and that mismatched expectations vary significantly based on the type of website, as well as user characteristics, such as privacy concern, knowledge, and age. Below, we discuss potential limitations of our study, followed by implications of our results.

5.1 Limitations

We conducted an online study to elicit user expectations. This line of research could benefit from further in-lab studies conducted under more controlled conditions. We compared user expectations with websites' data practices, as disclosed in websites' privacy policies. However, how a website actually handles personal information of their users could potentially be different, but this is difficult to assess in practice.

We recruited participants from Amazon Mechanical Turk. Compared to the general population, they may have higher privacy concern [20], computer knowledge and exposure to privacy-related surveys. Our participants were limited to the United States, and it would be interesting to study expectations of users in other countries or cultures. Nevertheless, our results show that even for potentially more privacy-concerned MTurk participants privacy expectations can be at odds with websites' data practices.

User characteristic (IV)	User expectation (DV)	Model			IV		
		R^2	$\chi^2(6, N{=}240)$	p	Odds(No)	$\chi^2(1, N{=}240)$	p
Privacy knowledge	Collect health info without account	0.10	14.52	0.024	2.09	7.60	0.0058
Privacy concern	Collect location info with account	0.13	13.80	0.0319	0.49	7.22	0.0072
	Share contact info for core purpose	0.09	18.47	0.0052	0.64	5.94	0.0148
	Share location info for core purpose	0.08	15.34	0.0177	0.58	7.67	0.0056
Age	Allow deletion	0.13	30.53	<0.0001	1.77	10.88	0.0010
Trust in website	Share location info for core purpose	0.08	15.34	0.0177	0.65	4.44	0.0352
	Share financial info for other purpose	0.07	21.33	0.0016	1.80	16.82	<0.0001
	Share health info for other purpose	0.05	14.54	0.0241	1.68	1124	0.0008
	Allow deletion	0.13	30.53	<0.0001	0.53	13.64	0.0002
Recent use	Collect location info with account	0.13	13.80	0.0319	1.56	4.01	0.0451
	Share contact info for core purpose	0.09	18.47	0.0052	1.50	6.67	0.0098
	Allow deletion	0.13	30.53	<0.0001	1.56	7.83	0.0051

Table 6: Regression models in which specific user characteristics (IV) significantly impact user expectations (DV). *Odds(No)* indicates, for one unit increase in the IV value, the increase in likelihood that a user will not expect a website to engage in that data practice (*Odds(Yes)*=1 / *Odds(No)*).

We studied collection, sharing and deletion data practices. We asked participants (n=240) if they wanted to know about other data practices; nearly half did not (47.5%). Among the rest, the top three requests were as follows: Participants wanted additional details about sharing (14%). They wanted to know with whom – partners, affiliates and third-parties – their data was being shared. They wanted to know about data security (12%) and how long their data was retained (7%). We plan to extend our research to cover these and other data practices of interest in the future.

We further plan to study more website categories. However, eliciting user expectations for websites with broad or multiple purposes, for example search or social networking websites, is challenging. For example, users may use Google.com for searching, shopping, directions, etc. Along similar lines, it would be interesting to study how accessing multiple websites via a single sign-on impacts expectations. We are studying the impact of additional expectation types, such as the "should" (Ideal) expectation type. Lastly, we are investigating expectations and mismatches in the context of mobile and Internet of Things data practices.

5.2 Highlighting Unexpected Practices

As we discussed earlier, simplified user-facing privacy notices [49] could complement comprehensive privacy policies. Existing simplified privacy notices, for example privacy nutrition labels [22], although an improvement over privacy policies, are themselves too complex. By identifying mismatches in users' privacy expectations, one could selectively highlight or display elements of a privacy nutrition label or other notice format that are most relevant to users. Our results suggest that the number of mismatches is small compared to the total number of website data practices. Thus, likely unexpected data practices should be especially emphasized, and the overall amount of provided privacy information could potentially be reduced. Effectiveness of such highlighting, however, needs to be validated with end users. Different types of mismatches (Yes–No vs. No–Yes) could have different consequences on user privacy, and privacy notices should consider that as well.

Although website operators could themselves generate simplified notices, the low adoption of simplified and standardized notice mechanisms [11] indicates that many website operators may not do so. An alternative approach is for a third-party to highlight unexpected data practices based on mismatched expectations. For example, a browser extension could generate and display a simplified notice [48, 55]. Such a notice could highlight snippets of text from the natural language privacy policy, corresponding to mismatched data practices. Currently third-party browser extensions, such as Ghostery[1] and Privacy Badger,[2] generate and display information regarding online tracking practices. Similarly, a third-party browser extension could display information regarding unexpected data practices. Extensions could use just-in-time notifications or static icons that users can click to gain more information. At installation time, the extension could gather user characteristics such as privacy knowledge, concerns and demographics in order to tailor which practices are emphasized to individual users.

Organizations could also use our approach to obtain a competitive advantage by making their website's data practices and privacy policies easier to understand. In the past, organizations such as Google, have tried to organize information within their policy along dimensions that are important to people, with the intent of making information easier to access. Mismatches in expectations are important, and highlighting them can aid in such efforts. Regulatory agencies such as the Federal Trade Commission work on protecting users' privacy, and mismatched expectations could indicate to them important public policy issues that need attention.

A number of factors are contributing to the growing complexity of website privacy policies. In particular, as websites collect and share more data, policies have to describe more diverse and often more complex data practices. With a growing number of ways to access websites – for example, computers, smart phones, smart cars etc. – policies have to describe data practices that may vary by access mechanisms. Hence, simplified privacy notices that reduce the amount of information to be processed could significantly improve the likelihood of users understanding relevant elements of privacy policies.

[1] www.ghostery.com
[2] www.eff.org/privacybadger

5.3 Generating Simplified Notices

Privacy policies could be potentially simplified or shortened by highlighting data practices that do not match user expectations. For example, consider BankofAmerica.com's privacy policy, which is one of the 16 policies in our study. A full website privacy notice has to include information about all the 17 data practices that we studied. However, for six data practices, user expectations match the website's data practices. Focusing on mismatches, it may be sufficient to highlight those 11 data practices, which is 35% less information. We could further simplify the notice by prioritizing the impact of mismatches. For example, if we determine that Yes–No mismatches are more concerning to users than No–Yes mismatches, the notice could highlight five Yes–No mismatches among the 11 mismatches, which results in 70% less information. This approach could be used in a layered notice approach [49] to determine what practices to include in a high-level summary of the full privacy policy.

Our results indicate that the data practices users expect, as well as respective mismatched expectations, vary significantly by website type. For example, users expect health websites to collect health information, but not finance websites. Therefore, website type could serve as a simple and practical feature to contextualize privacy notices in order to highlight those practices unexpected for the respective website type. Third party tools or browser extensions could further predict, based on website type, which data practices my be unexpected and emphasize or warn about them. Practices that are likely expected for websites of a given type, may not require explicit warnings. For example, in case of the BankofAmerica.com banking website, the extension could signal a mismatch with regard to the website's collection of health information. However, such a warning would not be necessary for health website that collects health information, as most users seems to expect such a practice.

User expectations and mismatches further vary based on user characteristics. Hence, we could personalize privacy notices based on user characteristics. For example, younger users are significantly more likely to expect a website to allow deletion of user data. Hence, when the website does not allow deletion, the likelihood of a mismatch is higher in case of younger users. Thus, privacy decision support tools could highlight a mismatch for younger users only.

Note, that the goal is not to replace or substitute privacy policies, but rather complement them with more targeted notices and tailored warnings to make users aware of those data practices they likely do not expect.

5.4 Semantics and Impact of Mismatches

We discussed mismatches concerning "will" expectations, corresponding to Miller's "Expected" expectation type [32]. We can extend our analysis to additionally include "should" expectations, which are more subjective, as they describe expectations of what would be "Ideal" [32], and are therefore closer to preferences of desired privacy. Users may answer Yes or No to whether a website *should* engage in a data practice. Considering "should" expectations in addition to "will" expectations, would add an additional dimension to the assessment of the implications stemming from matched or mismatched expectations.

For instance, consider when a user's "will" expectation matches the website's data practices (Yes–Yes). When combined with the "should" expectation type, only Yes–Yes–Yes is a perfect match, whereas Yes–Yes–No is a mismatch, i.e., users may expect the practice but prefer it to be different. For example, for data collection, a Yes–Yes–No indicates that a user is correctly aware that a website will collect information, but feels that it should not. The user may continue to use the website due to lack of awareness of other websites that do not collect information. It may also imply market failure due to monopoly or due to all websites in the website category being equally privacy invasive. An example of such market failure may be search engine websites; although users may know that Google's search website collects certain information about them, they may continue to use Google for convenience and utility reasons, despite the availability of privacy-friendly alternatives (e.g., DuckDuckGo.com).

Similarly, in case of a mismatch due to a website engaging in unexpected practices, the "should" expectation type may change the meaning of the mismatch. For example, when a Yes–No mismatch is combined with a "should" expectation. In a Yes–No–No mismatch, users both incorrectly think that a website will not engage in a data practice and feel that it should not. They may decide to use the website and lose data privacy. For Yes–No–Yes, users want the website to engage in a practice, but do not expect it to do so at the moment. For instance, users may want a website to provide personalized services based on their data. In this scenario, users may decide not to use the website and lose utility, but not data privacy.

The examples discussed above demonstrate the importance and potential of distinguishing and capturing the meaning of different expectation types in privacy research. In the case of website privacy notices, by distinguishing between expectation types, we may be able to better identify user needs and display appropriate information. For example, in case of a Yes–Yes–No mismatch, a privacy tool could display alternative websites with more privacy-friendly practices. In case of a Yes–No–Yes mismatch, such a tool could display whether an opt-in option for personalization is available.

Lastly, in addition to the semantics of mismatches, we need to consider which mismatches matter to users. Some mismatches may surprise users, but not really concern them. When designing simplified notices, we could focus on the subset of mismatches that are concerning to users.

6. CONCLUSION

We identified mismatches in user expectations regarding online data practices. Further, we identified factors that impact such mismatches. We believe that emphasizing such mismatches in privacy notices could help users make better privacy decisions. Further, given the small number of mismatches compared to the overall number of data practices, it could be possible to generate simplified user-facing privacy notices as summaries of full privacy policies. Based on the factors that impact mismatches, we identified future research opportunities for contextualizing and personalizing privacy notices and privacy tools to ameliorate the effect of mismatched expectations.

7. ACKNOWLEDGMENTS

This research has been partially funded by the National Science Foundation under grants CNS-1330596 and CNS-1012763. The authors thank Birendra Jha, Ivan Ruchkin and members of the Usable Privacy Policy Project for their useful feedback.

8. REFERENCES

[1] M. S. Ackerman, L. F. Cranor, and J. Reagle. Privacy in e-Commerce: Examining User Scenarios and Privacy Preferences. In *Proc. EC '99*, pages 1–8. ACM, 1999.

[2] A. Acquisti, L. Brandimarte, and G. Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, Jan. 2015.

[3] I. Altman. The environment and social behavior: Privacy, personal space, territory, and crowding. 1975.

[4] Amazon. Alexa website rankings. `http://www.alexa.com`, 2015.

[5] Amazon. Mechanical turk. `https://www.mturk.com/`, 2015.

[6] J. Bhatia, T. D. Breaux, and F. Schaub. Mining privacy goals from privacy policies using hybridized task recomposition. *ACM Trans. Softw. Eng. Methodol.*, 25(3):22:1–22:24, May 2016.

[7] N. M. Bradburn, S. Sudman, and B. Wansink. *Asking Questions: The Definitive Guide to Questionnaire Design – For Market Research, Political Polls, and Social and Health Questionnaires*. John Wiley & Sons, 2004.

[8] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5, 2011.

[9] F. Cate. The Limits of Notice and Choice. *IEEE Security & Privacy*, 8(2):59–62, Mar. 2010.

[10] Center for Information Policy Leadership. Ten steps to develop a multilayered privacy policy, 2006.

[11] L. F. Cranor. Necessary but Not Sufficient: Standardized Mechanisms for Privacy Notice and Choice. *Journal on Telecommunications and High Technology Law*, 10:273, 2012.

[12] L. F. Cranor, K. Idouchi, P. G. Leon, M. Sleeper, and B. Ur. Are they actually any different? comparing thousands of financial institutions' privacy practices. In *Proc. WEIS 2013*, 2013.

[13] J. B. Earp, A. I. Antón, L. Aiman-Smith, and W. H. Stufflebeam. Examining internet privacy policies within the context of user privacy values. *Transactions on Engineering Management,*, 52(2):227–237, 2005.

[14] Federal Trade Commission. Internet of things: Privacy & security in a connected world. FTC staff report, Jan. 2015.

[15] M. C. Gilly, W. L. Cron, and T. E. Barry. The expectations-performance comparison process: An investigation of expectation types. In *Proc. Conf. Consumer Satisfaction, Dissatisfaction, and Complaining Behavior*, pages 10–16, 1983.

[16] J. Gomez, T. Pinnick, and A. Soltani. Know privacy. Technical report, UC Berkeley School of Information, 2009. `http://knowprivacy.org/report/KnowPrivacy_Final_Report.pdf`.

[17] R. M. Hogarth. *Judgement and Choice: The Psychology of Decision*. John Wiley & Sons, 1987.

[18] C. Jensen and C. Potts. Privacy policies as decision-making tools: An evaluation of online privacy notices. In *Proc. CHI '04*, pages 471–478. ACM, 2004.

[19] A. N. Joinson, U.-D. Reips, T. Buchanan, and C. B. P. Schofield. Privacy, Trust, and Self-Disclosure Online. *Human-Computer Interaction*, 25(1):1–24, Feb. 2010.

[20] R. Kang, S. Brown, L. Dabbish, and S. B. Kiesler. Privacy attitudes of mechanical turk workers and the us public. In *Proc. SOUPS '14*, pages 37–49, 2014.

[21] R. Kang, N. Fruchter, L. Dabbish, and S. Kiesler. "my data just goes everywhere:" user mental models of the internet and implications for privacy and security. In *Proc. SOUPS '15*. USENIX, 2015.

[22] P. G. Kelley, J. Bresee, L. F. Cranor, and R. W. Reeder. A nutrition label for privacy. In *Proc. SOUPS '09*. ACM, 2009.

[23] P. G. Leon, B. Ur, Y. Wang, M. Sleeper, R. Balebako, R. Shay, L. Bauer, M. Christodorescu, and L. F. Cranor. What matters to users? factors that affect users' willingness to share information with online advertisers. In *Proc. SOUPS '13*. ACM, 2013.

[24] J. Lin, S. Amini, J. I. Hong, N. Sadeh, J. Lindqvist, and J. Zhang. Expectation and Purpose: Understanding Users' Mental Models of Mobile App Privacy through Crowdsourcing. In *Proc. UbiComp '12*. ACM, 2012.

[25] F. Liu, R. Ramanath, N. Sadeh, and N. A. Smith. A step towards usable privacy policy: Automatic alignment of privacy statements. In *Proc. COLING 2014*, pages 884–894, 2014.

[26] Y. Liu, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Analyzing facebook privacy settings: User expectations vs. reality. In *Proc. IMC '11*, pages 61–70. ACM, 2011.

[27] N. K. Malhotra, S. S. Kim, and J. Agarwal. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research*, 15(4):336–355, 2004.

[28] S. T. Margulis. On the Status and Contribution of Westin's and Altman's Theories of Privacy. *Journal of Social Issues*, 59(2):411–429, June 2003.

[29] G. Marx. Murky conceptual waters: The public and the private. *Ethics and Information technology*, pages 157–169, 2001.

[30] A. M. McDonald and L. F. Cranor. The cost of reading privacy policies. *ISJLP*, 4, 2008.

[31] A. M. McDonald, R. W. Reeder, P. G. Kelley, and L. F. Cranor. A comparative study of online privacy policies and formats. In *Proc. PETS 2009*, pages 37–55. Springer, 2009.

[32] J. A. Miller. Studying satisfaction, modifying models, eliciting expectations, posing problems, and making meaningful measurements. *Conceptualization and Measurement of Consumer Satisfaction and Dissatisfaction*, pages 72–91, 1977.

[33] G. R. Milne and S. Bahl. Are there differences between consumers' and marketers' privacy expectations? a segment and technology level analysis. *Public Policy & Marketing*, 29(1), 2010.

[34] H. Nissenbaum. Privacy as contextual integrity. *Washington Law Review*, 79:119, 2004.

[35] H. Nissenbaum. *Privacy in Context - Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2009.

[36] P. A. Norberg, D. R. Horne, and D. A. Horne. The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs*, 41(1):100–126, 2007.

[37] Office of the Australian Information Commissioner. Community attitudes to privacy survey, 2013.

[38] J. S. Olson, J. Grudin, and E. Horvitz. A study of preferences for sharing and privacy. In *Proc. CHI '05*, pages 1985–1988. ACM, 2005.

[39] L. Palen and P. Dourish. Unpacking "privacy" for a networked world. In *Proc. CHI '03*, pages 129–136. ACM, 2003.

[40] G. Paolacci and J. Chandler. Inside the turk understanding mechanical turk as a participant pool. *Current Directions in Psychological Science*, 23(3):184–188, 2014.

[41] G. Paolacci, J. Chandler, and P. G. Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419, 2010.

[42] I. Pollach. What's wrong with online privacy policies? *Commun. ACM*, 50(9):103–108, Sept. 2007.

[43] President's Concil of Advisors on Science and Technology. Big data and privacy: A technological perspective. Report to the President, Executive Office of the President, May 2014.

[44] L. Rainie, S. Kiesler, R. Kang, and M. Madden. Anonymity, privacy, and security online. *PEW Research Center*, September 2013.

[45] A. Rao, F. Schaub, and N. Sadeh. What do they know about me? contents and concerns of online behavioral profiles. In *Proc. PASSAT '14*. ASE, 2014.

[46] J. Reidenberg, A. M. McDonald, F. Schaub, N. Sadeh, A. Acquisti, T. Breaux, L. F. Cranor, F. Liu, A. Grannis, J. T. Graves, et al. Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Technology Law Journal*, 30(1):39–88, 2015.

[47] J. R. Reidenberg, N. C. Russell, A. J. Callen, S. Qasir, and T. B. Norton. Privacy harms and the effectiveness of the notice and choice framework. *ISJLP*, 11, 2015.

[48] N. Sadeh, A. Acquisti, T. D. Breaux, L. F. Cranor, A. M. McDonald, J. R. Reidenberg, N. A. Smith, F. Liu, N. C. Russell, F. Schaub, et al. The usable privacy policy project. Technical report, CMU-ISR-13-119, Carnegie Mellon University, 2013.

[49] F. Schaub, R. Balebako, A. L. Durity, and L. F. Cranor. A design space for effective privacy notices. In *Proc. SOUPS '15*, pages 1–17. USENIX, 2015.

[50] J. E. Swan and I. F. Trawick. Satisfaction related to predictive vs. desired expectations. *Refining Concepts and Measures of Consumer Satisfaction and Complaining Behavior*, pages 7–12, 1980.

[51] Y. Wang, H. Xia, and Y. Huang. Examining american and chinese internet users' contextual privacy preferences of behavioral advertising. In *Proc. CSCW '16*. ACM, 2016.

[52] G. B. Willis. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Sage Publications, 2004.

[53] S. Wilson, F. Schaub, R. Ramanath, N. Sadeh, F. Liu, and N. A. Smith. Crowdsourcing annotations for websites' privacy policies: Can it really work? In *WWW*, 2016.

[54] V. A. Zeithaml, L. L. Berry, and A. Parasuraman. The nature and determinants of customer expectations of service. *Academy of Marketing Science*, 21(1):1–12, 1993.

[55] S. Zimmeck and S. M. Bellovin. Privee: An architecture for automatically analyzing web privacy policies. In *Proc. USENIX Security '14*, 2014.

APPENDIX

A. DEFINITION OF INFORMATION TYPES

Contact Information: Examples include (but are not limited to) email address, postal address, phone number, home phone number, etc.

Current location: Current, real-time location of a user accessing the website (city-level or more precise)

Health information: Examples include (but are not limited to) user's medical history, family medical history, user's health insurance information, etc.

Financial information: Examples include (but are not limited to) bank account details, credit/debit card numbers, credit ratings/history etc.

B. SURVEY QUESTIONNAIRE

The complete survey questionnaire is reproduced on the next pages.

[Interview/Survey Questionnaire]

Thank you for your interest in our study.

Your answers are important to us. Please read the instructions carefully so that you can answer our questions as accurately as possible. Take your time in reading and answering the questions.

Peoples' opinions about websites may or may not vary depending on the type of website (news, health, finance etc.) and past experience (not heard of website, heard of, not visited, visited etc.)
While answering questions about a website, **think about your interactions only with the website**. Your interactions could be through a computer, mobile phone or other device. Ignore any interactions with mobile apps, physical stores, businesses or other websites related to the website.
For each website listed below, select the option that best indicates your answer.

	I have not heard of it	I have heard of it, but not visited it	I have visited it, but not in the last 3 months	I have visited it in the last 3 months	Don't know/Not sure
[website]	○	○	○	○	○

I would like to understand your opinions regarding Internet websites. For any question, it is okay to say that you don't know the answer. If you are guessing an answer, please say so. It would be very helpful, if you explain your reasoning behind your answers.

[For each website assigned to a participant, ask the following questions]

Now, I would like your opinions regarding [website name] website. Please interact with the website (provide URL) for 2-3 minutes and get familiar with it. Please let me know when you are ready to provide your opinions.

1. As far as you can recall, have you used any websites similar to [website name]?
   Yes (please specify) / No

[Omit questions 2 and 3 if the participant has not used the website]

2. I would like you to think about the last time you visited [website name]. As far as you can recall, what did you do on the website?
3. What other things have you done on this website?

To help you answer my questions, I will explain a few terms. Please use this handout to follow along. You can refer back to the handout at any time.

[Provide handout containing definitions for contact/health/financial/current location information]
[Read definitions for contact/health/financial/current location information]

4. Consider the following scenario to answer the next question.

   Imagine that you are browsing [website name] website. You **do not have a user account** on [website name], that is, you have not registered or created an account on [website name].

   What is the likelihood that [website name] would **collect your information** in this scenario? Each row in the table below, lists a specific type of information about you. For each information type, select the likelihood that [website name] would collect that information in the scenario described above.

		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Collects your **Contact information**	Email address	○	○	○	○
	Postal address	○	○	○	○
	Phone number	○	○	○	○
	Other Please specify	○	○	○	○

		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Collects your **Health information**	Medical history	○	○	○	○
	Health insurance information	○	○	○	○
	Other Please specify	○	○	○	○

		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Collects your **Financial information**	Bank account details	O	O	O	O
	Credit or debit card number	O	O	O	O
	Credit rating	O	O	O	O
	Other Please specify	O	O	O	O

		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Collects your **Location information**	Current location (city-level or more precise)	O	O	O	O

5. What leads you to think that [website name] would collect your information when you do not have an account? Please explain.
6. Now, consider an alternate scenario.

   Imagine that **you have a user account** on [website name], and you **have logged in** to your account while browsing [website name].

   What is the likelihood that [website name] would **collect your information** in this scenario?
   Each row in the table below, lists a specific type of information about you. For each information type, select the likelihood that [website name] would collect that information in the scenario just described.

		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Collects your **Contact information**	Email address	O	O	O	O
	Postal address	O	O	O	O
	Phone number	O	O	O	O
	Other Please specify	O	O	O	O

		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Collects your **Health information**	Medical history	O	O	O	O
	Health insurance information	O	O	O	O
	Other Please specify	O	O	O	O

		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Collects your **Financial information**	Bank account details	O	O	O	O
	Credit or debit card number	O	O	O	O
	Credit rating	O	O	O	O
	Other Please specify	O	O	O	O

		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Collects your **Location information**	Current location (city-level or more precise)	O	O	O	O

Thank you. As you may know, companies that own websites may handle information collected on websites in different ways. Some companies share the collected information with other companies, and some companies do not share. Companies may have to share your information in order to provide you a service that you requested on a website.

7. In your opinion, what services can you get from [website name]? Please explain.

8. In order to provide you services, [website name] may have to share your information with other companies. In your opinion, what are those companies, if at all any? Please explain.

9. A website may share your information for purposes unrelated to providing you a service that you requested from the website. What do you think are such unrelated purposes for which [website] can share your information? Please explain.

Before sharing your information, companies may or may not ask for your permission. Some companies assume that the permission is implied because you are using the website. Other companies may explicitly ask you for permission before sharing information, for example, via an explicit written or oral consent.

10. Consider the following scenario to answer the next question.

Imagine that [website name] is sharing your information with another company, but **only for the purpose of providing you a service you requested** on [website name]. Since [website name] has to provide you a service that you requested, [website name] assumes that it has your permission to share information, that is, your permission is implied. [Website name] will share only the information required to provide you the requested service.

What is the likelihood that [website name] would **share your information** with your implied permission in this scenario?

		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Shares your **Contact information**	Email address	O	O	O	O
	Postal address	O	O	O	O
	Phone number	O	O	O	O
	Other Please specify	O	O	O	O
		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Shares your **Health information**	Medical history	O	O	O	O
	Health insurance information	O	O	O	O
	Other Please specify	O	O	O	O
		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Shares your **Financial information**	Bank account details	O	O	O	O
	Credit or debit card number	O	O	O	O
	Credit rating	O	O	O	O
	Other Please specify	O	O	O	O
		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Shares your **Location information**	Current location (city-level or more precise)	O	O	O	O

11. Consider the following alternate scenario to answer the next question.

Imagine that [website name] is sharing your information with another company for **a purpose unrelated to providing you a service you requested**. Since you are using [website name], it assumes that it has your permission, that is implied permission, to share your information for any purpose.

What is the likelihood that [website name] would **share your information** with your implied permission in this scenario?

		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Shares your **Contact information**	Email address	O	O	O	O
	Postal address	O	O	O	O
	Phone number	O	O	O	O
	Other Please specify	O	O	O	O
		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Shares your **Health information**	Medical history	O	O	O	O
	Health insurance information	O	O	O	O
	Other Please specify	O	O	O	O
		Likely	Somewhat	Somewhat	Unlikely

			likely	unlikely	
Shares your **Financial information**	Bank account details	O	O	O	O
	Credit or debit card number	O	O	O	O
	Credit rating	O	O	O	O
	Other Please specify	O	O	O	O
		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Shares your **Location information**	Current location (city-level or more precise)	O	O	O	O

Thank you. As you may know, websites may allow users to delete or remove their data from the website e.g. by closing an account. Allowing users to edit or modify their data is not same as deleting data.

12. Do you think that [website name] would allow you to delete your personal data?
    - O Yes, it will allow me to delete all of my data
    - O Yes, but it will only allow me delete some of my data
    - O No, it will not allow me to delete my data

13. We discussed data practices such as collection and sharing of four types of information, and also deletion of information. What else would you like to know about [website name]?

[End of the interview]

Thank you. That was all I had to discuss. Would you care to add anything?

Thank you. Please take a few minutes to fill out the following questionnaire. That would be the end of our study.

Different users may have different opinions regarding websites. To help us understand how user opinions vary, please answer the following questions.

Please tell us about your experience with <u>[website name]</u> website.

As far as you know, do you have a user account on the website?
( ) Yes, I have an account
( ) No, I don't have an account
( ) Not sure

How many times have you visited the website in the last 30 days? Exclude the visit as part of today's study.
(Please specify a number equal to or greater than zero) _____

In your opinion, how much have you used the website in the last 30 days? Exclude use as part of today's study.

1 – Not at all	2 – Very little	3 - Somewhat	4 – Quite a bit	5 – A great deal
O	O	O	O	O

Do you know someone else who uses the website?
( ) Yes, I know someone
( ) No, I don't know anyone
( ) Not sure

In your opinion, how familiar are you with the website?

1 – Not at all	2 – Slightly	3 - Somewhat	4 – Moderately	5 – Extremely
O	O	O	O	O

In your opinion, how trustworthy is the website?

1 – Not at all	2 – Slightly	3 - Somewhat	4 – Moderately	5 – Extremely
O	O	O	O	O

As far as you know, do you have a user account on a website similar to [website name]?
( ) Yes, I have an account
( ) No, I don't have an account
( ) Not sure

Please tell us about your background.

What is your year of birth (4-digit, yyyy format)?
_____

What is your gender?
( ) Male ( ) Female ( ) Decline to answer

Which of the following best describes your primary occupation?
[List of occupations here]

Which of the following best describes your highest achieved education level?
[List of education levels here]

Do you have a college degree or work experience in computer science, software development, web development or similar computer-related fields?
( ) Yes ( ) No ( ) Decline to answer

Do you currently work or reside in the state of California?
( ) Yes ( ) No ( ) Decline to answer

While using the Internet, have you ever done any of the following things? Please check all that apply.
[ ] Used a temporary username or email address
[ ] Used a fake name or untraceable username
[ ] Given inaccurate or misleading information about yourself
[ ] Set your browser to disable or turn off cookies
[ ] Cleared cookies and browser history
[ ] Used a service that allows you to browse the web anonymously, such as a proxy server, Tor software, or a virtual private network
[ ] Encrypted your communications
[ ] Decided not to use a website because they asked for your real name
[ ] Deleted or edited something you posted in the past
[ ] Asked someone to remove something that was posted about you online
[ ] Used a public computer to browse anonymously

How would you rate your familiarity with the following concepts or tools?

	I've never heard of this.	I've heard of this but I don't know what it is.	I know what this is but I don't know how it works.	I know generally how this works.	I know very well how this works.
IP address	O	O	O	O	O
Cookie	O	O	O	O	O
Incognito mode / private browsing mode in browsers	O	O	O	O	O
Encryption	O	O	O	O	O
Proxy server	O	O	O	O	O
Secure Sockets Layer (SSL)	O	O	O	O	O
Tor	O	O	O	O	O
Virtual Private Network (VPN)	O	O	O	O	O
Privacy settings	O	O	O	O	O

Please indicate whether you think each statement is true or false. Please select "I'm not sure" if you don't know the answer.

	True	False	I'm not sure
Incognito mode / private browsing mode in browsers prevents websites from collecting information about you.	O	O	O
Website cookies can store users' logins and passwords in your web browser.	O	O	O
Tor can be used to hide the source of a network request from the destination	O	O	O
A VPN is the same as a Proxy server.	O	O	O
IP addresses can always uniquely identify your computer.	O	O	O
HTTPS is standard HTTP with SSL to preserve the confidentiality of network traffic.	O	O	O
A request coming from a proxy server cannot be tracked to the original source.	O	O	O

In order to protect your personal information, how often have you done the following?

Check that a website is secure before providing personal information.

1 - Never	2 - Rarely	3 - Sometimes	4 - Often	5 - Always
O	O	O	O	O

Ask public or private sector organizations why they need your information.

1 - Never	2 - Rarely	3 - Sometimes	4 - Often	5 - Always
O	O	O	O	O

Read privacy policies and notifications before providing personal information.

1 - Never	2 - Rarely	3 - Sometimes	4 - Often	5 - Always
O	O	O	O	O

As far as you know, have you ever had any of these bad experiences as a result of your online activities?

	Yes	No
Something happened online that led you into physical danger	O	O
Been stalked or harassed online (sexually harassed, physically threatened)	O	O
Got into trouble with local authorities, or government because of your online activities	O	O
Experienced trouble in a relationship between you and a family member or a friend because of something you posted online	O	O
Had your personal information leaked by a company	O	O
Lost a job opportunity or educational opportunity because of something you posted online or someone posted about you online	O	O
Had your reputation damaged because of something that happened online	O	O
Been the victim of an online scam and lost money	O	O
Had important personal information stolen such as your Social Security Number, your credit card, or bank account information	O	O
Something else bad happened (please explain)	O	O

You are almost done. Please share your opinion about Internet consumer experience.
Please indicate how much you agree or disagree with the following statements:

Consumer online privacy is really a matter of consumers' right to exercise control and autonomy over decisions about how their information *is collected, used, and shared.*

Strongly disagree 1 2 3 4 5 6 7 **Strongly agree**

Consumer control of personal information lies at the heart of consumer privacy.

Strongly disagree 1 2 3 4 5 6 7 **Strongly agree**

I believe that online privacy is invaded when control is lost or unwillingly reduced as a result of a marketing transaction.

Strongly disagree 1 2 3 4 5 6 7 **Strongly agree**

Companies seeking information online should disclose the way the data are collected, processed, and used.

Strongly disagree 1 2 3 4 5 6 7 **Strongly agree**

A good consumer online privacy policy should have a clear and conspicuous disclosure.

Strongly disagree 1 2 3 4 5 6 7 **Strongly agree**

It is very important to me that I am aware and knowledgeable about how my personal information will be used.

Strongly disagree 1 2 3 4 5 6 7 **Strongly agree**

It usually bothers me when online companies ask me for personal information.

Strongly disagree 1 2 3 4 5 6 7 **Strongly agree**

When online companies ask me for personal information, I sometimes think twice before providing it.

Strongly disagree 1 2 3 4 5 6 7 **Strongly agree**

It bothers me to give personal information to so many online companies.

Strongly disagree 1 2 3 4 5 6 7 **Strongly agree**

I'm concerned that online companies are collecting too much personal information about me.

Strongly disagree 1 2 3 4 5 6 7 **Strongly agree**

Thank you for participating in our study.

Do or Do Not, There Is No Try: User Engagement May Not Improve Security Outcomes

Alain Forget*, Sarah Pearman*, Jeremy Thomas*
Alessandro Acquisti*, Nicolas Christin*, Lorrie Faith Cranor*
Serge Egelman†, Marian Harbach†, Rahul Telang*
*Carnegie Mellon University, †International Computer Science Institute
{aforget, spearman, thomasjm, acquisti, nicolasc, lorrie, rtelang}@cmu.edu
{egelman, mharbach}@icsi.berkeley.edu

ABSTRACT

Computer security problems often occur when there are disconnects between users' understanding of their role in computer security and what is expected of them. To help users make good security decisions more easily, we need insights into the challenges they face in their daily computer usage. We built and deployed the Security Behavior Observatory (SBO) to collect data on user behavior and machine configurations from participants' home computers. Combining SBO data with user interviews, this paper presents a qualitative study comparing users' attitudes, behaviors, and understanding of computer security to the actual states of their computers. Qualitative inductive thematic analysis of the interviews produced "engagement" as the overarching theme, whereby participants with greater engagement in computer security and maintenance did not necessarily have more secure computer states. Thus, user engagement alone may not be predictive of computer security. We identify several other themes that inform future directions for better design and research into security interventions. Our findings emphasize the need for better understanding of how users' computers get infected, so that we can more effectively design user-centered mitigations.

1. INTRODUCTION

Humans are critical to the security of computing systems [8]. Unfortunately, computer security problems frequently arise because of the disconnect between what users do and what is expected of them, sometimes with disastrous consequences. For example, the Conficker botnet was successfully taken down in 2009 and abandoned by its operators. Yet, six years later we can still find evidence of over one million infected machines that are attempting to re-infect other vulnerable machines [2]. This may be due to users not following elementary security precautions, such as ignoring warnings or using out-of-date software.

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22–24, 2016, Denver, Colorado.

Some suggest that greater computer security can be achieved with greater user involvement [1, 4, 5]. To help users make better security decisions, we need to identify specific insecure behaviors and understand how often and why users behave insecurely. Unfortunately, we still lack a holistic understanding of how users process and address security threats. Past work [7, 12, 15, 19] has explored how users model computer security threats and use them to make decisions. While informative, this work has largely relied on surveys or lab studies rather than users' actual computing behaviors or focused on narrow behaviors and scenarios rather than comprehensively capturing end-users' *in situ* usage. We know of no work that longitudinally examines user behavior and directly maps users' decisions and self-reported understandings to the observed security states of their machines.

As part of an ambitious research project attempting to answer these questions, we developed the Security Behavior Observatory (SBO) [14], which is a panel of participants consenting to our monitoring of their general computing behaviors, with an eye toward understanding what constitutes insecure behavior. Technically, the SBO consists of a set of "sensors" monitoring various aspects of participants' computers to provide a comprehensive overview of user activity that regularly reports (encrypted) measurements to our secure server. Our monitoring provides us with the opportunity to characterize *which* user actions led to insecure computing states. We can also directly interact with our participants to solicit insights into their behaviors that may have led to their machines' states.

We present an initial study conducted with the SBO. After observing 73 users over the course of 9 months, we conducted interviews with 15 users whose computers were in a variety of security states to better understand users' attitudes and motivations toward computer security and to understand why their computers were in a state of (in)security. Qualitative inductive thematic analysis of the interviews produced "engagement" as the overarching theme.

We found that some *engaged* users actively maintain their computers' security, while other *disengaged* users prefer to ignore or delegate security tasks. Surprisingly, we found that engaged users' computers were not necessarily more secure than those of disengaged users. Thus, for user engagement with computer security to be effective, it has to be done correctly. Otherwise, it may be better that users not even try, lest they inadvertently subvert their machines' security.

Due to the SBO population at the time, our 15 interviewees had a median age of 63 and were mostly female. This gave us a unique opportunity to examine an often understudied population. Future work will test the extent to which the theme of engagement is applicable across demographics.

Our study's primary insight is that user engagement alone may not be predictive of computer security, which challenges past assumptions [1, 4, 5]. We also found that misunderstanding computer security leads users to adopt ineffective (though perhaps rational [18]) security postures. This *in situ* finding validates similar observations that have been made previously in other security contexts [7, 18]. Finally, we also found that disengaged and engaged users seem to have distinct sets of behaviors, needs, and problems. As such, our findings suggest that both types of users may not find the same type of computer security interventions effective (i.e., one size may not fit all).

2. RELATED WORK

While the SBO is distinct in its breadth and longevity, our study's qualitative approach is similar to past work [35, 37, 38]. Our findings both confirm and build upon results from many past publications regarding users' difficulties in understanding computer security, observing their challenges, and applying software updates to eliminate vulnerabilities.

Problematic understanding of security. Wash [37] conducted interviews to investigate how people conceptualize home computer security threats. The "folk" models Wash identifies do not match actual threats, which may explain why users inadvertently put themselves at risk when ignoring or misunderstanding expert advice. Wash recommended that security advice include recommendations of appropriate actions as well as explanations of why the actions are effective. Howe et al.'s [19] literature review highlighted that users get advice from relatives, friends, or co-workers much more frequently than from experts. Ion et al. [20] found that non-experts' security advice is less likely to overlap with that of experts. Dourish et al.'s [10] interviews found that users frequently delegate security to others (e.g., friends or family) who are perceived as more knowledgeable.

Observing end users' security challenges. Multiple surveys [4, 5, 26] show that home users have difficulty securing their computers, either because of lack of knowledge or ignoring (or misunderstanding) security advice. Furnell et al.'s [15] survey respondents had difficulty understanding the security feature interfaces of various Microsoft software, despite their respondents having above average technical expertise. This parallels our observation that users more engaged with their computers' security (and perhaps more knowledgeable) may still have poor security outcomes.

A few user studies have focused on specific aspects of personal computing behavior "in the wild." Christin et al. [7] found a large number of people were willing to download, execute, and give administrative access to untrusted software, since they felt protected by their antivirus software. We also observed an over-reliance on security software and lack of attention to other advisable security practices.

Perhaps most closely related to our work is Lalonde Lévesque et al.'s [22] 50-subject, 4-month study focusing on the effectiveness of antivirus software. Participants were given an instrumented Windows 7 laptop with antivirus software. Every month, researchers collected data from the machines and met with participants to complete a survey about their computer usage. The authors found that participants with greater computer expertise were *more* at risk of being exposed to threats than less knowledgeable users, which resonates with our findings about the disconnect between user engagement in computer security and observed security issues. The SBO differs from this study in that we are observing user behavior across a broader spectrum of security- and privacy-related issues over a longer period of time.

To our knowledge, the only existing work on older users and computer security examined their knowledge of Internet hazards [16]. They found that older, particularly female, participants had less knowledge of security hazards. This motivates our work to better understand the challenges faced by the understudied population of older (female) computer users, who may be particularly vulnerable to security risks.

Trouble with updates. Timely installation of software updates and use of security software are generally considered by experts to be essential security practices. Non-experts are often aware that using security software is advisable, but are less likely to perceive updates as important for security [20].

Wash et al. [38] surveyed 37 users about their understandings of Windows updates, comparing those self-reports to participants' Windows update logs. The majority of their participants were unaware of their update settings or of when updates were being installed, and the states of their machines often did not reflect the users' intentions, for better or worse. In 12 cases, users' machines were actually *more* secure than intended, in part because some users had intended to turn off automatic updates but had not done so successfully. Other users successfully turned off automatic updates due to the inconvenience of automatic reboots, causing them to install updates less promptly. Wash et al. focused solely on update logs at the time of the interview, whereas we collected data over a longer period and cover a broader range of computer security attitudes, behaviors, and outcomes.

Comprehension is not the only updating barrier. Vaniea et al. [35] found non-experts often fail to install updates due to prior bad experiences, such as unexpected user interface changes, uncertainty about their value, and confusion about why updates to seemingly-functioning programs are needed. Fagan et al. [13] report on negative emotional responses to update messages, including annoyance and confusion.

Wash et al.'s study [38] indicates that automatic operating system updates (such as those now required by default in Windows 10) do increase the security of machines in many cases. However, they and others [6, 11, 29, 36] also highlight problems that prevent automatic and opaque update systems from being panaceas, including possible negative effects on users' understanding, trust, convenience, and/or control. Some users may object to and override such systems, preferring manual updates. Tian et al. [33] present survey results indicating that Android smartphone users preferred manual app updates for reasons including desiring control, wanting to know more about updates before installing them, preferring to apply updates only to certain apps, and wishing to work around system performance limitations (e.g., primary tasks being slowed by updates in the background).

3. SECURITY BEHAVIOR OBSERVATORY

Time- and scope-focused lab and online computer security studies have yielded valuable insights over the past 20 years. However, such experiments often do not reflect users' actual behavior in their natural environments [31], while large-scale field studies can capture users' security and privacy behaviors and challenges with greater ecological validity. This is the objective of our IRB-approved Security Behavior Observatory (SBO) [14], which longitudinally monitors user and computer behavior *in situ*. We can also interview participants to better understand their computer security attitudes and behaviors, to compare with the security state of their machines over time.

Participant recruitment. We recruit SBO participants from a university service that telephones individuals to notify them about ongoing experiments in Pittsburgh, Pennsylvania. Potential participants are contacted to complete a brief pre-enrollment survey to ensure they are over 18 and own a Windows Vista, 7, 8, or 10 personal computer. A member of our research team then calls participants to walk them through the following tasks while they are in front of their computers:

1. Read and complete a consent form, which clearly informs participants that the researchers may collect data on all activity on their computer, except personal file contents, e-mails sent or received, contents of documents on Google Docs, and bank card numbers.
2. Provide the names and e-mail addresses of others users of the computer to be instrumented, so we may obtain their consent.
3. Download and install the SBO data collection software.
4. Complete an initial demographics questionnaire.

Once all the computers' users have consented and we begin receiving data, we send participants a $30 Amazon.com gift card. Participants are then paid $10 per month their computers continue transmitting data to our server. Data transmission occurs in the background, requiring no user action. We encourage and promptly respond to questions about the study via phone or e-mail. We assert that maintaining the confidentiality of their data is our primary concern. Participants may withdraw from the SBO at any time. If we unexpectedly stop receiving data from a machine, we contact the participant to attempt to resolve the issue.

SBO data is complemented by optional questionnaires and interviews that elicit participants' perspectives on issues, events, and behaviors we observe throughout the study, for which participants receive additional compensation.

Data collection architecture. The SBO relies on a client-server architecture with several client-side sensors collecting different types of data from participants' machines [14]. Examples of collected data include processes, installed software, web browsing behavior, network packet headers, wireless network connections, Windows event logs, Windows registry data, and Windows update data. The SBO data collection architecture is implemented with multiple technologies: Java, C#, C++, Javascript, SQL, Python, PHP, WiX, and command-line batch scripts.

The SBO architecture provides security and confidentiality of participants' data as follows. All communication between users' machines and our collection server is authenticated and encrypted using unique client-server key pairs. The server only accepts connections from authenticated machines on one specific port. Finally, the data collection server is not used for analysis. Instead, a data analysis server retrieves participants' data from the collection server for long-term storage. The data analysis server is only accessible from within our institution's network. All data analysis must be performed on the server. No collected data is authorized for transfer from the data analysis server.

4. METHODOLOGY

To explore the challenges users face in protecting themselves from and addressing security problems, we conducted semi-structured interviews with a subset of SBO participants in which we asked about security-related beliefs, practices, understandings, and challenges. We chose interviews because they provide more detailed information than other methodologies (e.g., surveys). We also examined the SBO data collected from interviewees' machines to compare users' understandings of their machines' states to reality. This qualitative analysis leverages the SBO's unique nature to acquire insights that are not normally available in interview studies.

We have been enrolling SBO participants since November 2014. As of March 2016, we had collected data from 131 participant machines. As the SBO is a long-term endeavor, participants are continuously recruited and may leave any time, so the amount of data collected from each participant varies. For this paper, we analyzed data from the 73 participant computers that had sent us data for at least 3 months within a 9-month window. We sent interview invitations to 28 active participants whose machines had been regularly sending us data and who had previously responded to our e-mail and phone communications. We interviewed the 15 participants who responded to our invitations.

4.1 Interviews

We conducted 15 pre-scheduled voluntary semi-structured one hour phone interviews. We asked participants about their and others' use of their computers, computer maintenance, precautions taken to reduce computer security risks, and whether they performed a variety of insecure computing behaviors (Appendix A). We used follow-up questions to elicit information about the beliefs that informed users' security-related decisions, as in similar qualitative usable security interview studies [37]. Our questions were phrased to not imply positive or negative behaviors, not be leading, and generally avoid biases [35]. We did not ask interviewees about specific events observed on their computers, since we were concerned about participants' possible difficulty in recalling particular event details. Our questions did not allude to our knowledge of their machines' states through the SBO-collected data, to avoid influencing participants' responses.

The interviewer also established a remote session to the interviewee's computer as a common frame of reference for portions of the interview. Throughout the interview, the interviewer (with the participant's permission) verified whether or not the computer reflected the state reported by the participant. The remote session also allowed the researcher to show participants examples of Internet browser warning

messages to ask participants about their past experiences with such messages (if any), understanding of the source of such messages, and actions taken after seeing such messages. After each interview, we sent the interviewee a $50 Amazon.com gift card and a debriefing e-mail explaining the purpose of our interview and provided information on reputable free security software and tips for avoiding malware.

4.2 Qualitative Coding Methodology

Each interviewee was assigned a pseudonym. Similar to past exploratory qualitative studies in this area [32, 35, 37], we performed an inductive thematic analysis. One researcher first open-coded the transcripts, building a low-level detailed codebook. After identifying main themes, that researcher drafted a higher-level codebook of 25 codes related to a single main emergent theme. That researcher then worked iteratively with a second coder to code the interviews with that high-level codebook. The second coder was instructed to note problems with, unclear distinctions between, or possible missing codes. Both coders initially met after each coded interview to reconcile discrepancies and refine the codebook. After iteratively coding the first 8 interviews in this way, both coders agreed on a final version of the codebook. During this process, the coders agreed to adding three new codes and remove two codes by collapsing them into other existing code categories. Using the final codebook of 27 codes (Table 4 in Appendix C), both coders coded the remaining 7 transcripts independently and then met to resolve any remaining discrepancies in their codes.

Cohen's kappa, a measure of inter-coder agreement over categorical items, was calculated to be 0.64, which is considered "substantial" agreement [23]. The coders reached consensus on all codes. The reconciled codes were used for all analyses.

4.3 Examination of SBO Data

In addition to interviews, we also inspected the SBO data collected from interviewees' machines to compare participants' understanding of their computers' states (from the interviews) to the actual states of their machines. We investigated configurations and behaviors parallel to the types of interview questions asked, including:

1. Presence or absence of security software[1]
2. Presence or absence of outdated vulnerable software, particularly Adobe Flash Player and Adobe Reader
3. Presence of known malicious software or other software displaying suspicious behaviors
4. Windows update settings
5. Regularity and promptness of installation of Windows updates

Installed Software All software-related data was regularly collected from participants' machines' Windows registry, including the software name, publisher, version, and date of installation. To determine if historically-vulnerable software (e.g., Adobe Flash Player, Adobe Reader)[2] was outdated, we manually collected update and version release data from the software publishers' official websites. To determine if any of the installed software was malicious or suspicious, we manually researched the online reputation of each of around 2,900 distinct software packages found on clients' machines. In doing so, we found that the website `ShouldIRemoveIt.com` was an excellent resource for this software categorization task, since it provides scan results from multiple security software suites, as well as information about the software's known behaviors, purpose, publisher, and more. Thus, we categorized any software as *malicious* if `ShouldIRemoveIt.com` reported, "multiple virus scanners have detected possible malware." We otherwise categorized software as *suspicious* if our online research revealed any of the following:

- The software's primary purpose was to show advertising to the user (via popups, injected advertising, etc.).
- The majority of search results were complaints about the software and requests for assistance in its removal.
- The software's rating on `ShouldIRemoveIt.com` was extremely negative (based on subjective user ratings and their data on how many users remove the software).
- The software was reported as changing settings unbeknownst to the user in undesirable ways (e.g., changing default browsers, homepages, or search engines).
- The software disguised itself, such as using false names in program or plug-in lists.
- The software was known to re-install itself or to be difficult to remove.

We acknowledge that our identification of malware and suspicious software is limited by including only software listed in the registry. A deeper examination of SBO machines for more insidious and covert malware is left to future work.

Windows Updates We examined the SBO computers' operating system updating behavior in two ways. First, we determined whether Windows settings were set to automatically install updates. Second, we examined the download and installation timestamps for Windows updates and noted cases where SBO computers failed to install security updates for long periods of time or installed updates sporadically despite the computer being in regular use.

4.4 Demographics

Table 1 lists the self-reported demographics of each of the 15 interviewees. Our interviewees were a median age of 63 (SD=11), 73.3% female, and earned a median household annual income of $50,000 (SD=$83,333). This group of mostly older women provided a unique perspective of an understudied population (who may be at particular risk against security threats [16]), versus the typical demographics of other studies in our field of young and/or technically-savvy users (often university students).

All users reported performing sensitive tasks on their computers. All but one interviewee, Monica, explicitly reported performing financial tasks (e.g., online banking, e-commerce). However, Monica reported performing other sensitive activities, such as searching for medical information online. Table 3 in Appendix B summarizes interviewees' reported computer usage. This self-reported data establishes how participants *perceive* themselves using the computer.

[1] Security software is strongly recommended [34, 37].

[2] While Java could also be considered historically-vulnerable software, we excluded it since our data collection software (which is partially-written in Java) automatically updates Java upon installation on participants' machines out of necessity. Thus, Java being up-to-date is not necessarily indicative of user behavior in this case.

Pseudonym	Age	Sex	Occupation	Annual income
Agnes	63	F	Travel	$50K-$75K
Betty	68	F	Homemaker	$200K-$500K
Carl	55	M	Tradesman	$25K-$50K
Denise	50	F	Psych. Tech.	$50K-$75K
Ed	66	M	Retired	$25K-$50K
Fiona	46	F	Education	$75K-$100K
Gina	80	F	Retired	$75K-$100K
Hailey	67	F	Retired	$25K-$50K
Ingrid	65	F	Retired	$25K-$50K
John	62	M	Clergy	$100K-$200K
Katrina	72	F	Retired	$25K-$50K
Laina	45	F	Admin.	$25K-$50K
Monica	42	F	Medical	$25K-$50K
Nancy	61	F	Medical	$50K-$75K
Oscar	70	M	Retired	Declined to respond

Table 1: Self-reported demographics of interviewees.

5. FINDINGS

The primary emergent theme from the interviews was that users had differing degrees of computer security *engagement*: a desire to control and manage their computer's functionality and security.[3] Interviewees' security engagement was distinct from their level of technical expertise. Some users with relatively little technical or security-related knowledge still expressed a desire to actively engage in computer security behaviors, while some relatively technically-knowledgeable users seemed to be largely disengaged. Furthermore, when participants' *perceived* levels of computer expertise were misaligned with their actual levels of expertise, their computers were likely to exhibit poorer security states. We also highlight additional themes expressed by our interviewees, including issues related to name recognition, trust, and legitimacy; update behavior; problematic gaps in users' knowledge; and an over-reliance on security software.

Table 4 in Appendix C lists the high-level codes in the final codebook. Our codes ultimately focused on traits, expressed beliefs, and self-reported decision-making related to user engagement. During the iterative coding process, the two coders grouped the high-level codes in the final codebook into *engaged* and *disengaged* categories. Interviewees were split into *engaged* and *disengaged* categories based on which code group was more common during their interviews. All interviewees clearly belonged in one of the two categories. When relevant, we use qualifiers such as "highly engaged" or "moderately disengaged" to highlight an interviewee's degree of (dis)engagement. Table 2 lists which interviewees were engaged versus disengaged, as well as other findings discussed in Section 5.2.

5.1 Security Engagement

We found that some users reported *disengaged* attitudes and behaviors regarding computer security. These users were likely to respond passively to events on their computers, either by ignoring them entirely or by requesting outside assistance for all but their most habitual tasks. They generally avoided making choices or independently seeking out information about their computers' functionality. They tended to make (often incorrect and dangerous) assumptions about their computers' default states. Their assumption that their computers would "just work" led to dangerous behaviors (e.g., accepting most or all prompts indiscriminately, assuming all security updates installed automatically).

In contrast, other users were relatively *engaged*. They seem to desire control and choice in computer security and maintenance tasks. They independently sought information on which to base their computer- and security-related decisions. However, more engaged users were not necessarily more knowledgeable. Some users who seemed fairly knowledgeable displayed disengaged behaviors, while some engaged users showed severe gaps in expertise.

Disengaged and engaged users alike desired to prevent security and functionality problems, but they differed in how they addressed these problems. Disengaged users did nothing or relied on automated features or outside help, while engaged users sought information and attempted to control both functionality and security.

5.1.1 Disengaged: "I just don't do anything."

Disengaged participants exhibited several similar behaviors and attitudes. Seven interviewees were classified as primarily disengaged: Betty, Fiona, Gina, Hailey, Laina, Nancy, and Katrina. Hailey and Nancy seemed to be especially disengaged, with no segments from their interviews corresponding to the "engaged" code group at all.

Outsourcing maintenance and security tasks. First, many of these users outsourced computer maintenance to a *resident expert*: a person (typically a family member) to whom the user entrusted the responsibility of performing computer security and maintenance tasks. When asked about how her computer was maintained, Hailey said, "It's my daughter who always fixes all my mistakes, I don't know." Hailey indicated that her daughter performs a variety of maintenance tasks for her, including organizing files, deleting unwanted e-mails, and offering remote troubleshooting: "she's installed [a firewall]. And I don't know if there's anything else other than the firewall. She checks it to make sure that I'm not being hacked or something?" However, we did not find any third-party security software running on Hailey's computer during her participation in the SBO.

Unfortunately, in some cases, we found evidence that these resident experts' technical expertise was lacking, which put participants and their computers at risk. Betty's spouse maintains her computer (and its security). Betty and her spouse (who was offering additions to Betty's responses in the background during the phone interview) thought it had security software named "Fix-it," but no such software could be found on the machine during the interview's remote session. According to the SBO data, this machine did have Avanquest's Fix-It Utilities Professional[4] installed at one time, but it does not provide anti-virus protection and was uninstalled months before the interviews.

Several users in this group outsourced computer maintenance to paid services, whether via remote sessions or physically taking their machines to a computer store for either regular maintenance or to fix problems (e.g., too slow, annoying behavior, malfunctioning). Users who outsourced computer maintenance were often oblivious to what types of changes their "resident experts" or paid technicians made.

[3] We define engagement more broadly than some sources in the HCI literature [27]. A more deconstructed analysis of security engagement is left for future work.

[4] `http://www.avanquest.com/USA/software/www.avanquest.com/USA/software/fix-it-utilities15-professional-501513`

For example, when asked questions about how she maintained her computer, Katrina simply replied, "I'm not sure what that is, unless you're talking about [paid technicians] taking over my computer [with a remote session]."

When asked similar questions, Hailey said, "all [the technician] does is take over the computer like you do [with a remote session]."

Passive responses to problems. Left alone to use and manage their computers, disengaged users were more likely to avoid taking action than to try to investigate or resolve problems independently. Betty, Gina, and Hailey tended to avoid unfamiliar tasks and those that their resident experts or paid services had advised against, such as installing software.

In the case of problems or warnings, disengaged users stated that they would often cease their tasks entirely. When asked what she would do if she saw a web browser warning, Betty replied, "I should not click on it; I just don't do anything."

Some disengaged participants indicated that they would also contact their resident experts without attempting to independently resolve problems. When asked about her response to browser warnings, Hailey said, "I'd call my daughter... I'd close Google Chrome, I'd just close the computer."

When asked a question about her response to scareware-style pop-up messages, Laina indicated her response would be, "call my dad, tell him what I saw, and then he would tell me what to do," rather than independently performing any action, such as closing the web browser or navigating away from the web page.

Lack of technical awareness and interest. In some cases, disengaged users' awareness of their own knowledge limitations seemed to protect them from exploratory but risky behaviors. They reported a reluctance to download or install new software, visit unknown websites, or change default settings that may put their machines at risk. When asked about whether Hailey had ever disabled her anti-virus or firewall, she replied, "I would not know how to do that."

Some disengaged users also reported that they found computer maintenance unenjoyable. For example, Gina recalled when Binkiland adware needed to be removed, and stated, "[My husband] enjoys that garbage. I don't... My husband and the folks at McAfee sort of sorted through that."

It is important to note that disengaged users did not necessarily lack *motivation* to keep their computers secure. All of our users reported performing sensitive tasks (Section 4.4) and disengaged users reported being affected by and concerned about computer security problems. For example, Laina was a highly disengaged user, but ransomware seizing her personal files was catastrophic for her work-related tasks. While she desired to avoid such an outcome in the future, she still did not express any desire for additional personal control over her computer's security and instead continued to outsource all maintenance to a family member. This illustrates that users could be highly motivated to keep their computers secure while still having little interest in performing such management themselves.

5.1.2 Engaged: "I'm trying to be self-taught"

Eight interviewees (Agnes, Carl, Denise, Ed, Ingrid, John, Monica, and Oscar) seemed to be more engaged. These users were more wary of specific security risks and more likely to respond proactively to problems indicative of potential security breaches. Engaged users desired more granular control of their computers, displayed more complex approaches to maintaining the security and functionality of their computers, and exhibited more tendencies to troubleshoot problems and research topics independently.

However, these more engaged users did not seem to be substantially more knowledgeable or to make better decisions in all cases. In fact, their engagement sometimes caused them to make risky decisions in situations where the less-engaged groups might have been protected by inaction. For example, Agnes reported that she uninstalled her Norton security software about a year before the interview because she did not feel it was necessary, and she had not installed any other security software since. SBO data showed Norton was still present on Agnes's computer, but was not running. We suspect she simply chose not to renew a subscription without actually removing the software.

Proactive maintenance and responses to problems. Proactive maintenance to prevent problems and active responses to perceived problems were both hallmarks of engaged users. We specifically asked all interviewees whether they performed any regular maintenance tasks, and while disengaged users generally only performed maintenance in reaction to a problem that halted other tasks, engaged users sometimes had specific routines that they reported performing regularly to maintain their computers.

The routines described by engaged users seemed to reflect their intentions to proactively maintain their computers. However, some aspects of engaged users' routines indicated incomplete understandings of the computer's functionality. For example, every time Denise logs into her Windows machine, which she reportedly uses for approximately three hours every day, she will "perform virus checks" and "clean the internet files." Both of these are probably good habits, but she also mentioned that she defragments her hard drive with the same frequency, which is likely unnecessary and possibly even detrimental to the drive's functionality.

Engaged users also reported more active responses to past scenarios such as scareware messages or when asked what they would do in response to browser warnings (examples of which were displayed to users by the interviewer via remote session). Rather than "just doing nothing," engaged users often offered examples of ways in which they sought the source of the problem and/or tried to prevent it from recurring. However, being engaged did not imply that participants had an accurate technical understanding of the problem or how to resolve it. For example, Denise's default response to perceived security threats while browsing was to try deleting her browser history and cache because she believed that would keep malicious sites or pop-ups from "popping up again."

A common (and possibly somewhat more effective) default response to any perceived threat or problem was to "run a security scan" manually with whatever security software was present on the machine. However, this behavior was also taken too far as a default response in some cases. For

example, Oscar described having network connectivity problems (which, given his description, we believed were likely to be hardware or ISP problems), to which he reportedly conducted "a thorough manual scan." Two other users had also installed multiple conflicting security applications during past attempts to troubleshoot problems, likely making any existing performance problems worse and possibly hindering the programs' effectiveness as they compete with each other for access to the client machine's resources.

Information-seeking behaviors. Engaged users also tended to mention seeking out and reading product reviews and other kinds of publicly-available information about software and operating systems. In Oscar's words, "I'm trying to be self-taught." They seemed motivated to proactively seek information for a variety of reasons, including a desire for granular control, to preemptively avoid potentially problematic software, or simple curiosity. When making computer-related decisions (e.g., choosing software to purchase, whether to upgrade to Windows 10), engaged users commonly stated, "I Google it," and regularly read reviews from `CNET.com` or similar sources. The SBO data confirmed that at least four engaged participants (Carl, Denise, Ed, and Monica) and one less-engaged participant (Fiona) had searched online for information about their computers and their performance.

The tendency to perform independent research resulted in largely positive outcomes for engaged users. For example, it seemed to help users choose reputable software to install. Ed described how he chose Kaspersky as his security suite: "I checked out reviews, I read articles and PC magazines and CNET-type reviews to get an idea of what was the best security suite for the money, what offered the best protection for the lowest cost. What was the most reliable, what had the best customer service, things of that nature. And that's how I decided to go with the Kaspersky Security Suite." Carl also mentioned various kinds of research that he might perform to find information about software, including reading Internet forums.

In some cases, these investigations may have had negative impacts on users' attitudes and behaviors towards legitimate security products or upgrades. For example, Agnes said she avoids updates with negative reviews: "you'll hear people say 'don't install version 8.1.2 because... my computer slowed down immensely or my printer isn't functioning right,' so I usually [read reviews] before I install it." When participants discussed research performed before installing updates, they mentioned factors such as compatibility and performance, but not security.

Aware of and involved in updates. Engaged users were more actively involved with the update process overall, for better or worse. In some cases, this had positive effects: some engaged users mentioned actively and habitually checking for updates. On the other hand, some engaged users were more likely to "pick and choose" updates in strategic ways, and their strategies for doing so did not always seem to be well-informed. Many engaged users were at least aware that updates could be helpful in resolving problems with software in general, but not all were fully aware of the security purposes of some updates.

Unlike disengaged users, engaged users sometimes searched for updates without being prompted by their software. Some reported doing so as part of habitual, proactive maintenance. Monica, for example, said that she normally spent about half an hour performing a list of habitual maintenance tasks each time she logged onto the computer to "run my internet security, [do] my updates." Monica reported using the computer for five to six hours per day, three to four days per week.

Some would also look for updates manually to troubleshoot problems with specific programs. For example, Oscar described a situation in which a piece of software was not functioning as desired, and part of his response was to "check just to make sure that they didn't sneak a new version in that I didn't know about." Ed also mentioned troubleshooting his Kaspersky security software by searching Kaspersky's site and finding a download that resolved a conflict between Kaspersky and Windows 10.

However, engaged users' more active relationships with updates also resulted in sometimes explicitly choosing to avoid operating system and software updates that may fix critical security vulnerabilities. The reasons users cited for this behavior included prior negative experiences with updates or aversion to feature changes, confirming findings of past studies [33, 35, 36, 38].

Ed said that his behavior differs depending on whether the update seems to be critical or optional: "Sometimes I'll have something that, I don't know if they call it critical or what, and then there's recommended...or maybe it'll say 'recommended,' and it'll say 'in addition to,' and sometimes I'll ignore those, where it's an option of yes or no."

John said that he "has the update button set to contact me to let me know. I'm real careful about updating," citing past negative experiences with updates. This matched SBO data from his machine: Windows was set to notify him before downloading updates and multiple important updates had not been installed throughout his participation. John also noted, "What I tend to do is read the descriptions of the updates and pick and choose what seems to me to be of value." This is a distinct contrast from disengaged users' tendencies towards blanket approaches to updates and prompts: disengaged users tend to either ignore or avoid updates entirely or to accept prompts rather indiscriminately.

5.2 Computer Security State

We used the information available to us from the SBO data collection software to assess the states of interviewees' machines both in terms of their compliance with some of the most common points of standard end-user security advice (e.g., install updates regularly, run security software) and in terms of the presence or absence of undesirable software. These findings are summarized in Table 2.

5.2.1 Prevention: security software and updates

Three interviewees (Gina, Katrina, and Nancy) had machines that were relatively secure in their configurations, with security software running and updated versions of the vulnerable programs we examined. The remaining interviewees all had evidence of at least one of the following: a lack of third-party security software, outdated versions of vulnerable programs, or problematic Windows update behavior. Betty, Carl, and John possessed the machines with the most problems. Betty's machine lacked security software, was not installing Windows security updates regularly, and

		Security deficiencies						
		No	Updates OS		Out of date		Presence of	
	User	security software	Manually	Irregularly	Reader	Flash	Suspicious	Malicious
Disengaged	Betty	●		●	●		●	●
	Fiona				●			
	Gina						●	●
	Hailey	●				●	●	●
	Katrina						●	●
	Laina				●	●	●	●
	Nancy						●	●
Engaged	Agnes	●				●		
	Carl	●	●			●		
	Denise					●	●	●
	Ed					●	●	
	Ingrid					●	●	●
	John		●	●		●	●	
	Monica					●	●	
	Oscar	●						

Table 2: List of interviewees' machines' security deficiencies. ● denotes interviewee machines with *no security software*, *manual* or *irregular operating system (OS) updates*, *out of date* versions of Adobe *Reader* or *Flash*, or the presence of *suspicious* or *malicious* software.

was running an outdated and vulnerable version of Adobe Reader. Carl and John were not automatically installing Windows updates, which past work has shown can result in users installing updates more slowly and leaving vulnerabilities unpatched longer [38]. Carl was still manually installing operating system updates fairly regularly, but John had failed to install multiple important updates. Carl's machine also had no third-party security software.

In our sample, we observed a variety of combinations of levels of engagement and computer security states. Both engaged and disengaged users had machines that were generally configured according to common security advice such as installing updates and running antivirus software [20, 34].

Conversely, other engaged and disengaged users alike had very poorly-configured machines, including Carl, who was one of the most engaged, and Betty, who was especially disengaged and reliant on a "resident expert."

As one might expect, some disengaged users' computers were less secure. It seemed these users' lack of engagement resulted in a lack of awareness of (and/or interest in) their machine's security state. Betty and Hailey, for example, believed that their resident experts were maintaining security software on their computers, but we found that both of their machines lacked third-party security software and had malicious programs installed.

However, disengagement sometimes led to more secure states. For example, disengaged users seldom changed their Windows update settings from the default automatic installation (typically resulting in security updates being installed as soon as they are available). When asked whether she usually installed Windows updates, Fiona replied, "I don't know if it's a choice. I mean, I could make it a choice, I guess. But it doesn't. It just, automatically, it updates stuff."

On the other hand, since less-engaged users felt ill-equipped to make security decisions when their resident experts were unavailable to assist them, their inaction sometimes put their machines at risk. For example, they seemed less likely to install software updates, including those with security patches. Hailey mentioned several times that she sometimes delayed or refused updates for fear of making a mistake: "Sometimes Java sends me updates, and I don't really know what it is, so I don't download it 'cause I'm always afraid I'm gonna do something wrong." This type of response from disengaged participants also seemed to indicate that they sometimes went too far in taking advice to avoid installing unknown software: they sometimes seemed to conflate this with the installation of updates and as a result might not patch vulnerable software if they did not recognize it. In these cases, their intentions are to avoid security problems, but the effect is exactly the opposite.

Carl and John are examples of different security states between two engaged users. They were the only two interviewees who set their Windows update settings to notify them before installing updates so they could choose which to install. They cited previous bad experiences where updates were perceived to "change things" (undesirably) or "break things" (requiring troubleshooting). Despite their similar attitudes, the resulting states of their computers were quite different. The SBO data showed that Carl installs Windows updates very regularly, but John does not. John's interview responses confirmed that he is averse to updates that do not seem useful to him, even though he also understands that updates to software can sometimes be important for security. While he reported periodically installing software updates, it was unclear if he was aware that Windows operating system updates could also contain security updates.

5.2.2 *Evidence of outcomes: presence of suspicious and malicious software*

Both disengaged and engaged users exhibited good outcomes as measured by the lack of undesirable software found by the SBO's sensors (to the extent that we could detect it). Fiona and Oscar, for example, display very different approaches to security: Fiona is quite disengaged, while Oscar aims to be "self-taught" and is actively involved in configuration and

troubleshooting of his computer. Regardless, both seem to be successful, with no suspicious software detected on their machines.

Denise had relatively negative outcomes in terms of the unwanted software detected on her computer, despite being relatively engaged and having a computer with security software running regularly and software kept up-to-date (other than Flash Player). We detected three malicious and six suspicious programs on Denise's computer. Denise did not report awareness of the unwanted programs detected in the SBO data. However, she did have vague memories of having some sort of "Trojan" or "worm" in the past. She noted, "[her] icons were doing weird things, so I ran Norton," but she did not seem to remember how malicious programs had gotten installed, nor did she remember whether past problems were resolved fully or exactly why she chose particular courses of action, implying a lack of awareness of the actual state of her machine as a contributor to her problems.

Misdirected application of security advice may have also been a factor in Denise's case. When asked about hypothetical or actual past responses to situations such as scareware messages or browser warnings, Denise's preferred default response was to delete her temporary internet files and/or browser history. Denise may have learned that deleting cached files can solve certain kinds of problems or that removing the browser history might be beneficial for privacy, and she seemed believe this same solution might prevent more potential security problems than it actually does. Denise simply seemed to be trying to take any kind of action she could think of to address problems at the time. Accordingly, Denise may have installed undesirable programs like "BrowserSafeguard with RocketTab, Ad-Aware Security Toolbar," "RegCure Pro," and "Hardware Helper" while trying to troubleshoot security or performance problems. Her poor outcomes might have been mitigated if the operating system and software required fewer decisions from the user, or if she had been provided with more comprehensive advice about what actions to take in which situations.

In some cases, less-knowledgeable engaged users were sometimes more likely to take the wrong actions and put themselves at risk of security problems (for example, by picking and choosing types of updates that they deem unnecessary without understanding that those updates might contain security content). In contrast, sometimes the computers of certain disengaged users appeared to be more secure due to their users' inaction and deferral to defaults. Fiona, for example, describes an approach in which she generally clicks update prompts whether or not she fully understands their purpose. She also reports that she simply avoids installing new software altogether because she recognizes that she lacks the knowledge to know "what's safe and what's not safe." These factors may be contributors to the relatively clean state of her machine (mostly up-to-date software other than Adobe Reader and no detectable unwanted software). In this type of scenario, users may be protected by their recognition that the system might be more equipped to make security-related decisions and their reluctance to override system defaults.

On the other hand, sometimes disengaged users had poor outcomes, which frequently seemed to be due to over-reliance on their "resident experts" or professional help. This left disengaged users disempowered to resolve problems or make decisions independently. For example, Betty seemed to think that her husband was maintaining her computer, including keeping security software running, but this was not the case. Betty and her husband chose to seek additional paid assistance to resolve problems related to unwanted software on at least one occasion during the course of the study.

The worst observed outcome was on Laina's computer, which became infected with ransomware. Through an in-depth analysis of her SBO data, we identified this ransomware as "Ransom:Win32/Tescrypt.A," reported by Microsoft.[5] This type of ransomware has been frequently observed throughout 2015 and is most commonly spread through known vulnerabilities in out-of-date versions of Adobe Flash Player, Adobe Reader, and Java. In the few days before the ransomware seized her machine, Laina was both browsing the web and opening e-mail attachments with out-of-date versions of Adobe Flash Player and Adobe Reader. This disastrous outcome occurred in spite of her father, described as an IT expert, maintaining her computer. This illustrates that delegating computer security to a trusted third-party is not without considerable risk, suggesting that effective solutions tailored for disengaged users are essential.

In summary, disengaged users had machines in a variety of security states, since their lack of involvement or action had both positive and negative consequences. More engaged users also had machines in a variety of states, but for different reasons. Highly-engaged users might have been expected to have more secure machines because they were making more proactive efforts to manage their computer security (and were sometimes noticeably more knowledgeable). However, since these users were not experts, their efforts may have backfired at times when they made dangerous choices in configuring their machines. They took more action, but not always the correct action. They sought out and acquired more information, but sometimes that information was flawed or not reputable.

5.2.3 Discussion

A major insight revealed from our findings above is that users' levels of engagement in computer security tasks do ***not*** necessarily imply:

- how knowledgeable they are about correctly securing and maintaining their computers;
- how interested or motivated they are to keep their computers and data secure;
- the importance of the tasks performed on their computers (e.g., all users performed financial tasks, regardless of engagement); or
- how secured and/or compromised their computers will be.

One possible explanation for our observations here is that the state of a machine, both its configuration and theoretical risk and its actual health, is likely determined in some part by a *combination* of a user's level of technical expertise, her own ability to evaluate her expertise, and her subsequent engagement. On the one hand, we have noted users

[5] `http://www.microsoft.com/security/portal/threat/encyclopedia/entry.aspx?Name=Ransom:Win32/Tescrypt.A`

like Oscar, who demonstrated greater computer expertise and confidence in his technical ability than other interviewees. He was more engaged as a result of feeling that he was sufficiently knowledgeable to find information and make decisions himself. He also had fairly good outcomes: despite choosing not to install security software, the relatively malware-free state of his machine may be evidence that he was making correct security decisions.

There are also users like Fiona, who states that she does not have much technical expertise. She is an archetype of a disengaged user, whereby her approach is largely to "set it and then let it go." She mostly avoids installing software altogether: "I don't get a lot of new software, partly 'cause I don't know that I really need anything, but partly I don't know enough about computers to be a good judge of what's safe and what's not safe so I tend to just kinda shy away from doing much of anything." Since she is running an operating system that automatically updates by default (Windows 10), this approach seems to work well. Besides a lone outdated version of Adobe Reader, her self-assessment of her limited technical ability appears to have led her to a successful and relatively secure course of action.

In contrast, we have both engaged and disengaged users who have have unsuccessful outcomes. For example, John is highly engaged, but may place too much faith in his own ability to micromanage decisions about updates, since he does not install some important security updates. Thus, users at both ends of engagement spectrum can have positive security states and outcomes, if their levels of expertise and awareness of their (lack of) expertise are in alignment. We will suggest possible security solutions that respectively cater to engaged and disengaged users' needs and expectations in Section 6, inspired by some of the additional themes we identified in our data (Section 5.3).

5.3 Other Themes, Codes, and Findings

In addition to the concept of engagement with computer security and its varying relationships with users' computer security states and outcomes, we also identified some other themes below that warrant further mention since they impact users' participation in computer security.

5.3.1 Name recognition, trust, and legitimacy

Multiple participants reported that *legitimacy* was a major factor in their decisions to trust or not trust specific websites, software, or prompts. Participants generally defined legitimacy as a function either of the familiarity of a program or website's name or of subjective visual cues (e.g., the appearance of logos, the grammatical accuracy of a message).

A good example is Hailey, who will download and install updates from sources she recognizes and trusts, "...the Epson, I know that's my printer, so I, um, I download whatever they send me, and HP used to be my printer, but they still have some kind of thing on my computer, so I download that."

However, in some cases, interviewees did not recognize or trust legitimate software or brands, which can lead to poorer computer security. Hailey is again a good example: "Sometimes Java sends me updates, and I don't really know what it is, so I don't download it 'cause I'm always afraid I'm gonna do something wrong." As a result of not updating Java, her computer has unpatched Java vulnerabilities.

Oscar trusts his online news sources to not send him anything malicious, "If I'm on a site, like let's say [main local newspaper] or [another well-known local news source], and they're blocking something, I kinda trust that they wouldn't have something that's super bad." Unfortunately, Oscar seemed unaware that legitimate websites can still be a vector for malicious behavior, such as through malicious ads served by less reputable third parties (unbeknownst to the website owner) [24].

Participants had some difficulty clarifying specifically how they decide whether or not a digital event is from a trustworthy or legitimate source. For example, Agnes suggested she would only click on requests that either are related to her primary tasks or are from sources she recognizes (e.g., Adobe): "I'm just not gonna click on an e-mail and install somethin' that's gonna trash my computer, so I would say it has to be something legitimate. I can't say every time something comes up, 'if you wanna please click here to install,' I do it. It has to be related to what function I'm doing on the computer, and it has to just be legitimate. Usually it's Adobe, Adobe something..."

Similarly, Monica trusts messages that she recognizes from personal experience, and will override her computer's security settings if she feels that the request comes from a trusted source: "It all depends what it is. I've been using Adobe and Java for a long time, so I kinda know what's a good message and what's a bad message, so as long as it's something that's common to me, then I just ignore it, change my firewall settings, and let it run. A lot of times, the way my firewall's set up, it says maybe it didn't recognize the company, [but] I know the company. So I just let it slide through. Now if it was to say something like 'malicious' or 'malware,' I actually don't install it." It remained unclear how Monica would decide what to do if she was presented with conflicting information, whereby the request appeared to be from a legitimate source she trusted, but was flagged as malicious.

In addition to familiar names, users also mentioned relying on subjective cues to determine legitimacy. John, for example, noted that he paid careful attention to logos, "if the colors are right, to see if it's crisp and clear, because a lot of bogus stuff is copied, and every time you copy something you lose some fidelity." Unfortunately, these are typically very superficial cues that malicious sources can fake. Indeed, semantic attacks, such as phishing, rely on spoofing these types of untrustworthy trust indicators. Platforms and web browsers have attempted to combat this by creating trust indicators that cannot be modified by third parties, such as OS-level permission dialogs and web browser SSL indicators. However, previous research has found that many operating system and web security indicators fail because users do not notice them [40], cannot tell the difference between application-controlled content and immutable chrome [21], or view any professionally-designed logo as being trustworthy [25]. Our study is consistent with those findings in that none of our participants mentioned noticing the Windows User Account Control (UAC) prompts,[6] which they would have seen anytime that third-party software requested administrator privileges. Instead, they relied on ineffective cues that could have been spoofed.

[6] `http://windows.microsoft.com/en-us/windows/what-is-user-account-control`

5.3.2 Update behavior

Both *engaged* and *disengaged* interviewees mentioned avoiding and delaying updates. Some *engaged* users also discussed turning off automatic updates and manually picking and choosing updates to install.

These interviews contained a variety of examples that confirmed findings from past work on update behavior. Reasons for not installing updates may include, according to our interviewers' codes as well as previous work:

- Aversion to change (e.g., to UI changes) [35, 36]
- Inconvenience; interruption of tasks [38]
- Belief that that updates are not important, especially for software that is not used regularly [35, 36, 38]
- The "if it ain't broke, don't fix it" mentality [35]
- Past problems with updates (bugs, crashes, etc.) [33]
- Updates and upgrades with negative online reputations (e.g., from consumer reviews and forums) [33]
- Technical issues encountered during installation [36]

Some of our *disengaged* users also reported not installing updates for fear of making mistakes, and some, including Hailey and Laina, also mentioned relying on their resident experts to tell them when updates should be installed.

5.3.3 Problematic knowledge gaps

Basic concepts and terminology. Our interview questions avoided technical language whenever possible, but interviewees seemed unaware of some computing terminology. For example, when asked, "What web browser do you normally use?", three interviewees replied, "What's a web browser?" Furthermore, even those who were able to offer answers to the question sometimes answered by describing the appearance of the program's icon but were unable to give the name of the program.

One participant was also confused by our question regarding frequently-visited websites, asking, "What's a website?" Once offered examples, this participant did report using the Internet and visiting a few websites (e.g., Facebook, e-mail) primarily via AOL Desktop.

Terms such as "USB drive" or "flash drive" were confusing for some users. For example, one user was confused about whether her USB mouse would be considered a USB or flash drive. We advise that security interventions targeting end users be careful not to assume users are aware of what may be considered basic computing terminology.

Browser extensions. We asked participants about each of their installed browser extensions (including plug-ins and add-ons) to learn about users' decisions to install, uninstall, enable, or disable extensions. However, most participants were unaware and seemed unconcerned about their browser extensions. At best, a few users were vaguely aware of extensions' presence or purpose. We suspect this was partly a terminology issue as discussed previously. We also showed the participants their lists of extensions through the remote session, and multiple participants remarked that they did not know how to find such a list in their browser.

All participants had multiple browser extensions installed, but few could offer even vague information regarding experiences with extensions. Katrina installed an extension called Blur without fully understanding what it would do or the risks of providing her passwords to an extension: "[The Blur extension] just says it protects your passwords. It supposedly puts them in some type of an encryption, I think, [but] I didn't really see the value of it. I just kept getting prompts that I didn't want." She couldn't remember how she had gotten the extension, "I think it popped up. I was doing something with passwords. It said 'do you want to encrypt your passwords' or something like that...or maybe my email?" This illustrates that people in real-life situations will install software claiming to improve security (without verifying said claims) from unknown sources or e-mails, which is dangerous behavior that has been observed in previous experiments [7, 9]. We recommend that the capabilities of browser extensions, the risks of installing them, and methods of managing them be more clearly communicated to users.

5.3.4 Over-reliance on security software

When asked if she took any precautions when downloading files, Denise said "Norton checks all that out. It tells me if it's safe." This may have been an incorrect assumption, since we found that her Norton browser extension was disabled, preventing it from scanning downloaded files. She also recalled having a, "worm [or] I think it was a trojan. My icons were doing weird things, so I ran [Norton]." She did not know how her computer contracted the malware. Clearly Norton was insufficient to protect Denise from getting infected in the first place. This illustrates that, while using reputable security software is necessary, it alone is not sufficient. In fact, as Christin et al. previously observed [7], it is possible that the presence (or even the perception) of security software results in the Peltzman effect [28], whereby users engage in even riskier behaviors because they believe they are being protected.

6. DISCUSSION

Although our interviewees did not offer specific recommendations, our observations suggest that users with different degrees of engagement may benefit from distinct types and styles of interventions. Disengaged users, who want to minimize time spent on maintenance and security tasks, probably need concise, precise, simple, and easy-to-perform security instructions, as well as "fire-and-forget," "all-in-one" security solutions that, once applied, will remain effective without any user effort. Such solutions might also be effective for more engaged people, but they may want configurable settings to personally manage their systems and additional information supporting any suggested interventions. Still, engaged users are not computer security experts, so any information provided to them should use language that non-experts can understand, leveraging their existing understanding [3] and empowering them to make informed choices that avert dangerous errors [39].

The application of updates is a prime example of how this can be accomplished with varying degrees of success. Many of our users failed to install security updates for Adobe Reader and Flash Player, which are prone to security vulnerabilities. Modern software that updates automatically by default may overcome these problems in some cases in which users are not equipped to make good updating decisions. However, some of our users set Windows to prompt them before installing updates, because they did not want to risk updates changing how system features work (which

supports prior findings [35, 37]). Thus, we recommend that feature and interface modification updates be completely decoupled from security updates whenever possible. It may also be desirable for most security updates to be installed automatically, but we would not recommend automatic updating as a universal solution: our interviews and past studies [33,38] show that automatic updates can cause significant frustration and true functionality problems for users.

Disengaged interviewees reported that they would stop their primary tasks if their computers warned them of security problems. While doing so may sometimes be a safe course of action, it remains a severe usability problem, and it is not clear what users would do if time-critical tasks were halted while immediate assistance was unavailable. Thus, we recommend security warnings be designed to allow the user's task to proceed in a safe manner, rather than the typical all-or-nothing approach that forces users to proceed with risk, deal with the problem, or abort. Similarly, options presented to users should be framed with disengaged users in mind, offering concise recommendations that are more prominent than less-secure alternatives [12]. For example, when warning the user they may be accessing a dangerous website, secure alternative websites that may satisfy the user's primary goals should be suggested.

Past work [30] has shown that differences in technical training and knowledge may result in women being more at risk for falling for phishing attacks than men. While all four of our male interviewees fell into the engaged group, some women, such as Denise, were also highly engaged. Furthermore, Hailey said that her husband requests her help with the computer. Our sample is too small for us to draw conclusions regarding gender, so further research is warranted.

6.1 Limitations

Our analysis has some limitations. Given our small sample size, a distinction in engagement might not be as clear in a larger sample. However, given the marked distinction between groups within this exploratory study of a relatively small and homogeneous sample, we feel our main findings remain a valuable contribution worth further study.

Studies like ours may suffer from "observation effects," whereby subjects who know they are being observed alter their behavior. However, past work [17, 22] suggests that in-situ data collection does not affect users' natural behavior, and we believe SBO users are unlikely to significantly alter their computer usage since our software runs transparently in the background for months on their computers without affecting daily usage.

Our study is further limited by the fact that, when inviting participants for interviews, we ruled out some participants in the SBO panel who had previously been unresponsive to our communications. This may have biased our sample towards more extroverted participants or those with whom we had previous contact. While future work should attempt to reach out to all people in the target population, most user studies inherently have a similar selection bias whereby the data are collected from people who volunteered to participate.

7. CONCLUSION

In this paper, we explored the relationships between users' attitudes, behaviors, and understandings of computer security (collected from interviews) and the actual configurations and security outcomes observed on their computers (collected via the Security Behavior Observatory). Our interview analysis revealed that users vary in their degree of *engagement* in securing their machines. We then examined the relationship between each participant's level of security engagement and the actual state of their computer's security. Security experts might assume that greater user engagement in computer security would result in more secure machines and vice versa. However, our qualitative findings suggest that the relationship among users' security engagement and their computers' security states may be more complex. Engaged users desire more control and decision-making power, and thus have different needs from disengaged users who prefer delegating decisions to the machine or someone they trust. In addition to engagement, another important factor that may affect computer security is not only the user's own technical expertise, but also their *awareness* of their level of expertise. We found that, when an interviewee's estimation of their computer expertise was misaligned with their actual expertise, their computer's security was likely to suffer.

Our findings suggest a need for a more critical evaluation of the content, presentation, and functionality of security interventions we provide to users. Future research should also examine how to design security interventions tailored to users with differing levels of (perceived versus actual) technical expertise and computer security engagement, since they all have different information needs and expectations from computer security solutions.

This is the first of many studies leveraging the Security Behavior Observatory (SBO). The SBO provides a window into *in situ* computer usage, which can then be augmented with explanatory qualitative data from interviews and surveys. This provides multiple research communities (e.g., HCI, computer security and privacy, behavioral sciences) the opportunity to understand people's personal computing behaviors in the wild. As evidenced by the ransomware incident (Section 5.2.2), the SBO empowers researchers to observe critical events in real-time and reconstruct the sources and sequences of past events that led to incidents. The SBO's longitudinal data collection will provide more such critical insights in the years to come.

8. ACKNOWLEDGEMENTS

This work was partially funded by the NSA Science of Security Lablet at Carnegie Mellon University (contract #H9823014C0140); the National Science Foundation, Grant CNS-1012763 (Nudging Users Towards Privacy); and the Hewlett Foundation, through the Center for Long-Term Cybersecurity (CLTC) at the University of California, Berkeley. We also thank (Daisy) Xi Dai for her assistance with SBO data analysis, and the reviewers for their assistance in improving the paper.

9. REFERENCES

[1] C. L. Anderson and R. Agarwal. Practicing safe computing: a multimedia empirical examination of home computer user security behavioral intentions. *MIS Quarterly*, 34(3), 2010.

[2] H. Asghari, M. Ciere, and M.J.G. van Eeten. Post-mortem of a zombie: Conficker cleanup after six years. In *USENIX Security Symposium*, 2015.

[3] F. Asgharpour, D. Liu, and J. Camp. Mental models of computer security risks. In *Workshop on Usable Security (USEC)*. Springer, 2007.

[4] K. Aytes and T. Connolly. Computer security and risky computing practices: A rational choice perspective. *Journal of Organizational and End User Computing*, 16(3), 2004.

[5] P. Bryant, S. Furnell, and A. Phippen. Improving protection and security awareness amongst home users. In *Advances in Networks, Computing and Communications 4*. University of Plymouth, April 2008.

[6] J. Camp. *Trust, Reputation and Security: Theories and Practice*, chapter Designing for Trust. Springer-Verlang, 2003.

[7] N. Christin, S. Egelman, T. Vidas, and J. Grossklags. It's all about the benjamins: An empirical study on incentivizing users to ignore security advice. In *International Conference on Financial Cryptography and Data Security (FC)*. Springer, 2011.

[8] L. Cranor. A framework for reasoning about the human in the loop. In *Usability, Psychology, and Security (UPSEC)*. USENIX, 2008.

[9] R. Dhamija, J. Tygar, and M. Hearst. Why phishing works. In *Conference on Human Factors in Computing Systems (CHI)*. ACM, 2006.

[10] P. Dourish, R. Grinter, J. D. D. L. Flor, and M. Joseph. Security in the wild: User strategies for managing security as an everyday, practical problem. *Personal and Ubiquitous Computing*, 8(6), 2004.

[11] T. Dumitras, P. Narasimhan, and E. Tilevich. To upgrade or not to upgrade: Impact of online upgrades across multiple administrative domains. In *International Conference on Object Oriented Programming Systems Languages and Applications (OOPSLA)*. ACM, 2010.

[12] S. Egelman, L. Cranor, and J. Hong. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *Conference on Human Factors in Computing Systems*. ACM, 2008.

[13] M. Fagan, M. Khan, and R. Buck. A study of users' experiences and beliefs about software update messages. *Computers in Human Behavior*, 51, 2015.

[14] A. Forget, S. Komanduri, A. Acquisti, N. Christin, L.F. Cranor, and R. Telang. Security Behavior Observatory: Infrastructure for long-term monitoring of client machines. Technical Report CMU-CyLab-14-009, CyLab, Carnegie Mellon University, July 2014.

[15] S. Furnell, A. Jusoh, and D. Katsabas. The challenges of understanding and using security: A survey of end-users. *Computers & Security*, 25(1), February 2006.

[16] G. Grimes, M. Hough, E. Mazur, and M. Signorella. Older adults' knowledge of internet hazards. *Educational Gerontology*, (3), 2010.

[17] M. Harbach, E. von Zezschwitz, A. Fichtner, A. De Luca, and M. Smith. It's a hard lock life: A field study of smartphone (un)locking behavior and risk perception. In *Symposium on Usable Privacy and Security (SOUPS)*. USENIX, 2014.

[18] C. Herley. So long, and no thanks for the externalities: the rational rejection of security advice by users. In *New Security Paradigms Workshop (NSPW)*. ACM, 2009.

[19] A. Howe, I. Ray, M. Roberts, M. Urbanska, and Z. Byrne. The psychology of security for the home computer user. In *Symposium on Security and Privacy*. IEEE, 2012.

[20] I. Ion, R. Reeder, and S. Consolvo. "...no one can hack my mind": Comparing expert and non-expert security practices. In *Symposium on Usable Privacy and Security (SOUPS)*. USENIX, 2015.

[21] C. Jackson, D. Simon, D. Tan, and A. Barth. An evaluation of extended validation and picture-in-picture phishing attacks. In *Financial Cryptography and Data Security*. Springer, 2007.

[22] F. Lalonde Lévesque, J. Nsiempba, J. Fernandez, S. Chiasson, and A. Somayaji. A clinical study of risk factors related to malware infections. In *Conference on Computer and Communications Security (CCS)*. ACM, 2013.

[23] J. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1), March 1977.

[24] Z. Li, K. Zhang, Y. Xie, F. Yu, and X. Wang. Knowing your enemy: understanding and detecting malicious web advertising. In *Conference on Computer and Communications Security (CCS)*. ACM, 2012.

[25] T. Moores. Do consumers understand the role of privacy seals in e-commerce? *Communications of the ACM*, 48(3), 2005.

[26] National Cyber Security Alliance and Symantec. 2010 NCSA / Norton by Symantec Online Safety Study, October 2010. `http://und.edu/cio/it-security/awareness/_files/docs/2010-ncsa-home-user-study.pdf`.

[27] H. L. O'Brien and E. G. Toms. What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 59(6), 2008.

[28] S. Peltzman. The effects of automobile safety regulation. *Journal of Political Economy*, (4), August 1975.

[29] E. Rescorla. Security holes...who cares? In *USENIX Security Symposium*, 2003.

[30] S. Sheng, M. Holbrook, P. Kumaraguru, L. Cranor, and J. Downs. Who falls for phish?: A demographic analysis of phishing susceptibility and effectiveness of interventions. In *Conference on Human Factors in Computing Systems (CHI)*. ACM, 2010.

[31] A. Sotirakopoulos, K. Hawkey, and K. Beznosov. On the Challenges in Usable Security Lab Studies: Lessons Learned from Replicating a Study on SSL Warnings. In *Symposium on Usable Privacy and*

Security (SOUPS). ACM, 2011.

[32] E. Stobert and R. Biddle. The password life cycle: user behaviour in managing passwords. In *Symposium on Usable Privacy and Security (SOUPS)*. USENIX, 2014.

[33] Y. Tian, B. Liu, W. Dai, B. Ur, P. Tague, and L. Cranor. Supporting privacy-conscious app update decisions with user reviews. In *Conference on Computer and Communications Security (CCS) Workshop on Security and Privacy in Smartphones and Mobile Devices*. ACM, 2015.

[34] US-CERT. Security tip (ST15-003): Before you connect a new computer to the internet, 2015. `https://www.us-cert.gov/ncas/tips/ST15-003`.

[35] K. Vaniea, E. Rader, and R. Wash. Betrayed by updates: How negative experiences affect future security. In *Conference on Human Factors in Computing Systems (CHI)*. ACM, 2014.

[36] K. Vaniea and Y. Rashidi. Tales of software updates: The process of updating software. In *Conference on Human Factors in Computing Systems (CHI)*. ACM, 2016.

[37] R. Wash. Folk models of home computer security. In *Symposium on Usable Privacy and Security (SOUPS)*. ACM, 2010.

[38] R. Wash, E. Rader, K. Vaniea, and M. Rizor. Out of the loop: How automated software updates cause unintended security consequences. In *Symposium on Usable Privacy and Security (SOUPS)*. USENIX, 2014.

[39] A. Whitten and J. Tygar. Why Johnny can't encrypt: A usability evaluation of PGP 5.0. In *USENIX Security Symposium*, 1999.

[40] M. Wu, R. Miller, and S. Garfinkel. Do security toolbars actually prevent phishing attacks? In *Conference on Human Factors in Computing Systems (CHI)*. ACM, 2006.

APPENDIX

A. INTERVIEW TOPICS

Although this paper only discusses participant responses that were of most interest, our questions and discussion with interviewees focused on several broad topics related to computer usage, behavior, and security, including:

- Who uses the computer and for what purpose
- Computer accounts and use of authentication
- Software installation and updating practices
- File sharing practices
- Use of security software
- Involvement in previous security incidents
  - Experiences with scareware messages
  - Experiences with browser warnings
  - Experiences with adware or malware
  - Experiences with being "hacked," identity theft, or other compromise of sensitive information
- Web browser usage and use of extensions
- Use of wired and wireless networks

B. SELF-REPORTED COMPUTER USAGE

Self-reported computer usage is presented in Table 3.

Pseudonym	*Communication*	*Education*	*Entertainment*	*Financial*	*Productivity*	*Programming*	*Research*	*Social*
Agnes				✓	✓			
Betty	✓	✓		✓			✓	
Carl	✓		✓	✓	✓		✓	✓
Denise	✓		✓	✓	✓		✓	✓
Ed	✓	✓	✓	✓	✓		✓	✓
Fiona	✓		✓	✓	✓		✓	✓
Gina	✓		✓	✓	✓		✓	✓
Hailey	✓			✓	✓		✓	
Ingrid	✓	✓		✓	✓		✓	✓
John	✓		✓	✓	✓		✓	✓
Katrina	✓		✓	✓	✓			✓
Laina	✓		✓	✓	✓		✓	✓
Monica		✓	✓		✓		✓	✓
Nancy	✓	✓	✓	✓	✓		✓	✓
Oscar	✓			✓	✓		✓	✓

Table 3: Summary of self-reported computer usage (based on initial SBO demographic survey and interview responses) for *communication* (e.g., e-mail, chatting), *education*, *entertainment* (e.g., gaming, watching videos), *financial* (e.g., online banking, e-commerce), *productivity* (e.g., Office-type applications and tasks), *programming* (i.e., building software), *research*, and online *social* networking.

C. CODEBOOK

Table 4 describes our codebook.

Primary	Secondary	Tertiary
Engaged	Active response to problem	-
	Actively seeking updates	-
	Actively selecting updates	-
	Independently installing software	-
	Independently removing software	-
	Learning from experience	-
	Other	-
	Proactive maintenance	-
	Self-education	-
	Takes specific software precautions	-
Neutral	Neutral response to problem	-
	Updates cause problems	-
	Other maintenance	-
Disengaged	Accepts prompts indiscriminately	-
	Avoids updates or installations	Change averse
		Fear of making mistake
		Inconvenient or unimportant
	No maintenance	-
	No specific software precautions	-
	Other	-
	Outsourcing maintenance	Friends or family
		Professional
	Overly reliant on security software	-
	Passive response to problem	-
	Rarely or never installs software	-
	Reactive maintenance	-
	Reliance on outside advice	-

Table 4: Final reconciled high-level codebook (organized by spectrum of engagement).

An Inconvenient Trust: User Attitudes Toward Security and Usability Tradeoffs for Key-Directory Encryption Systems

Wei Bai, Doowon Kim, Moses Namara, Yichen Qian, Patrick Gage Kelley,* and Michelle L. Mazurek
University of Maryland, *University of New Mexico
{wbai, doowon, mnamara, yqian1, mmazurek}@umd.edu, *pgk@unm.edu

ABSTRACT

Many critical communications now take place digitally, but recent revelations demonstrate that these communications can often be intercepted. To achieve true message privacy, users need end-to-end message encryption, in which the communications service provider is not able to decrypt the content. Historically, end-to-end encryption has proven extremely difficult for people to use correctly, but recently tools like Apple's iMessage and Google's End-to-End have made it more broadly accessible by using key-directory services. These tools (and others like them) sacrifice some security properties for convenience, which alarms some security experts, but little is known about how average users evaluate these tradeoffs. In a 52-person interview study, we asked participants to complete encryption tasks using both a traditional key-exchange model and a key-directory-based registration model. We also described the security properties of each (varying the order of presentation) and asked participants for their opinions. We found that participants understood the two models well and made coherent assessments about when different tradeoffs might be appropriate. Our participants recognized that the less-convenient exchange model was more secure overall, but found the security of the registration model to be "good enough" for many everyday purposes.

1. INTRODUCTION

As important communications become primarily digital, privacy becomes an increasingly critical concern. Users of communication services (e.g., email and chat) risk breaches of confidentiality due to attacks on the service from outsiders or rogue employees, or even government subpoenas. The only way to truly assure confidentiality is to use encryption so that the communication service has no access to the content. Despite considerable evidence of and front-page reporting about content breaches [3, 5, 9, 21, 24], encryption has generally not been widely adopted for person-to-person communications such as email and chat [20].

Researchers have given considerable thought to the reasons for this lack of adoption. More than 15 years of research have identified major usability problems with encryption tools, ranging from poorly designed user interfaces to the fundamental challenges of safe and scalable key distribution [16, 33, 35, 43].

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22–24, 2016, Denver, Colorado.

Recently, however, progress toward better usability and thus wider adoption has been made. Apple applied seamless end-to-end encryption to its iMessage and FaceTime services [2,25]. By centrally distributing public keys, Apple ensures the encryption is transparent to users, bringing end-to-end encryption to millions of iPhone, iPad, and Mac users. This design, however, leaves open the possibility that Apple itself could carry out a man-in-the-middle attack to break its users' privacy, for example at the request of law enforcement authorities [8, 42]. Popular messaging app WhatsApp has also implemented end-to-end encryption for text, voice, and video communications [27]. As with iMessage, WhatsApp centrally distributes public keys; however, users can optionally verify each other's keys manually or via QR code. Google and Yahoo! are currently developing similar approaches, with an added monitoring protocol that allows users and third parties to audit the key directory for consistency and transparency [19, 30, 37]. Some privacy experts have suggested that given this potential man-in-the-middle attack, these services should not be recommended to end users. As just one example, one security researcher suggests that "iMessage remains perhaps the best usable covert communication channel available today if your adversary can't compromise Apple. ... If one desires confidentiality, I think the only role for iMessage is instructing someone how to use Signal[1]" [42].

In a sense, the issue comes down to whether the benefit from many more people adopting encrypted communications is outweighed by the reduced security inherent in the central key distribution model. While security experts are best positioned to understand the technical differences between models, end users will ultimately be faced with the choice of which platforms and products to install and use. Researchers have considered the needs of some highly privacy-sensitive users, such as journalists and activists [18, 28]. To our knowledge, however, no one has asked average users for their opinions about these tradeoffs. This means that although security researchers may understand the risks and benefits of different tools, as a community we do not understand how an average user will weight different factors in deciding whether to adopt or ignore various encrypted communication technologies.

To understand how non-expert users feel about these tradeoffs, we undertook a 52-person lab study. We introduced participants to two encryption models: an *exchange* model in which participants manually exchange keys (analogous to traditional PGP) and a *registration* model in which participants sign up with a central service that distributes keys (analogous to iMessage). For each model, we asked them to complete several encrypted communication tasks; we also gave them a short, high-level explanation of each model's

[1]An encryption tool: https://whispersystems.org/. Last accessed on 05/16/2016.

security properties. (We varied the order of presentation to account for biases.) We then asked participants to comment on the security and usability of each model, as well as their overall opinion of the tradeoffs involved. The experiment was designed, insofar as possible, to avoid comparisons based on user-interface design and focus instead on the underlying properties of each encryption model.

We found that participants understood the two models fairly well and expressed nuanced insights into the tradeoffs between them. As predicted, participants found the registration model considerably more convenient than the exchange model. More interestingly, while the exchange system was considered more secure overall, the difference was slight: both general trust that large email providers would not risk their reputations by cheating and reasonable concerns about participants' own ability to correctly implement the exchange model mitigated this difference. Separately, we asked about half of our participants to evaluate the auditing model proposed in CONIKS [29], which is similar to that in development by Google and Yahoo!, and we found that for many users it provides a meaningful additional degree of confidence in the registration model's privacy.

Overall, our results suggest that users recognize the benefit of the exchange model for very sensitive communications, but find the more-usable registration model sufficient for the majority of everyday communications they engage in. While there are risks to this model, some of which can be alleviated by auditing, we argue that the marginal benefit of broad adoption will outweigh these risks. Historically, encryption schemes that require significant user effort have never gained broad popularity. Trying to convince average users to exclusively use more complicated schemes, when they often don't see a need for the added protection, may instead keep them away from using any encryption at all. Rather than spreading undue alarm about the risks of registration models, or forcing users into only exchange models, we recommend that policymakers and designers present tradeoffs clearly and encourage adoption of usable but imperfect security for the many scenarios where it may be appropriate.

2. BACKGROUND AND RELATED WORK

We briefly discuss the history of public-key-encrypted email systems and encryption usability studies.

2.1 A brief history of encrypted email

Diffie and Hellman proposed public-key cryptography in 1976, suggesting that a public directory would allow anyone to send private messages to anyone else; in 1978, the RSA algorithm made the idea practical [11]. In 1991, John Zimmerman developed PGP, which supported sending public-key encrypted email. In the second version, to alleviate the key verification problem, he proposed a "web of confidence" (later known as web of trust) for establishing key authenticity [44]. In a web of trust, users can sign each others' keys to endorse their authenticity, and can choose to accept keys that come with signatures from "trusted introducers." Despite this, key verification has remained problematic for many years.

In 1999, RFC 2633 defined Secure/Multipurpose Internet Mail Extensions (S/MIME), which takes a centralized approach to key distribution: all users have public-key certificates signed by a certification authority (CA), which are distributed along with any signed emails sent by that user [31]. S/MIME allowed straightforward integration of encryption to email clients like Microsoft Outlook and Netscape Communicator and was adopted by some corporate organizations with the capability to manage keys hierarchically, but was not adopted broadly by consumers.

More recently, several researchers and companies have explored ways to split the difference between completely decentralized and completely centralized key management. Gutmann proposed applying *key continuity management*, in which keys are trusted on first use but key changes are detected, to email [22]. In Apple's iMessage, private keys are generated on users' devices and the corresponding public keys are uploaded to Apple's proprietary directory service. To send a message to a user with multiple devices, the message is encrypted once for each device [1, 25]. WhatsApp uses a related approach based on the Signal Protocol, but allows users to confirm the authenticity of each other's keys if they choose to [27]. A recently reported vulnerability in the iMessage encryption mechanism points to the importance of validating the security of any end-to-end-messaging system [17]; however, this is orthogonal to our consideration of the underlying key exchange model.

In *certificate transparency*, publicly auditable append-only logs can be used to determine whether rogue certificates have been signed using a stolen CA key [26]. Ryan extended this approach for end-to-end email encryption [34]. CONIKS extends certificate transparency to allow users to efficiently monitor their own key entries and to support privacy in the key directory [29]. Google and Yahoo! are adopting a variation of certificate transparency for their end-to-end encryption extension [19, 30]. Each of these approaches trades off a different amount of security for convenience.

Other researchers have considered alternatives to standard public-key encryption that are designed to be more usable. Fahl et al. proposed Confidentiality as a Service (CaaS) [13], which operates on a registration model mostly transparent to users. This approach uses symmetric cryptography and splits trust between the communications provider and the CaaS provider. Neither individually can read private messages, but if the two collude they can.

2.2 The usability of encrypted email

In 1999, Whitten and Tygar published the now-seminal *Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0* [43]. This paper evaluated the interface for PGP 5.0 and found that most users (two-thirds) were unable to successfully sign and encrypt an email in the 90 minute session. This led to a series of follow-on papers: evaluating PGP 9 (key certification is still a problem) [35], S/MIME and Outlook integration (KCM seems promising) [16], Facebook encryption (using CaaS) [14], and several others (e.g., [33,39]). These studies largely ask users to do tasks they are unfamiliar with and focus on success rates (key pairs generation and collection, sending and decrypting messages, etc.). They provide valuable insight into how effectively novices can learn a particular system, how specific user interface design choices impact users, and where the difficulties lie. However, users are rarely presented with multiple potential encryption infrastructure models. Ruoti et al. compared the usability of three email encryption systems using pairs of novice users [32], but this work did not consider the security tradeoffs of the systems users were evaluating.

Tong et al. re-evaluated the test of *Johnny* with a different set of terms and documentation, including using a lock-and-key metaphor for public and private keys [40]. In preliminary results, they found that the metaphors aided understanding. We adopt the lock metaphor in our study, as detailed below.

Researchers have also studied social and cultural norms that also lead to aversion to encryption. Often users believe that they have no reason to encrypt their email because they have "nothing to hide," or because they cannot imagine anyone being interested in the messages they are sending [36]. In an interview study at an unnamed

non-violent, direct-action organization (which one might expect to be more interested and aware of the benefits of encryption), Gaw et al. found that employees believed "routine use of encryption [was] paranoid [behavior]" [18]. In this work, we do not directly address social norms regarding encryption, but several participants did discuss paranoia and suggested using different systems to accommodate different levels of privacy concern.

McGregor et al. considered the specific security and encryption needs of journalists protecting confidential sources, finding that adoption is frequently driven by source preferences and that existing models do not meet some important journalistic needs, such as verifying the authenticity of sources [28]. Considering the needs and preferences of users with critical privacy sensitivity, such as activists and journalists, is an important topic, but is orthogonal to our emphasis on general users.

3. METHODOLOGY

We used a within-subjects lab study to examine participants' concerns and preferences regarding the usability and security of end-to-end email encryption. Each participant was introduced to two general models for key management, *exchange* and *registration*. For both models, we described a public key as a *public lock*. This approach, inspired by Tong et al., avoids overloading the term "key" and was used to provide a more intuitive understanding of how public-key pairs operate [40].

In the exchange model, similar to traditional PGP, participants generate a key pair and then distribute the public locks to people they want to communicate with. We offered participants several methods for exchanging locks: the same email account they would use for encrypted communication, a secondary email account, posting the public lock on Facebook or sending via Facebook Messages, or using a simulated "key server" to upload their lock to a public directory. (These options were presented to each participant in a random order.) Simulated correspondents (played during the study by a researcher) sent back their own public locks via the same mechanism the participant chose, or via the mechanism the participant requested.

In the registration model, participants again generate a key pair. In this case, they "register" their public lock with a simulated key directory service; correspondents' locks were pre-installed to simulate automatically retrieving them from the directory. Participants were thus able to send and receive encrypted email from all simulated correspondents immediately upon creating and registering their own keys. In iMessage, the key generation step itself is completely transparent to users, who may never realize a key was created; we chose instead to make key generation explicit to help users understand the process.

Within each model, participants were asked to complete a series of simulated tasks, such as exchanging encrypted emails in a role-playing scenario (see details below); they were also introduced to a brief, non-technical review of the security properties of each model. Participants were asked to give their opinions about each model immediately after completing the tasks and security learning for that model. We also conducted an exit interview regarding the overall usability and security of each model, whether participants would use it themselves or recommend it to others, and in what circumstances it might or might not be appropriate.

We chose a within-subjects study because we were primarily interested in how participants would understand and value the tradeoffs among the options. As shown in Table 1, we varied the order of activities to account for ordering effects. Participants were assigned round-robin to one of these four possible orders of activities.

First activity	Second	Third	Fourth
ET (Exchange, Tasks)	ES	RT	RS
ES (Exchange, Security learning)	ET	RS	RT
RT (Registration, Tasks)	RS	ET	ES
RS (Registration, Security learning)	RT	ES	ET

Table 1: The order of activities varied across participants. Each participant worked with either the Exchange (E) or the Registration (R) model first. Within each model, participants either completed the encryption Tasks (T) first or learned about Security properties (S) first. Throughout the paper, participants are labeled by first activity; e.g., participant RT3 completed encryption tasks for the registration model first.

3.1 Encryption tasks

The set of encryption-related tasks for each model is shown in Table 2. In both models, participants were asked to generate a key pair locally. In the exchange model, participants then exchanged public locks with simulated friend Alice, including both sending Alice their lock and importing the lock received in return. In the registration model, participants registered with a simulated central service and had their public lock automatically "uploaded" and others' locks automatically "imported." After the locks were exchanged or the participant registered, participants composed and sent an encrypted email to Alice. A researcher, posing as Alice, sent an encrypted response. As a slightly more complex task, participants were asked to send an encrypted email to a group of two recipients. This task was designed to get participants to consider how the two models scale. Finally, we asked participants to consider how they would handle several other situations, including communicating with larger groups of people and various possible errors related to losing or publicizing one's own private key or losing other users' public locks. The possible errors were specific to each model and are shown in Table 2. In the interest of simplicity, we did not include any email signing (or signature verification) tasks.

Encryption tasks were completed using a Gmail account created especially for the study and a Chrome browser extension based on Mailvelope.[2] We modified Mailvelope to remove its branding, change the labels to match our lock/key metaphor, and reduce the interface to include only those features relevant to the study tasks. Figure 1, right shows a screenshot of sending encrypted email with our extension. As in Mailvelope, users of our extension compose an email and then use an "Encrypt" button to select recipients. Upon receiving encrypted email, users are prompted to enter their password to decrypt it (with the option to save the password and avoid future prompting).

We created two versions of our extension, one for exchange and one for registration, taking care to make them as similar as possible. The only two visible differences were (1) changing the "Generate lock/key pair" menu item and subsequent screen (exchange model, Figure 1, left) to read "Register" (registration model) and (2) a lock import screen (Figure 1, center) that was only relevant in the exchange model.

We also provided participants with detailed instructions to help them use the Chrome extension. By simplifying the interface, keeping it consistent, and providing detailed instructions, we hoped participants' reactions would better reflect the inherent properties of

[2]https://www.mailvelope.com/. Last accessed on 05/16/2016.

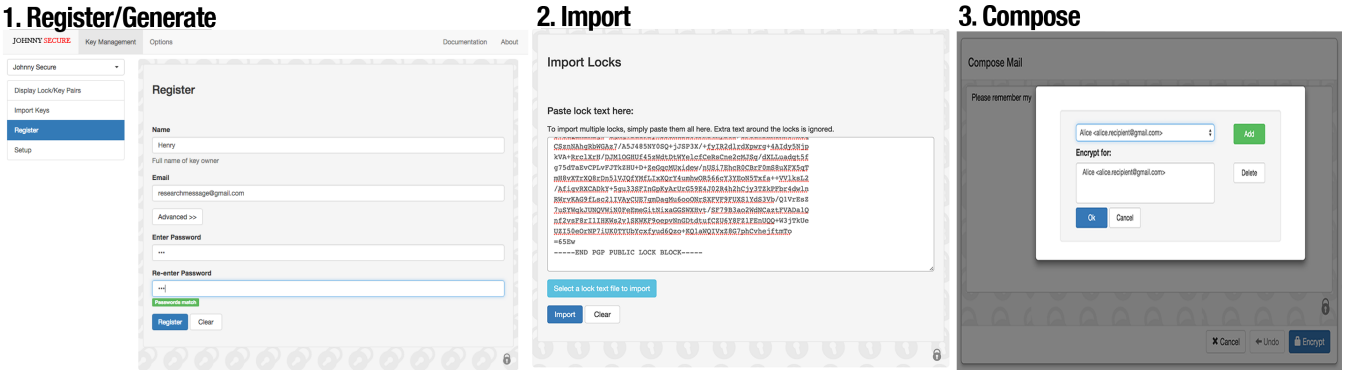


Figure 1: To use our extension, participants first generated (or registered) a key pair. Participants using the exchange model then needed to import recipients' locks. Finally, when composing encrypted emails, they clicked the Encrypt button (shown in the lower right of Step 3) to bring up a modal dialog to select recipients.

Task #	Exchange Model	Registration Model
1	Generate public lock/private key pair	Register public lock/private key pair
2	Exchange public locks with Alice	N/A
3	Send encrypted email to Alice	Send encrypted email to Alice
4	Decrypt received email from Alice	Decrypt received email from Alice
5	Exchange public locks with Bob and Carl	N/A
6	Send encrypted email to Bob and Carl	Send encrypted email to Bob and Carl
7	Decrypt received email from Bob and Carl	Decrypt received email from Bob and Carl
8	Imagine sending encrypted email to 10 people.	Imagine sending encrypted email to 10 people.
9	Consider misconfigurations: a. Lose Alice's public lock b. Lose own private key c. Publicize own private key	Consider misconfigurations: N/A b. Lose own private key c. Publicize own private key

Table 2: The encryption-related tasks completed by participants. The tasks differed slightly in the two models.

each model rather than idiosyncrasies of a particular interface.

3.2 Description of security properties

We provided participants with short, non-technical descriptions of possible attacks on each model.

Exchange model

For the exchange model, we described a man-in-the-middle attack in which the attacker could intercept or replace keys during the exchange process: "For example, when you try to get the public lock from Dave, the attacker secretly switches the public lock to his own. You think you have Dave's public lock, but in fact you have the attacker's. ... As a result, the attacker can read your email. The attacker will then use Dave's public lock and send the encrypted email to Dave, so that neither you nor Dave realize the email has been read." We also showed participants the illustration in Figure 2.

We decided not to include an option for key signing in our exchange model both because we thought it would add unnecessary complexity to our explanations and because it does not change the underlying requirement to trust some keys that are manually exchanged.

Registration model

For the registration model, we primarily described a man-in-the-middle attack enabled by the key directory service: "When you try to send encrypted emails to Dave, you think the database will return Dave's public lock to you. But in fact, it returns the attacker's lock, so the attacker can read your email. Therefore, you need to trust the email provider in this system." We showed participants the illustration in Figure 3.

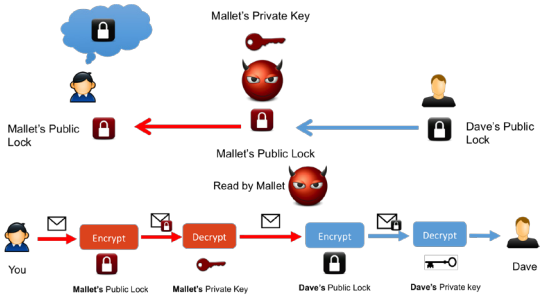


Figure 2: Possible attacks on the exchange model

In addition, we described two variations on the basic key directory approach: the Confidentiality as a Service (CaaS) variation [13,14], and an auditing model similar to the one proposed by Google and CONIKS [19, 29]. Because these approaches are not currently in wide use the way the iMessage-analogous system is, they were treated as secondary options. The auditing model was added (to the end of the interview, to maintain consistency with earlier interviews) during recruiting, and was therefore presented only to 24

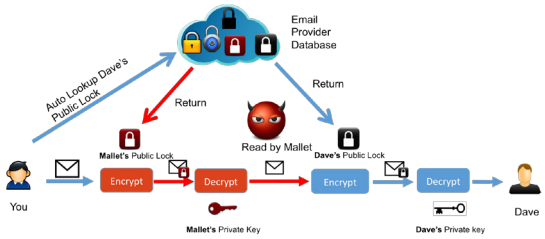


Figure 3: Possible attacks on the registration model

participants.

The security of the CaaS variation was described as follows: “There is a third-party service (not the email provider) as an intermediary. In this version, neither the third-party service nor your email provider can read your email themselves. However, if your email provider and the third-party service collaborate, they can both read your email. Therefore, you need to trust that the two services are not collaborating."

We described the auditing variation as follows: “The email provider stores all users’ public locks, just like [the primary registration model]. But there are other parties (auditors) who audit the email provider, to ensure it is giving out correct public locks. These auditors may include other email providers, public interest groups, and software on your devices. If the email provider gives you a public lock that doesn’t belong to the recipient, or gives someone else the wrong public lock for you, these auditors will notify you. You (or someone else) may use the wrong lock temporarily (for an hour or a day) before you are notified. In this model, you don’t need to trust your email provider, but you need to trust the auditors and/or the software on your device. Because there are several auditors, even if one auditor does not alert you another one probably will."

3.3 Participant feedback

Participants were asked questions after completing tasks for each model and at the end of the process. After completing tasks and learning about security for each model, participants were asked for their agreement (on a five-point Likert scale) with the following statements:

- The task was difficult (for each task).
- The task was cumbersome (for each task).
- The system effectively protected my privacy.

The first two questions were repeated for each task in Table 2. Before answering, participants were reminded that difficult tasks would require intellectual effort or skill, while cumbersome tasks would be tedious or time-consuming. After each Likert question, we asked participants to briefly explain their answer choice (free response).

After completing all tasks and learning about all security models, participants were asked several summative questions, including:

- Willingness to use each system, on a five-point Likert scale, and why.
- Willingness to recommend each system, on a five-point Likert scale, and why.
- What the participant liked and disliked about each system.

3.4 Recruitment

We recruited participants 18 or older who were familiar with Gmail and Chrome and who send and receive email at least 3 times per week. We placed flyers around our university campus and the surrounding area, advertised via email listservs for the university, and advertised on web platforms like Craigslist. All interviews were conducted in person at our university campus; interviews were video recorded with the explicit consent of participants. Participants were paid $20 for a one-hour study and were reimbursed for parking if utilized. Our study protocol was approved by the university’s Institutional Review Board.

Participants took part in the study in multiple batches between August 4, 2015 and Feb 5, 2016. For context, all of the participants engaged in the study well after Edward Snowden revealed details of the National Security Agency’s broad surveillance of digital communications [12], but before Apple publicly fought the Federal Bureau of Investigation to not weaken the security of a locked and encrypted iPhone [4]. We include this to note that average users’ views of security, privacy, and here specifically, encryption are a moving target. Future events may continue to shift public opinion on the importance of encrypted communications.

3.5 Data analysis

We used statistical analysis to investigate participants’ responses to the exchange and registration models. To account for our within-subjects design, we used the standard technique of including random effects to group together responses from each participant. We used a cumulative-link (logit) mixed regression model (CLMM), which fits ordinal dependent variables like the Likert scores we analyzed [23]. We included three covariates: whether the participant performed tasks or learned about security first, whether the encryption model she was evaluating was seen first or second, and the encryption model itself (exchange or registration). This approach allows us to disentangle the ordering effects from the main effects we are interested in. For each encryption model, we tested regression models with and without the obvious potential interaction of encryption type with order of exposure to that type, selecting the regression model with the lower Akaike information criterion (AIC) [6].

Qualitative data was independently coded by two researchers using textual microanalysis [38]. After several iterative rounds of developing a coding scheme, the researchers each independently coded the full set of participant responses, with multiple codes allowed per response. The researchers originally agreed on more than 94% of the codes, then discussed the instances of disagreement until consensus was reached. Where appropriate, we report prevalence for the final qualitative codes to provide context.

3.6 Limitations

Our methodology has several limitations. Our lab study participants had only limited exposure to the different encryption models, and their opinions might change after working with the models for a longer period. Participants also only imagined their responses to misconfigurations, rather than actually handling them. Nonetheless, we argue that first impressions like the ones we collected influence whether people will try any tool for long enough to develop more-informed opinions. It is well known that study participants may rate tools they examine more favorably (acquiescence bias) [41], which may explain the high rate of participants reporting they wanted to use or recommend each model. Because we are primarily interested in comparing results between models, we believe this has limited impact on our overall results; however, the absolute ratings should be interpreted as a ceiling at best.

In order to provide participants with any understanding of the security properties of each model, we had to prime them with descriptions of possible attacks. While this priming was unavoidable, we endeavored to keep the security descriptions as neutral as possible so that priming would affect both models approximately equally.

To avoid overwhelming participants, we evaluated a limited subset of possible encryption models and possible tasks; in particular, we left out key signing as well as any email signing or signature verification tasks. We did this because we believe signing to be the most difficult aspect of cryptography for non-experts to understand (see e.g., [43]), but including it might have provided a broader spectrum of user opinions.

Our registration model, unlike for example iMessage, was not completely invisible to participants. We believe it was necessary to give participants something to do other than just sending a normal email, in order to help them think through the tradeoffs involved. While presumably using a fully transparent variation would only have increased the convenience gap between the two models, prior work indicates that taking any steps at all increases feelings of security [33]. This may have contributed to the small observed security gap between the two models, but we argue that a version with no intervention required would lead to underestimations of security. Because we added the auditing model late, we were not able to get as much feedback about it or to compare it quantitatively to the other models we examined. In addition, because all participants encountered it last, their responses may reflect some ordering effects. Nonetheless, we believe the qualitative data we collected does provide interesting insights. Future work can examine all these alternatives in more detail.

As with many lab studies, our participants do not perfectly reflect the general population, which may limit the generalizability of our results.

4. PARTICIPANTS

A total of 96 people completed our pre-screening survey. We interviewed the first 55 who qualified and scheduled appointments. Three participants were excluded for failing to understand or respond coherently to any directions or questions.

Demographics for the 52 participants we consider are shown in Table 3. Among them, 60% were male and 80% are between the ages of 18-34, which is somewhat maler and younger than the general American population. Almost 85% of participants reported "primarily" growing up in the United States, South Asia, or East Asia. 40% of participants reported jobs or majors in computing, math, or engineering.

Despite this high rate of technical participants, most had little experience with computer security. We measured security expertise using a slightly adapted version of the scale developed by Camp et al. [7]. Higher scores indicate security expertise; the maximum score is 5.5 and the minimum score is zero. Only two of our participants scored 3 or higher.

Using a Kruskal-Wallis omnibus test, we found no significant differences among our four conditions in age, gender, country of origin, or security expertise ($p > 0.05$).

ID	Gend.	Age	Occupation	Security Expertise	Where grew up
ET1	F	25-34	Other	0	United States
ET2	F	45-54	Education	0.5	United States
ET3	M	21-24	Education	1.5	United States
ET4	M	25-34	Education	2	Middle East
ET5	M	21-24	Computers/math	1	South Asia
ET6	M	25-34	Engineering	2	East Asia
ET7	M	45-54	Life Sciences	2	United States
ET8*	M	18-21	Engineering	0.5	East Asia
ET9*	F	21-24	Computers/math	1	South Asia
ET10*	F	35-44	Computers/math	2	United States
ET11*	M	35-44	Transportation	0.5	United States
ET12*	M	21-24	HealthCare	1.5	United States
ET13*	M	21-24	Social Service	0.5	Western Europe
ES1	M	35-44	Engineering	0	United States
ES2	M	21-24	Sales	0.5	United States
ES3	F	25-34	Health Care	0.5	United States
ES4	M	21-24	Computers/math	4	South Asia
ES5	M	21-24	Computers/math	1	East Asia
ES6	M	25-34	Computers/math	1.5	South Asia
ES7	F	21-24	Education	0.5	United States
ES8*	M	25-34	Engineering	0.5	East Asia
ES9*	F	21-24	Engineering	1	South Asia
ES10*	M	25-34	Engineering	1	United States
ES11*	F	45-54	Business	0.5	United States
ES12*	F	21-24	Communications	0	United States
ES13*	F	25-34	Education	0.5	Latin America
RT1	M	25-34	Computers/math	3	East Asia
RT2	F	25-34	Sales	0.5	United States
RT3	M	21-24	Engineering	2.5	South Asia
RT4	F	21-24	Engineering	1.5	United States
RT5	M	21-24	Business	2	East Asia
RT6	F	25-34	Other	1.5	United States
RT7	F	25-34	Health Care	0	United States
RT8*	F	18-20	Sales	0.5	United States
RT9*	M	18-20	Education	0.5	United States
RT10*	M	25-34	Engineering	2	Middle East
RT11*	F	35-44	Admin. Support	0.5	United States
RT12*	M	35-44	Admin. Support	0.5	United States
RT13*	M	21-24	Production	0	United States
RS1	M	21-24	Other	1	East Asia
RS2	M	25-34	Life Sciences	1.5	Middle East
RS3	M	21-24	Computers/math	0	Africa
RS4	M	21-24	Computers/math	0.5	South Asia
RS5	M	25-34	Life Sciences	2	Middle East
RS6	M	25-34	Other	0.5	United States
RS7	F	25-34	Health Care	0	United States
RS8*	F	45-54	Sales	0	United States
RS9*	F	25-34	Engineering	1.5	East Asia
RS10*	M	21-24	Engineering	1	United States
RS11*	M	25-34	Architecture	0.5	United States
RS12*	F	25-34	Life Sciences	0.5	United States
RS13*	M	25-34	Construction	0	United States

Table 3: Participant Demographics. The columns show: participant identifiers (coded by activity order), gender, age, occupation, security expertise, and place where the participant grew up. The * indicates participants who were exposed to the auditing model.

5. RESULTS AND ANALYSIS

We present participants' reactions to the convenience and security of each model, followed by a discussion of their overall preferences among the models.

5.1 Registration is more convenient

Unsurprisingly, our participants found the registration system considerably more convenient, rating the exchange system as significantly more cumbersome and more difficult. Figure 4 and Tables 4 and 5 show the results of the CLMM for cumbersome and difficult, respectively, for Task 8: imagining sending email to a group of 10 people. In reading the CLMM tables, the exponent of the coefficient indicates how much more or less likely participants were to move up one step on the Likert scale of agreement.

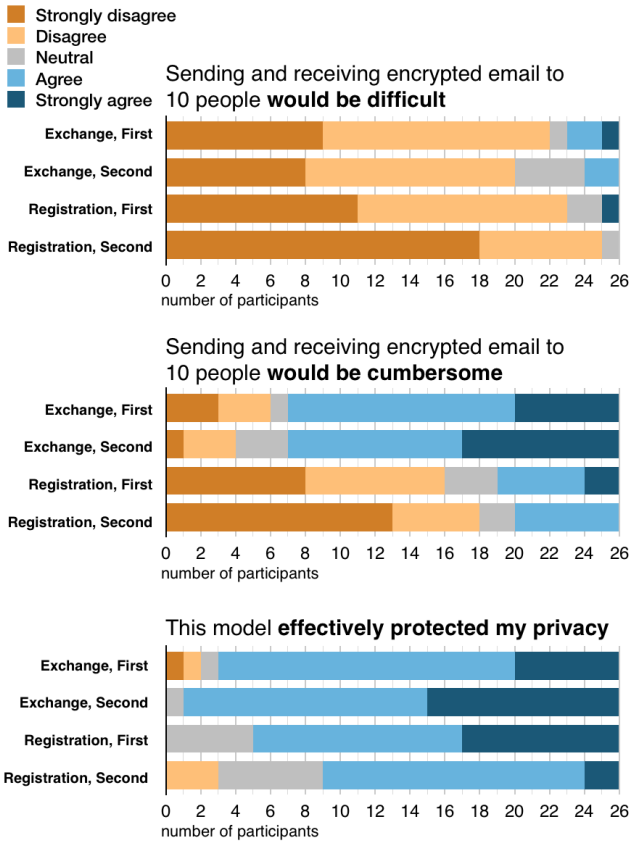


Figure 4: Participants' ratings of difficulty and cumbersomeness (Task 8) as well as whether participants thought the model protected their privacy. Labels indicate which model participants evaluated along with whether they saw that model first or second; e.g., "Exchange, First" indicates ratings for the exchange model among those who saw it first, which includes ET and ES participants.

For cumbersomeness the exchange model was associated with almost a 20x increase in likelihood of indicating more agreement. The exchange model was also about 5x more likely to be perceived as more difficult.

Factor	Coef.	Exp(coef)	SE	p-value
tasks first	0.010	1.010	0.589	0.986
second model	-0.166	0.847	0.393	0.673
exchange	2.978	19.656	0.567	<0.001*

Table 4: Regression table for cumbersomeness, Task 8. The non-interaction model was selected. Non-significant values are greyed out; significant values are indicated with an asterisk.

Factor	Coef.	Exp(coef)	SE	p-value
tasks first	-0.282	0.754	0.747	0.706
second model	-0.726	0.484	0.460	0.115
exchange	1.674	5.333	0.520	0.001*

Table 5: Regression table for difficulty, Task 8. The non-interaction model was selected. Non-significant values are greyed out; significant values are indicated with an asterisk.

Participants' comments generally supported this finding: that the exchange model was dramatically more cumbersome and somewhat more difficult. Within the exchange model, the most tedious task was manually exchanging locks and the most commonly mentioned reason was waiting for a correspondent's public lock. ES9 was concerned that the exchange model was "time-consuming, especially sending urgent emails. I have no choice but to wait for" the correspondent's public lock. RS5 agreed, saying "There are so many steps to exchange locks." RS13 mentioned that the cumbersomeness of exchanging locks was mainly related to initialization: "If their locks are already there, it would not be cumbersome. But if I have to ask them to send me locks person by person, it's more cumbersome." One participant (ET10) worried it would be additionally cumbersome to use the exchange model on a phone.

Several participants expressed concern that users with low digital literacy might have trouble with the exchange model or prefer the registration model. For example, RS12 recommended the registration model "especially to people that don't know very well how to use a computer . . . old people, like my father." ET2 agreed that the registration model is "easy to teach others to use."

While few participants considered any of the tasks very difficult, choosing a mechanism for exchanging locks was considered the most difficult step by a few participants, such as RS4, who mentioned having to "think about a safe way to exchange public locks," and RS10, who was concerned about making an exchange error while multitasking.

Other concerns related to the general issue of convenience included scalability and misconfiguration. As RT9 said, "When I send to more people, I have to be very careful, especially when I choose to send them my public locks separately. I need to control every step is correct." ET13 said, "When I exchange locks with ten people, I can send my lock, which is kind of easy. But I have to get ten replies for their locks. I can easily get lost. And if I exchange with 100 people, it'll be a nightmare." A few participants were concerned about the difficulty of recovering from misconfiguration, and ET10 was particularly worried that others' mistakes could cause additional hassle for her: "If other people lose their private keys and send me new public locks, I will be overwhelmed." RS12 agreed that "if accidents or mistakes happen, it bothers both parties to do extra steps."

The inconvenience of the exchange model could potentially be mitigated somewhat by posting the key publicly or semi-publicly (on a key server or Facebook profile), rather than sending it individually to different recipients. About a third of our participants chose this option: 34 used the primary email, 20 used the secondary email, 10 used Facebook chat, five posted to the Facebook profile, and 13 used the key server. (Some participants chose multiple methods during different tasks.) However, few of the participants who used the public or semi-public methods mentioned the added convenience as a reason for their choice. RT12 said exchanging locks is "not too cumbersome, it's manageable through the lock server to exchange locks". On the other hand, a few participants chose the key server because they thought it was more secure than other choices we provided.

5.2 The perceived security gap is small

We found that participants understood and thought critically about the security properties we explained to them for each model. Surprisingly, they found the exchange model to be only marginally more secure than the registration model, for a variety of reasons.

Exchange: Manual effort may lead to vulnerability

Most participants believed the exchange model was most secure overall, with 48 (out of 52, 92.3%) agreeing or strongly agreeing that this model protected their privacy. Nonetheless, participants also expressed concern that managing key exchange themselves would create vulnerabilities. More than half (27 out of 52) of participants were concerned about the security of the medium used to exchange locks.—ET4 worried that 'the key server [could] be manipulated or compromised." RT7 had several such concerns, including that an attacker could break into her Facebook account to post an incorrect public lock, or that public Wi-Fi in a coffee shop could be unsafe for transmitting locks. Overall, she said, "There are too many exchanges between different people. Exchanging [locks] to many people may go wrong." Others, like RS5, worried that their internet service provider could "sit between my recipient and me" and switch locks to execute a man-in-the-middle attack. ET7 was one of several participants who noted that "If I send the public locks and encrypted emails using the same email provider, it's not very secure." RT9 thought the ability to choose from different mechanisms to exchange locks provided added security, but worried that "people may choose a particular way in real life. It's their habits, so that attackers may anticipate" their choices and take advantage of their known routine. ES10 asked his recipients to send back his public lock, both through Facebook and via email, so he could verify for himself that the received public locks were not altered.

Other participants were concerned about making a mistake during the ongoing responsibility of managing keys. As ET10 put it, "Every time when I send or get a public lock ... there is a probability, even though not high, that my privacy is compromised. Then when I exchange public locks with many people, this probability will increase exponentially." RS12 worried that "I don't know what I actually need to do when I lose or publicize my private key. I am not confident about my answers. Non-tech experts may make mistakes."

Other participants mentioned that careless or compromised users could ruin the security of a whole group. ES12 said, "If I send to Alice, and she decrypts and goes away, then other people can see the email or even copy that email." ET8 said that "Within a company, if one person is hacked, then the whole company is hacked. It's hard to track the source, just like rotten food in the refrigerator." ET4 agreed that "There can be attacks on users with weak security, which may impair the whole user system."

Registration: Some concern but generally trusted

As expected, many participants were concerned about the need to trust email providers in the registration model. As ES5 said, having the email provider store "all public locks ... is not very comfortable." Despite this, however, most participants (38 out of 52) trusted the system protecting their privacy. Also, the CLMM results in Table 6 and Figure 4 indicate that the order in which the models were introduced played a significant role. Participants who saw the registration model first were more comfortable with it: 9 of 26 who saw registration first strongly agreed that the model protected their privacy, compared to only 3 of 26 who had already heard about the more-secure exchange model. None of the participants who saw registration first disagreed that the model protected their privacy, while 3 did so after seeing the exchange model first.

This general confidence in the registration model reflects many participants' belief that even though email providers could compromise the security of the primary registration model, they would be unlikely to. Ten participants mentioned that they trust their own email provider (presumably if they didn't they would switch services). ET11 mentioned that his email provider "knows me, I have my name there," and ET12 said that "All public locks are stored in a database, and I trust the database. This database provides extra security." ET13 provided a slightly different view: that some email providers are untrustworthy in well-known ways. "Everyone knows the Gmail potential vulnerabilities. And some people who are particularly hiding some information from the U.S. government, they will choose Yandex email from Russia, because they'd rather be intercepted by the Russian government, instead of the U.S. government. ... If you are an activist in US, and you don't want the U.S. government to know what you are up to, so I will choose some email services I feel comfortable with."

Several (7 participants) were specific about which kind of providers they would trust: RT8 would trust "certain big companies, not small companies," because big companies must protect their reputations. RT10 felt similarly, with an important caveat, mentioning that big companies like "Google and Yahoo! don't do such things [violate users' privacy] usually, unless the government forces them to do so. In general, it's secure." ET11 would choose an email provider with many users since "the more people using it, the more reliable." RT2, on the other hand, preferred to trust institutions like universities that "own their own email server" to better protect her privacy.

Also contributing to the general comfort level with the registration model is that participants do not believe most or any of their communication requires high security. RT4 said "encryption is not necessary for me," and RS8 agreed, saying "If I have some private information, I won't put it on the Internet."

CaaS and auditing: Some additional perceived security for registration

Twenty-two participants preferred the CaaS variation to the primary registration model, and 12 preferred the primary model to CaaS; the rest rated the two variations the same. The most popular explanation for preferring CaaS was a belief that different companies would not collude. RS7 said that the two parties would not collude because they do not "even trust each other." ES12 was cautiously positive, saying "This separation makes me feel good. However, [the two parties] still can possibly collaborate." Relatedly, ES8 suggested that the CaaS approach was more secure because "If one party is screwed up, you have another one to protect [your email]. You are still safe." These comments have implications for the auditing model as well; belief that different parties are unlikely to collude and recognition that distributing trust spreads out possible points of failure would also point to more trust in the auditing model.

On the other hand, four users thought the primary registration model was more secure than the CaaS variation because adding more intermediate systems and providers reduces overall security. RS1,

Factor	Coef.	Exp(coef)	SE	p-value
tasks first	-0.332	0.717	0.527	0.528
second model	-1.684	0.186	0.699	0.016*
exchange	-0.288	0.750	0.670	0.668
second model :: exchange	2.818	16.740	1.124	0.012*

Table 6: Regression table for privacy. The interaction model was selected. Non-significant values are greyed out; significant values are indicated with an asterisk.

for example, said that "involving more systems may complicate the system, so it is less trustful." A few users said that the possibility of collaboration invalidates the entire model: for example, ES13 said "I don't trust that much the whole system. I am afraid they may collaborate." Two participants (ET4, RS13) were afraid that in CaaS the two parties might collaborate for a sufficiently large gain. For example, RS13 said "They will not collaborate for one or two person's email, but for many, a group of people."

Other participants were concerned about whether the third party in CaaS was trustworthy. ET11 worried that "the third party service is not verified," and RT9 said his opinion "depends on who the two entities are. If the two companies are big names, like Gmail and Facebook, it seem more secure. Also if they do different types of services [from each other], it's more secure."

The 24 participants who were briefly exposed to the auditing variation gave generally positive feedback. ES9 was happy that "somebody is supervising" lock distribution and watching for problems, and ET13 said "Obviously it's extra secure. Other parties are verifying it, like an anti-virus system telling me if something goes wrong." ET8 appreciated that "if something goes wrong, I will be notified." The presence of many auditors reassured participants that collusion was unlikely; for example, RT10 commented that "it's less likely that all auditors [would] do something bad," and RS12 appreciated that "there are many auditors who can notify me."

Several participants, however, were concerned about the reliability of the auditors: RS9 said, "I want to know who these auditors are, ... their reputations, and whether they are truly independent." Similarly, RT13 said, "Am I able to choose auditors? This is a big question. The principle is good ... but I want to know who they are and how to choose them, because I need to trust them." One user (ET10) was concerned that auditors from competing companies might have incentives to lie about each others' behavior, making it hard to know who to trust. According to ET11, involving more parties reduced the overall trustworthiness: "Putting trust to only one party is better."

Ten participants expressed concern about the time lag for notification, noting that "a lot emails have already been sent" with even an hour's delay (ES10). RT11 said "It should be immediate notification. Even an hour is too late. ... Something bad has already happened." Others, however, were more pragmatic: "Immediate notification is ideal, but I don't expect immediateness in reality" (RT9). ET13 said the time lag "is a vulnerability. It depends on how often I send encrypted emails. If I use it very often, then it's vulnerable." Similarly, RT12 pointed out that "If I don't send the email, it doesn't matter, but in this case, I don't receive the wrong locks. ... Notification happens after the fact that I already received the wrong lock."

5.3 Overall comparison between systems

After exposing them to both models, we asked participants whether they would use or recommend the exchange model, the primary registration model, or the CaaS registration model. Figure 5 shows that the exchange model and CaaS variation were slightly preferred to the primary registration model. The number of participants who agreed or strongly agreed to use or recommend each model were 27, 23, and 28 (use) and 29, 21, and 28 (recommend). The CLMM results (Tables 7 and 8), which take the exchange model as a baseline, show no significant difference between exchange and either variation of registration for would-use, but do show that the primary registration was recommended less frequently than the exchange model. The 95% confidence intervals for each model indicate no

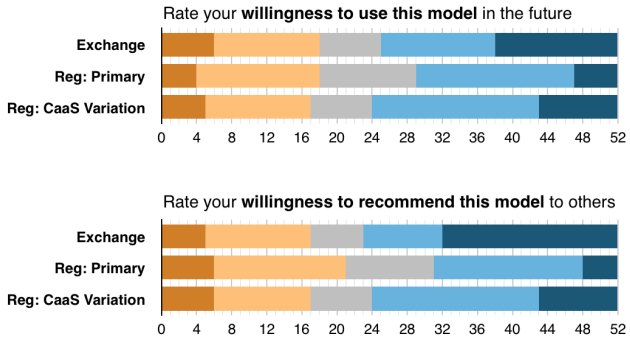


Figure 5: Participants' ratings of whether they would use or recommend each model.

significant differences between the primary and CaaS registration models in either case.

The regression models also indicate that participants who completed the encryption tasks before hearing about security properties were less likely to use or recommend any model than those who heard about security properties first. We hypothesize that participants who used the encryption extension before hearing about security anchored on the inconvenience of the tool rather than its privacy benefits. While this does not provide useful insight about comparing the different systems, it does underline the need for careful consideration about how new encryption tools are presented to the public.

Factor	Coef.	Exp(coef)	SE	*p*-value
tasks first	-0.606	0.546	0.308	0.049*
second model	-0.026	0. 975	0.291	0.930
registration (primary)	-0.376	0.687	0.358	0.294
registration (CaaS)	-0.077	0.926	0.360	0.823

Table 7: Regression table for whether participants would use each model. The non-interaction model was selected. Exchange is the base case for model type. Only whether participants completed tasks first or heard about security first was significant.

Factor	Coef.	Exp(coef)	SE	*p*-value
tasks first	-0.678	0.508	0.303	0.025*
second model	-0.198	0.820	0.291	0.496
registration (primary)	-0.915	0.401	0.368	0.013*
registration (CaaS)	-0.490	0.613	0.366	0.180

Table 8: Regression table for whether participants would recommend each model to others. The non-interaction model was selected. Exchange is the base case for model type. Primary registration is significant (less recommended vs. exchange), while CaaS is not significantly different from exchange. Participants who completed the encryption tasks before hearing about security properties were signficantly less likely to recommend any model.

We asked participants why they would or would not use each system, and categorized each participant's self-reported most important reason as related to security, usability, or both. (Details of participants' usability and security opinions for each system were discussed in Sections 5.1 and 5.2 respectively.) For participants who would not use a system, we also included having no need for

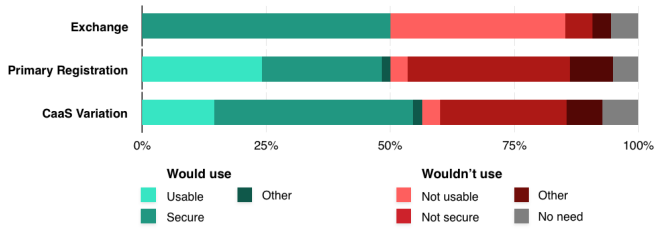


Figure 6: The most significant reason why participants would and would not use each model. We note here that while the number of participants who would use each system is similar, their reasoning varies. For example, prospective users of the exchange model uniformly cite security, while prospective users of the two registration models cite a mixture of security and usability.

encryption as a separate category. Figure 6 shows the results. Some participants gave more than one answer; a few did not give meaningful responses.

Unsurprisingly, the perception of better security attracted participants to the exchange model, while poor usability drove them away. Participants' reactions to the two registration models were more complicated. In both cases, insufficient security was the most common reason for rejecting the systems; however, participants who said they would use the primary registration model were evenly split between whether its usability or its security was more important. Participants who said they would use the CaaS model largely but not uniformly cited its security properties.

Participants who said they would use the exchange model generally described using it for high-security information only, or only at a small scale. ES6 exemplified this trend, saying the exchange model is "the safest one. I want to use it in small scale, like one or two people, ... like private and personal things. But I don't want to use it every day." RS9 felt similarly: "I think this system is more effective with fewer people, maybe under ten. I would use it when I send my credit card information to my Mom, instead of QQ or Wechat [two instant messaging services]." ES10 said he would use the exchange model for client projects, which should be kept secret until they are finished. Among the 27 participants who agreed they would want to use the exchange model, none mentioned using it with a large group; 16 said they would use it for very private information while only one said she would use it for general or everyday emails.

In contrast, participants who said they would use either variation of the registration model mentioned "contacting a large number of customers" for payroll or financial information (ET6) as well as "party and meeting announcements" (ET9, RS13). RT8 said she would use the registration model for information that was "overall private, but would not be a disaster if disclosed, e.g., my daughter is sick." ES7, a teacher, said she would use the exchange model only for "extremely sensitive information, such as SSNs," while she would use the registration model to send "location information or grade information." In total, 15 participants who wanted to use either variation of the registration model mentioned general email or large-scale communications.

These results suggest that although most participants said they would use both systems at least sometimes, quite a few wanted encryption only in specific circumstances. Between the exchange and registration models, however, our participants found the registration model useful in a broader variety of circumstances.

Using vs. recommending

As expected, most participants (44) who said they would use a system also said they would recommend it to others, and vice versa, but a few gave interesting reasons for answering differently. ET4 said he would not use the exchange model because it was too cumbersome, but would recommend it to others who have stronger privacy requirements. Similarly, RT4 said that "encryption is not necessary for me," but recommended the CaaS variation of the registration model because it is "easier to use [than the exchange model] and more secure than the vanilla [primary] registration system."

Registration vs. no encryption

We did not explicitly ask participants to compare these encryption models to unencrypted email. However, 5 participants who had concerns about the security of the registration model (total 14 rated less than 4) also mentioned that it does provide a valuable security improvement over unencrypted email. ET7 said "The email is not raw, which is another layer of security. ... Doing encryption gives me a security sense that I lock my door myself." RT12, explaining why he would use the primary registration model, noted that "I have to trust the email provider, which is problematic, but . . . it's better than raw email."

In line with findings from prior work [33], for some participants the process of taking any manual steps (such as generating a key pair in either model) increased their confidence that protection was occurring; for example, RS6 said "extra steps give me a security sense."

Auditing model

We asked participants who heard about the auditing model whether they would use it; overall, it proved popular. Of the 24 participants who were introduced to the auditing model, 15 said they would like to use it. Of these, 10 preferred it to any other model discussed. For example, ES11 said, "It's best among all systems mentioned in the experiment, because somebody else is policing them, just like watchdogs. If someone is reading your email, they might be caught." RT8 preferred the auditing model to any other option because "unlike the other models . . . instead of using [the attacker's] public lock blindly, I will get the update, 'Oh, that's the wrong public lock, you should not use this.'"

Four found the auditing model superior to the other registration models, but preferred the exchange model in at least some circumstances. RS10 said he would send personal information including banking data using the auditing model, but "if I worked in a government department, I would still use the exchange model." RT12 said the audit model is "slightly better than [the primary] registration model . . . because in [the primary registration] model I don't know if wrong locks happened. But overall, the lock exchange system has extra steps, extra layers of security, so I like it best among all the systems." Several of these 15 participants noted the possible time lag in notification as an important disadvantage, but were willing to use the model anyway. This generally positive reaction, combined with the preference to split risk among different parties in the CaaS model, suggests that the auditing model has strong potential to meet with user approval.

Eight participants said they would not use the auditing model (one was unsure). One of these (RS11) preferred it to all other models but believed he had no need to encrypt his email, and three found it worse than the exchange model. Four said it was worst among all models discussed, either because they did not trust the auditors or

because the time lag was too great.

5.4 Participant understanding

Despite receiving only a short introduction to each encryption system, most of our participants demonstrated thoughtful understanding of key concepts for each, suggesting that they provided credible opinions throughout the experiment.

Handling misconfiguration

We asked participants to consider how they would handle various possible misconfigurations in each model. Our primary goal was to prompt them to consider usability issues related to longer-term key maintenance, but this section of the interviews also offered a chance to evaluate participants' understanding of the different security models. Most participants were capable of reasoning correctly about these error scenarios.

Participants were presented with five different misconfiguration scenarios across the two models (see Table 2). Thirty-nine of 52 participants (75%) responded to all five scenarios with a straightforwardly correct answer, such as asking Alice to resend a lost public key (task 9a, exchange) or generating a new lock-key pair and redistributing the lock to all correspondents (task 9b, exchange). Seven additional participants (13.5%) provided such answers to at least three of the scenarios. One participant (RS13) mentioned recovering keys from a backup (such as a USB drive) rather than generating a new key pair.

We note several interesting misconceptions among those participants who got at least one scenario wrong. Four participants responding to task 9c (accidentally publicizing their own private key, in either model) suggested changing their password within the encryption extension; the password unlocks access to the private key, but a new password would not help if the key has already been exposed. Another participant (RS7) suggested for 9c that "I will send my email to a third person I trust, and ask that person to encrypt the email for me and send to my recipients. Similarly, he will decrypt the [response] email for me and forward it to me." This shows interesting security thinking but misses the potential for the message to be captured during the first step. Other common answers included getting tech support from the company that developed the encryption extension[3] and simply "I don't know."

Overall, participants were largely able to engage with these misconfiguration scenarios, demonstrating working understanding of the encryption tools; remaining misconceptions highlight areas in which more education, clearer directions in the tools, and more frequent use of encryption may be helpful.

Thinking about security

Our participants made several thoughtful points about encryption, security, and privacy that apply across models. ES4 mentioned that an extra benefit (of any encryption model) is a reduction in targeted ads: The "email provider can collect data through my emails, and then present ads. . . . I don't want that. [Using this tool] the ads will not appear."

ES10 expressed concern that an email encryption provider (in either model) might collect your private key, especially if you are using Apple email on an Apple device or Google email in Chrome, etc. One participant (RS9) was concerned about using public computers. This is potentially a problem for both encryption models, which assume the private key is securely stored on the user's local device. She was also concerned that the act of sending a lock might itself catch the interest of attackers; another participant (RS11) liked the sense of security provided by both encryption models but thought it might seem paranoid to worry about others reading his emails. Similar concerns were raised in [18]. ES12 expressed concern that the centralized nature of the registration model would provide a juicier target for an attacker than many individuals participating in the exchange model. ET10 worried that encryption would bypass an email provider's virus-detection system.

Several (11) participants liked that the exchange model allowed them to explicitly control who would be able to send them encrypted email. ES2 said he would "know the person whom I sent the public locks to," and RT3 liked that "who can send me encrypted emails [is] controlled by myself." RS13 said that "if I communicate with a group of people, it's easy to kick someone out of the group." A similar level of control can be implemented in a registration model; our findings suggest this is a feature at least some users value.

Although many participants understood and reasoned effectively about the security properties we presented, some retained incorrect mental models that have implications for the ongoing design of encryption systems. RS1 incorrectly believed that since he could not understand an encrypted message, no one else (including his email provider) would be able to either. Others were concerned about keeping their public locks secret in the exchange model; three split their locks across different channels in an effort to be more secure. For example, RS2 sent half of his public lock through the secondary email account and posted the other half on the key server. RT7 thought it would be insecure to store public locks: "After I send my lock to other people, others may not delete my public lock. . . . I may also forget to do so after I import others' locks. The fewer people know my public lock, the safer." Relatedly, ES13 worried that in the auditing model, the auditors "are scanning my lock. It sounds like more people are watching me besides the email provider, and I don't feel good."

Several participants also had concerns and misconceptions about how keys are managed across multiple devices, regardless of model. System designers may want to provide help or information on these points.

Evaluating tradeoffs

In deciding which system(s) they preferred, participants explicitly and deliberately made tradeoffs among their security and usability features. For example, ES13 said he would use the exchange model because "Exchanging locks makes it more private for me", despite the fact that "it takes time to exchange locks". ES10 also preferred the exchange model: "Having something better than baseline is one approach. But if I compare to perfect security I am trying to get, it's another approach. . . . When you want to use it, you really want it to be very well protected."

On the other hand, RT13, who said he would not use the exchange model, commented that "The negotiating process maybe gives me safer feelings, more protection. But on the other hand . . . the disadvantage is it is time consuming, cumbersome, tedious, more complicated, and this is the price I have to pay for more protection."

RS7 said she would use the primary registration model because it is "easy to use, and I think most of us trust our email provider", al-

[3] While completely reasonable in practice, this answer does not demonstrate understanding of the encryption model's security properties and so was not counted as "correct" for this purpose.

though she understood that "there are some possible threats." ET8, in contrast, would not use the primary registration model because "It's easy to send encrypted emails, especially to many people. But security concern is the reason I don't want to use it." According to ES12, the exchange model "is more straightforward. Only I and the other person [recipient] get involved in the communication, and no others." These comments and others demonstrate that participants understood pros and cons of the different models and thought carefully about how to balance them.

6. DISCUSSION AND CONCLUSION

We conducted the first study examining how non-expert users, briefly introduced to the topic, think about the privacy and convenience tradeoffs that are inherent in the choice of encryption models, rather than about user-interface design tradeoffs.

Our results suggest that users can understand at least some high-level security properties and can coherently trade these properties off against factors like convenience. We found that while participants recognized that the exchange model could provide better overall privacy, they also recognized its potential for self-defeating mistakes. Similarly, our participants acknowledged potential security problems in the registration model, but found it "good enough" for many everyday purposes, especially when offered the option to audit the system and/or split trust among several parties. This result is particularly encouraging for approaches like CONIKS and Google's end-to-end extension, which spread trust among many potential and actual auditors. It is important to note that understanding the identities and motivations of third-party auditors was important to several of our participants, so making this auditing process as open and transparent as possible may prove important to its success.

We believe our results have important implications for designers of encryption tools as well as researchers, policymakers, journalists, and security commentators. First, our results suggest that it may be reasonable to explain in clear language what the high-level risks of a given encryption approach are and trust users to make decisions accordingly. The Electronic Frontier Foundation's *Secure Messaging Scorecard*, which tracks the security properties of encryption tools, provides an excellent start in this direction [15]. Of course, participants in our study were directly instructed to read the materials we gave them; real users often have neither the time nor the motivation to seek out this kind of information. This magnifies the role of journalists, security commentators, and other opinion-makers whose recommendations users often rely on instead.

As a result, alarmed denunciations of tools that do not offer perfect privacy may only serve to scare users away from any encryption at all, given that many users already believe encryption is either too much work or unnecessary for their personal communications. Instead, making clear both the marginal benefit and the risk can support better decision making. This also underscores the critical importance of making risks explicit up front, in plain non-technical language; users who are misled into a false sense of security may misjudge tradeoffs to their detriment.

We do, however, advise some caution. Although most participants understood the encryption models and their security properties at a high level, there were some smaller misunderstandings that impacted their ability to make informed decisions. Despite years of effort from the security community, effectively communicating these subtleties remains difficult; however, we believe our findings demonstrate the benefits of continuing to try. Continued education, discussions in the media, and more frequent engagement with encryption tools in daily life may all assist this effort. Our own educational materials were improved through early pilot testing but not rigorously developed into an ideal or standard format; there is room to develop better materials for those users who are interested in learning more about encryption.

As end-to-end encryption is increasingly widely deployed, designers and companies must make choices about which models to adopt. We believe our results can provide some additional context for making these decisions, relative to the targeted use cases and user population. Further work in this area—for example, testing how a completely transparent registration model affects decision making and perception of security, examining an auditing model in greater detail and with reference to specific trusted auditors and notification lags, and comparing different approaches to framing security properties for non-experts—can provide further insight into how to optimize these choices.

7. ACKNOWLEDGMENTS

The authors wish to thank Elaine Shi and Christopher Soghoian for discussions that helped lead to this work; Nikos Kofinas and Yupeng Zhang for contributing to an early version of this study; members of the University of Maryland HCI Lab for their helpful feedback; and Cody Buntain for suggesting the title.

8. REFERENCES

[1] Apple. iOS security guide, iOS 9.0 or later. https://www.apple.com/business/docs/iOS_Security_Guide.pdf, Sept. 2015. (Last accessed on 05/16/2016).

[2] Apple. The most personal technology must also be the most private. https://www.apple.com/privacy/approach-to-privacy/, Mar. 2016. (Last accessed on 05/16/2016).

[3] J. Ball. GCHQ views data without a warrant, government admits. *The Guardian*, Oct. 2014. http://www.theguardian.com/uk-news/2014/oct/29/gchq-nsa-data-surveillance (Last accessed on 05/16/2016).

[4] D. Bisson. A timeline of the Apple-FBI iPhone controversy. *The State of Security*, Mar. 2016. http://www.tripwire.com/state-of-security/government/a-timeline-of-the-apple-fbi-iphone-controversy/ (Last accessed on 05/16/2016).

[5] O. Bowcott. Facebook case may force european firms to change data storage practices. *The Guardian*, Sept. 2015. http://www.theguardian.com/us-news/2015/sep/23/us-intelligence-services-surveillance-privacy (Last accessed on 05/16/2016).

[6] K. P. Burnham and D. R. Anderson. Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2):261–304, Nov. 2004.

[7] L. J. Camp, T. Kelley, and P. Rajivan. Instrument for measuring computing and security expertise. Technical Report TR715, Indiana University, Feb. 2015.

[8] C. Cattiaux and gg. iMessage privacy. http://blog.quarkslab.com/imessage-privacy.html, Oct. 2013. (Last accessed on 05/16/2016).

[9] C. A. Ciocchetti. The eavesdropping employer: A twenty-first century framework for employee monitoring. *American Business Law Journal*, 48(2):285–369, 2011.

[10] K. Conger. Google engineer says he'll push for default end-to-end encryption in Allo, May 2016.

http://techcrunch.com/2016/05/19/google-engineer-says-hell-push-for-default-end-to-end-encryption-in-allo/ (Last accessed on 06/02/2016).

[11] W. Diffie and M. E. Hellman. New directions in cryptography. *Information Theory, IEEE Transactions on*, 22(6):644–654, Nov 1976.

[12] K. Elliott and T. Rupar. Six months of revelations on NSA. *Washington Post*, June 2013. http://www.washingtonpost.com/wp-srv/special/national/nsa-timeline/m/ (Last accessed on 05/16/2016).

[13] S. Fahl, M. Harbach, T. Muders, and M. Smith. Confidentiality as a Service – usable security for the cloud. In *Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference on*, pages 153–162, June 2012.

[14] S. Fahl, M. Harbach, T. Muders, M. Smith, and U. Sander. Helping johnny 2.0 to encrypt his facebook conversations. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, SOUPS '12, pages 11:1–11:17. ACM, 2012.

[15] E. F. Foundation. Secure messaging scorecard, 2016. https://www.eff.org/secure-messaging-scorecard (Last accessed on 05/16/2016).

[16] S. L. Garfinkel and R. C. Miller. Johnny 2: a user test of key continuity management with S/MIME and Outlook Express. In *Proceedings of the 2005 Symposium on Usable Privacy and Security*, SOUPS '05, pages 13–24. ACM, 2005.

[17] C. Garman, M. Green, G. Kaptchuk, I. Miers, and M. Rushanan. Dancing on the lip of the volcano: Chosen ciphertext attacks on Apple iMessage, 2016.

[18] S. Gaw, E. W. Felten, and P. Fernandez-Kelly. Secrecy, flagging, and paranoia: Adoption criteria in encrypted email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 591–600, New York, NY, USA, 2006. ACM.

[19] Google. Google End-To-End wiki. https://github.com/google/end-to-end/wiki, Dec. 2014. (Last accessed on 05/16/2016).

[20] Google. Email encryption in transit. *Transparency report*, Mar. 2016. https://www.google.com/transparencyreport/saferemail/?hl=en (Last accessed on 05/16/2016).

[21] B. Greenwood. The legality of eavesdropping in the workplace. *Chron*, Dec. 2012. http://work.chron.com/legality-eavesdropping-workplace-15267.html.

[22] P. Gutmann. Why isn't the Internet secure yet, dammit. In *AusCERT Asia Pacific Information Technology Security Conference 2004; Computer Security: Are we there yet?* AusCERT Asia Pacific Information Technology Security, May 2004.

[23] D. Hedeker. Mixed models for longitudinal ordinal and nominal outcomes, 2012. http://www.uic.edu/classes/bstt/bstt513/OrdNomLS.pdf (Last accessed on 05/16/2016).

[24] C. Johnston. NSA accused of intercepting emails sent by mobile phone firm employees. *The Guardian*, Dec. 2014. http://www.theguardian.com/us-news/2014/dec/04/nsa-accused-intercepting-emails-mobile-phone-employees (Last accessed on 05/16/2016).

[25] G. Kumparak. Apple explains exactly how secure iMessage really is. *TechCrunch*, Feb. 2014. http://techcrunch.com/2014/02/27/apple-explains-exactly-how-secure-imessage-really-is/ (Last accessed on 05/16/2016).

[26] B. Laurie, A. Langley, and E. Kasper. Certificate transparency. RFC 6962, RFC Editor, June 2013.

[27] N. Lomas. WhatsApp completes end-to-ed encryption rollout. http://techcrunch.com/2016/04/05/whatsapp-completes-end-to-end-encryption-rollout/, Apr. 2016.

[28] S. E. McGregor, P. Charters, T. Holliday, and F. Roesner. Investigating the Computer Security Practices and Needs of Journalists. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 399–414. USENIX Association, 2015.

[29] M. S. Melara, A. Blankstein, J. Bonneau, E. W. Felten, and M. J. Freedman. CONIKS: Bringing key transparency to end users. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 383–398. USENIX Association, Aug. 2015.

[30] I. Paul. Yahoo Mail to support end-to-end PGP encryption by 2015. *PCWorld*, Aug. 2015. http://www.pcworld.com/article/2462852/yahoo-mail-to-support-end-to-end-pgp-encryption-by-2015.html (Last accessed on 05/16/2016).

[31] B. Ramsdell. S/MIME version 3 message specification. RFC 2633, RFC Editor, June 1999.

[32] S. Ruoti, J. Anderson, S. Heidbrink, M. O'Neill, E. Vaziripour, J. Wu, D. Zappala, and K. Seamons. "We're on the same page": A usability study of secure email using pairs of novice users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '16, pages 4298–4308, New York, NY, USA, 2016. ACM.

[33] S. Ruoti, N. Kim, B. Ben, T. van der Horst, and K. Seamons. Confused Johnny: when automatic encryption leads to confusion and mistakes. In *Proceedings of the Ninth Symposium on Usable Privacy and Securit*, SOUPS '13, pages 5:1–5:12. ACM, July 2013.

[34] M. D. Ryan. Enhanced certificate transparency and end-to-end encrypted mail. In *21st Annual Network and Distributed System Security Symposium, NDSS'14*, 2014.

[35] S. Sheng, L. Broderick, C. A. Koranda, and J. J. Hyland. Why Johnny still can't encrypt: evaluating the usability of email encryption software. In *Proceedings of the Second Symposium on Usable Privacy and Security*, SOUPS '06, 2006.

[36] D. J. Solove. 'I've got nothing to hide' and other misunderstandings of privacy. *San Diego Law Review*, 44:745, 2007.

[37] S. Somogyi. Making end-to-end encryption easier to use. *Google online security blog*, June 2014. http://googleonlinesecurity.blogspot.com/2014/06/making-end-to-end-encryption-easier-to.html (Last accessed on 05/16/2016).

[38] A. L. Strauss and J. M. Corbin. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage Publications, Inc, Thousand Oaks, CA, USA, 1998.

[39] M. Sweikata, G. Watson, C. Frank, C. Christensen, and Y. Hu. The usability of end user cryptographic products. In *2009 Information Security Curriculum Development Conference*, InfoSecCD '09, pages 55–59. ACM, 2009.

[40] W. Tong, S. Gold, S. Gichohi, M. Roman, and J. Frankle.

Why King George III can encrypt. http://randomwalker.info/teaching/spring-2014-privacy-technologies/king-george-iii-encrypt.pdf, 2014.

[41] M. Viswanathan. *Measurement Error and Research Design*. Sage Publications, 2005.

[42] N. Weaver. iPhones, the FBI, and Going Dark. *Lawfare*, Aug. 2015. https://www.lawfareblog.com/iphones-fbi-and-going-dark (Last accessed on 05/16/2016).

[43] A. Whitten and J. D. Tygar. Why Johnny can't encrypt: A usability evaluation of PGP 5.0. In *Proceedings of the 8th Conference on USENIX Security Symposium - Volume 8*, SSYM'99, pages 14–14, 1999.

[44] P. Zimmermann. PGP version 2.6.2 user's guide. ftp://ftp.pgpi.org/pub/pgp/2.x/doc/pgpdoc1.txt, Oct. 1994.

APPENDIX

This appendix contains the full survey and instructional instrument used in our research.

- Section A introduces the task and role-play to the participant.
- Section B contains the introduction and explanation of the Exchange Model.
- Section C contains the introduction and explanation of the Registration Model.
- Section D contains the post-task survey instrument.
- Section E contains the demographic questionnaire.

A. OVERALL INTRODUCTION

Welcome to our experiment. Today you will use two systems. These systems are developed to encrypt your emails so that your emails can be protected from being read by email providers (such as Google and Yahoo!), governments (e.g. NSA), as well as malicious attackers.

In this experiment, **pretend you are Henry**, and you want to send and receive encrypted emails to some people. Below are email addresses you may use in this experiment.

- Henry: researchmessage@gmail.com
- Henry2: researchmessage2@gmail.com
- Alice: alice.recipient@gmail.com
- Bob: bobby.recipient@gmail.com
- Carl: carl.recipient@gmail.com

B. EXCHANGE MODEL

Below is how *Lock Exchange System* works.

1. Every user can get a public lock and a private key.

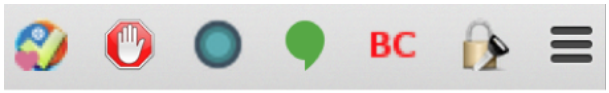
2. Users have to exchange their public locks in some way.

3. You can send encrypted emails with others' public locks, so that others' can read the emails with their private keys.

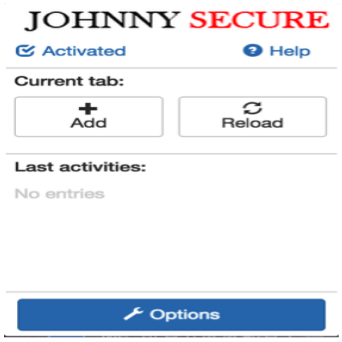
4. Similarly, you can also read any encrypted emails that are encrypted to you using your private key.

Task Instructions:

- Click the extension on upper right corner on tool bar in Chrome.



- Click "Options" for configuration.



1. Generate a public lock/private key pair.

   Go to "Generate Lock and Key" to generate a **public lock/ private key pair**. Note: The password is only for this study, and is NOT your email password. **DON'T** use your real passwords associated with any of your account in real life.

2. Exchange Public Locks with Alice.

   (a) Go to "Display Lock/Key Pair" and click the lock/key pair you just generated. Then export your public lock to Alice.
   The public lock will **start with** "-----BEGIN PGP PUBLIC LOCK BLOCK-----", and **end with** "-----END PGP PUBLIC LOCK BLOCK-----" (Note: there are FIVE "-" in the beginning and in the end).
   You can send your public lock by one or combination of ways that we **provide** you.

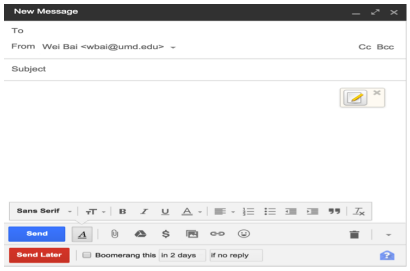
   (b) Then you will receive Alice's public lock.

   (c) Import Alice's public lock into the extension.

3. Send an encrypted email to Alice

   In the email interface, first click the encryption icon to write "What is your favorite color" to Alice. If the icon doesn't show up, please refresh the website.

   Note: you need to encrypt for Alice.



4. Decrypt the received email from Alice.

   Move your mouse to the email body. When a lock icon appears, click on the icon. You need your password (you created in step 3) to decrypt email.

   Next you will send encrypted email to two recipients Bob and Carl.

5. Exchange Public Locks with Bob and Carl.

   You can use the same way or different way provided in step 4 to exchange public locks with Bob and Carl.

6. Send encrypted email to Bob and Carl.

   Imagine that you are a financial secretary in your department, and you want to send the payroll reports to Bob and Carl by encrypted email. For simplicity, you can simply write "Here is your biweekly payroll summary: Salary is $888.88, Tax is $88.88. Your subtotal: $800.00." in the email body. You can refer to previous steps to send encrypted email.

7. Decrypt the received email from Bob and Carl.

8. Imagine that you are still the financial secretary in your department, and you will send the payroll reports to 10 people by encrypted email, what will you do? Please specify the steps.

9. Misconfiguration

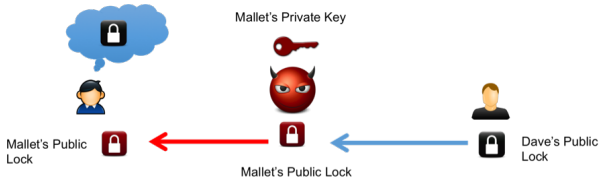
   (a) If you accidentally delete or lose Alice's public lock, what will you do if you want to send/receive encrypted email to/from Alice?

   (b) If you accidentally delete or lose your own private key, what will you do if you want to send/receive encrypted email to/from other recipients?

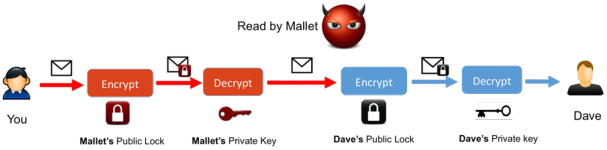
   (c) If you accidentally publicize your own private key, what will you do if you want to send/receive encrypted email to/from other recipients?

Possible Threats for *Lock Exchange System*:

These systems are developed to encrypt your emails so that your emails can be protected from being read by email providers (such as Google and Yahoo!), governments (e.g. NSA), as well as malicious attackers.



The threat may happen when you exchange public locks with others. When you try to get the public lock from Dave, Mallet (can be any type of attacker from above) secretly switches the public lock to his own. You think you get Dave's public lock, but in fact you get Mallet's.



Then when you send encrypted email to Dave, you actually use Mallet's public lock. As a result, Mallet can read your email. Mallet will consequently use Dave's public lock and send the encrypted email to Dave, so that both you and Dave don't realize the email has been read.

This threat doesn't happen usually, because it requires Mallet to have much power and resources to achieve this.

Please give your feedback about *Lock Exchange System:*

Note: We are evaluating these systems. We are not testing you. These systems are not developed by us. Please leave your feedback as honestly as you can. Your honest feedback, positive or negative, will help with our research.

For the first two questions, please note the difference between difficulty and cumbersomeness. Difficult tasks are intellectually challenging and need effort or skills to accomplish. Cumbersome tasks are tedious and need an unnecessarily long time to accomplish.

Please rate your agreement with the following statements.

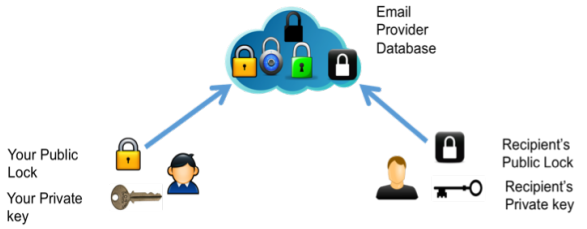
1. The following tasks were difficult.
   (a) Generate the public lock and private key pair
   (b) Exchange public lock with Alice
   (c) Send encrypted email to Alice
   (d) Decrypt email from Alice
   (e) Exchange public locks with Bob and Carl
   (f) Send encrypted email to Bob and Carl
   (g) Decrypt email from Bob and Carl
   (h) Send and receive encrypted emails to 10 people
2. The following tasks were cumbersome.
   (a) Generate the public lock and private key pair
   (b) Exchange public locks with Alice
   (c) Send encrypted email to Alice
   (d) Decrypt email from Alice
   (e) Exchange public locks with Bob and Carl
   (f) Send encrypted email to Bob and Carl
   (g) Decrypt email from Bob and Carl
   (h) Send and receive encrypted emails to 10 people
3. This system effectively protected my privacy.

C. REGISTRATION MODEL

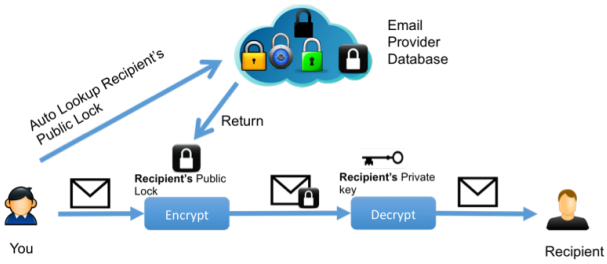
Instruction: Below is how *Registration System* works.

1. Every user can get a public lock and a private key when you register.

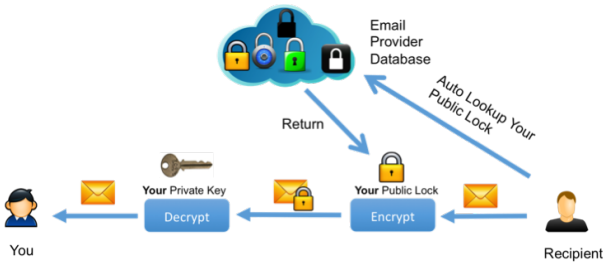
2. Every user's public lock will be automatically stored in a cloud database that is run by the email provider.



3. You can send encrypted emails with others' public locks, so that others' can read the emails with their private keys. The cloud database will return others' public locks for you.



4. Similarly, you can also read any encrypted emails that are encrypted to you using your private key.

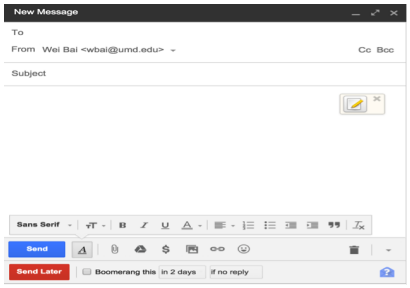


Task Instructions:

- Click the extension on the upper right corner on the tool bar in Chrome.

- Click "Options" for configuration.

1. Register
   Go to "Register" to register your email account to the email provider server. The registration will give you a public lock and a private key.
2. Send an encrypted email to Alice
   In the email interface, first click the encryption icon to write "What is your favorite color" to Alice. If the icon doesn't show up, please refresh the website.
   Note: you need to encrypt for Alice.
3. Decrypt the received email from Alice
   You need your password (you created in step 3) to decrypt email.
   Next you will send encrypted email to two recipients Bob and Carl.



4. Send encrypted email to Bob and Carl.

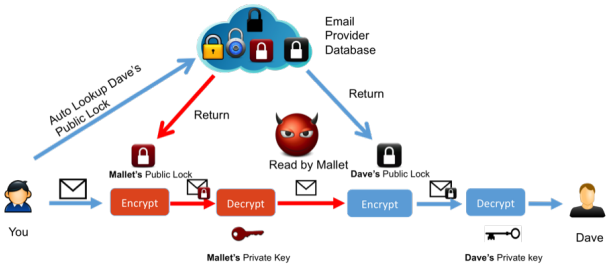
   Imagine that you are a financial secretary in your department, and you want to send the payroll reports to Bob and Carl by encrypted email. For simplicity, you can simply write "Here is your biweekly payroll summary: Salary is $888.88, Tax is $88.88. Your subtotal: $800.00." in the email body. You can refer to previous steps to send encrypted email.

5. Decrypt the received email from Bob and Carl
6. Imagine that you are still the financial secretary in your department, and you will send the payroll reports to 10 people by encrypted email, what will you do? Please specify the steps.
7. Misconfiguration
   (a) If you accidentally delete or lose your own private key, what will you do if you want to send/receive encrypted email to/from other recipients?
   (b) If you accidentally publicize your own private key, what will you do if you want to send/receive encrypted email to/from other recipients?

Possible Threats for *Registration System*:

These systems are developed to encrypt your emails so that your emails can be protected from being read by email providers (such as Google and Yahoo!), governments (e.g. NSA), as well as malicious attackers.

There are two prototypes for *Registration System*. For the first prototype (Model 1), the possible threats are as follows.



The threat may happen when you send encrypted emails to others. For example, when you try to send encrypted emails to Dave, you think the email provider database will return Dave's public lock to you. But in fact it returns Mallet's, so that Mallet can read your email. Therefore, you need to trust the email provider in this system.

In the second prototype (Model 2), there is a third-party service (not the email provider) as an intermediary. In this prototype, neither the third-party service nor your email provider can read your email themselves. However, if your email provider and the third-party service collaborate, they can both read your email. Therefore, you need to trust that the two services are not collaborating.

Please give your feedback about *Registration System*:

Note: We are evaluating these systems. We are not testing you. These systems are not developed by us. Please leave your feedback as honestly as you can. Your honest feedback, positive or negative, will help with our research.

For the first two questions, please note the difference between difficulty and cumbersomeness. Difficult tasks are intellectually challenging and need some effort or skills to accomplish. Cumbersome tasks are tedious and need an unnecessarily long time to accomplish.

Rate your agreement with the following statements.

1. The following tasks were difficult.
   (a) Register
   (b) Send encrypted email to Alice
   (c) Decrypt email from Alice
   (d) Send encrypted email to Bob and Carl
   (e) Decrypt email from Bob and Carl
   (f) Send and receive encrypted emails to 10 people
2. The following tasks were cumbersome.
   (a) Register
   (b) Send encrypted email to Alice
   (c) Decrypt email from Alice
   (d) Send encrypted email to Bob and Carl
   (e) Decrypt email from Bob and Carl
   (f) Send and receive encrypted emails to 10 people
3. This system effectively protected my privacy.

D. OVERALL FEEDBACK

Please give your overall feedback about these two systems:

Note: Again, please give your honest feedback to help with our research.

1. Please rate your willingness to use these two systems in the future.
   (a) I would like to use *Lock Exchange System*.
   (b) I would like to use *Registration System* with Model 1.
   (c) I would like to use *Registration System* with Model 2.
2. Below please rate your willingness to recommend these systems to others.
   (a) I would like to recommend *Lock Exchange System* to others.
   (b) I would like to recommend *Registration System* with Model 1 to others.
   (c) I would like to recommend *Registration System* with Model 2 to others.
3. Please rate your agreement with the following statements.
   (a) I think that I would need the support of a technical person to be able to use Lock Exchange System.
   (b) I think that I would need the support of a technical person to be able to use Registration System.

(c) I would imagine that most people would learn to use Lock Exchange System very quickly.
(d) I would imagine that most people would learn to use Registration System very quickly.
(e) I would need to learn a lot of things before I could get going with Lock Exchange System.
(f) I would need to learn a lot of things before I could get going with Registration System.

4. What do you like or dislike for each system? Why?

In Model 3, the email provider will still store all users' public locks, just like Model 1. But there are other parties (auditors) who audit the email provider, to ensure that the email provider is giving out correct public locks. These auditors may include other email providers, public interest groups, and software on your devices. If the email provider gives you a public lock that doesn't belong to the recipient, or gives someone else the wrong public lock for you, these parties will notify you. You (or someone else) may use the wrong lock temporarily (for an hour or a day) before you are notified.

In this model, you don't need to trust any email provider, but you need to trust the auditors and/or the software on your device. Because there are several auditors, even if one auditor does not alert you another one probably will.

E. DEMOGRAPHICS

1. Which of the following best describes your current occupation?
   (a) Healthcare Practitioners and Technical Occupations
   (b) Office and Administrative Support Occupations
   (c) Production Occupations
   (d) Farming, Fishing, and Forestry Occupations
   (e) Computer and Mathematical Occupations
   (f) Community and Social Service Occupations
   (g) Life, Physical, and Social Science Occupations
   (h) Management Occupations
   (i) Legal Occupations
   (j) Installation, Maintenance, and Repair Occupations
   (k) Food Preparation and Serving Related Occupations
   (l) Architecture and Engineering Occupations
   (m) Arts, Design, Entertainment, Sports, and Media Occupations
   (n) Building and Grounds Cleaning and Maintenance Occupations
   (o) Healthcare Support Occupations
   (p) Construction and Extraction Occupations
   (q) Education, Training, and Library Occupations
   (r) Protective Service Occupations
   (s) Sales and Related Occupations
   (t) Business and Financial Operations Occupations
   (u) Transportation and Materials Moving Occupations
   (v) Other (please specify)
2. Where did you grow up (primarily)?
   (a) United States
   (b) Other North America
   (c) South or Central America
   (d) Western Europe
   (e) Eastern Europe
   (f) Africa
   (g) South Asia (India, Bangladesh, Pakistan, etc.)
   (h) East Asia (China, Japan, Korea, etc.)
   (i) Central Asia
   (j) The Middle East
   (k) Australia / Oceania
   (l) Other: [please specify]
   (m) I prefer not to answer
3. What is your age?
   (a) 18-20
   (b) 21-24
   (c) 25-34
   (d) 35-44
   (e) 45-54
   (f) Above 54
   (g) I prefer not to answer
4. What is your gender?
   (a) Male
   (b) Female
   (c) I prefer not to answer
5. Please tell us whether you have the following experiences (yes or no).
   (a) I have attended a computer security conference in the past year.
   (b) I have taken or taught a course in computer security before.
   (c) Computer security is one of my primary job responsibilities.
   (d) I have used SSH before.
   (e) I have configured a firewall before.
   (f) I have a degree in an IT-related field (e.g. information technology, computer science, electrical engineering, etc.)?
   (g) I have an up-to-date virus scanner on my computer.

User Attitudes Toward the Inspection of Encrypted Traffic

Scott Ruoti*†, Mark O'Neill*†, Daniel Zappala*, Kent Seamons*
Brigham Young University*, Sandia National Laboratories†
ruoti@isrl.byu.edu, mto@byu.edu, {zappala, seamons}@cs.byu.edu

ABSTRACT

This paper reports the results of a survey of 1,976 individuals regarding their opinions on TLS inspection, a controversial technique that can be used for both benevolent and malicious purposes. Responses indicate that participants hold nuanced opinions on security and privacy trade-offs, with most recognizing legitimate uses for the practice, but also concerned about threats from hackers or government surveillance. There is strong support for notification and consent when a system is intercepting their encrypted traffic, although this support varies depending on the situation. A significant concern about malicious uses of TLS inspection is identity theft, and many would react negatively and some would change their behavior if they discovered inspection occurring without their knowledge. We also find that a small but significant number of participants are jaded by the current state of affairs and have lost any expectation of privacy.

1. INTRODUCTION

In early 2013, one of the authors received an email from a former student who expressed serious concerns after becoming aware that his employer was inspecting its employees' encrypted Internet traffic in order to protect the network from attackers. Though he was himself employed in the computer security industry, he expressed surprise and anger that this could happen, and also mentioned his serious concerns about the potential for employees to disclose personal information without being aware that their data was visible to their employer. He questioned whether this practice was legal and whether it was ethical to do this without notifying employees in advance.

In fact, it is common practice for companies to inspect employees' encrypted traffic to filter malware and viruses, prevent the leak of intellectual property, and block harmful websites [2, 25, 20]. This inspection is usually accomplished with a network device that acts as a TLS/SSL proxy, sitting in the middle of the communication between a browser and web server where it can intercept, decrypt, inspect, then re-encrypt and forward on the user's traffic to its original destination. This is all accomplished without any visible notification to the user that their encrypted traffic is being inspected.

While security experts overwhelmingly view the inspection of encrypted traffic by attackers and governments as undesirable, the practice of businesses and organizations inspecting their *own* encrypted traffic in order to secure their *own* network and intellectual property is more controversial. Many experts are alarmed by any use of a TLS proxy because it is deceptive; users' browsers continue to inform them they have a secure connection to the server, even though this is not the case. Most research in the literature treats all TLS proxies as undesirable and actively tries to prevent their use [4]. Still, a smaller number of researchers are investigating how the malicious uses can be prevented while still allowing for benevolent use of encrypted traffic inspection by businesses and organizations [17, 14].

While the opinions of businesses and security experts regarding the inspection of encrypted traffic are known, no prior work has measured general (i.e., non-expert) user attitudes and preferences toward the inspection of encrypted traffic. To better understand users' perspectives on this issue, we surveyed 1,976 people across two surveys regarding their opinions of TLS proxies and their use in inspecting encrypted traffic.[1] The results of the first survey of 1,049 individuals showed a surprising willingness by participants to accept the inspection of encrypted traffic, provided they are first notified. Based on the results of the first survey, we conducted a second survey of 927 individuals to further explore user attitudes towards inspection of encrypted traffic in specific situations.

Our contributions from these surveys include the following insights:

This work was supported by Sandia National Laboratories, a 2014 Google Faculty Research Award, and the National Science Foundation under Grant No. CNS-1528022. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22–24, 2016, Denver, Colorado.

[1]The full data from both surveys is available at `https://soups2016.isrl.byu.edu/`.

- User opinions toward TLS proxies and the inspection of encrypted traffic are nuanced. Many express concerns about privacy and identity theft from hackers (75.8%) or government surveillance (70.9%). Yet there is broad, general acceptance of TLS proxies when used by employers, schools, etc (71.7%).
- Most participants indicated support for the inspection of encrypted traffic as long as they were first notified of it (90.7%). Likewise, participants indicated strong support for legislation requiring notification or consent (83.2%).
- When asked about specific situations in which TLS proxies might be used (e.g., at work, at school, at a café, or at home), support for TLS proxies ranges from 65% to 90% of participants (including those who want notification or consent). Support for inspection of encrypted traffic without notification or consent is strongest at elementary schools (45.9%) and at businesses when employees are using company-provided computers (47.9%). Participants generally favor consent in cases when they feel in control (at home, free WiFi, their own device at work) versus notification when an organization is in control (public library, school, company computer). In nearly all the scenarios we posed, only a small minority of the participants indicated that using TLS proxies is not acceptable. The one exception is government surveillance, in which case 47.5% say that this is not acceptable.
- Many users would have a negative opinion if they discovered that the owner of their network used a TLS proxy without prior notification and/or consent (60.8%), though for some (34.2%) it would depend on who the owner was and how they were using the technology. Some would change their behavior on the network, either discontinuing to use it (17.2%) or changing which sites they visited (6%).
- We identify personas based on participants' responses regarding TLS proxies: pragmatic (76.5%), privacy fundamentalist (17.0%), jaded (5.0%), and unconcerned (1.0%). Jaded participants are interesting in that their opinions regarding privacy and security align with the privacy fundamentalist persona, but their practices align with the unconcerned persona. This dichotomy stems from the fact that these users feel that regardless of what steps they take, they are powerless to prevent compromise of their online information, and so choose to not do anything to protect themselves.

While several of our findings might seem intuitive, it is important to ground intuitions in data, and this paper provides the first survey of user opinions on this topic. In addition, participants showed a high level of engagement in the survey, notwithstanding the complexity of the topic. Many users shared in-depth analysis of trade-offs in open responses, demonstrating that they care deeply about this issue. User attitudes toward TLS proxies provide an important data point along the spectrum of discussion that is currently taking place regarding who should have access to encrypted information.

2. BACKGROUND

The focus of our surveys is on user attitudes towards the inspection of encrypted traffic (i.e., HTTPS), specifically with the use of TLS proxies. In this section we provide technical details regarding TLS proxies. We also discuss real-world examples of how TLS proxies are used. Finally, we present related work on measuring user sentiment towards online privacy.

2.1 TLS Proxies

When a web browser attempts to validate the identity of a website, it relies on certificate authorities (CAs) that digitally sign certificates vouching for the identity of servers. Web browsers authenticate a site by validating a chain of trust from the site's certificate back to one of a set of trusted root certificates. These certificates comprise the *root store* and are typically bundled with the operating system or browser.

This validation system is currently being co-opted by the use of TLS proxies that act as a man-in-the-middle (MitM) for TLS connections. A TLS proxy can issue a *substitute certificate* for any site the user visits, so that the user establishes an encrypted connection to the proxy rather than the desired web site. The proxy can then decrypt and monitor or modify all user traffic, before passing it along via a second encrypted channel to the desired web site. For example, when a user attempts to create a secure connection to Amazon by requesting Amazon's certificate, the proxy intercepts this request, generates a certificate for Amazon, and sends this substitute certificate back to the user's machine. The user's machine will then create a secure connection to the proxy (instead of Amazon) and send all of its data to the proxy, which has full access to it before forwarding it on to Amazon's servers.

TLS proxies can be used for both benevolent and malicious purposes. Some companies use TLS proxies to filter malware and viruses, prevent the leak of company secrets and intellectual property, block harmful websites, or catch malicious insiders. However, less scrupulous companies, government agencies, crime organizations, and others may also use proxies to steal a user's sensitive data, conduct surveillance, or commit identity theft. Currently, browsers and users have no method for distinguishing between benevolent and malicious TLS proxies, and the user is entirely unaware that an organization or attacker is intercepting encrypted traffic. Even when a TLS proxy is present, the browsers displays a reassuring lock icon that could mislead users to assume they are communicating securely with the website.

To avoid browser warnings that self-signed substitute certificates would trigger, TLS proxies generate substitute certificates signed by a CA that the user's machine trusts. This can be done in several ways:

- Purchasing an intermediate certificate authority certificate.
- Installing a new trusted root certificate on the user's machine. This can be done either by businesses (e.g., custom system image, manual installation, enterprise PKI system) or by malware.

- Including the certificate on a device's root store when it is manufactured. Nokia was recently found to be using TLS proxies on mobile devices [18] and Lenovo has pre-installed software using a TLS proxy on its laptops [21].
- Controlling a root certificate authority. Some governments are in this position, and evidence suggests that even when governments do not own the root they can coerce authorities into granting them certificates for domains they do not own [15, 23].
- Stealing existing root and intermediate certificate authority certificates [15, 7].

2.2 Real-world Examples

There are a variety of real-world scenarios, ranging from suspicious to malicious, where inspection of encrypted traffic is documented as having occurred.

Reports have notified the public that both Nokia and Lenovo used TLS proxies to decrypt customer (not employee) traffic for reasons other than security. Nokia decrypted cell phone data, allegedly to improve performance on their cellular network [18]. Some Lenovo laptops came with third party software that inserted ads into encrypted data [21]. Weaknesses in the adware implementation left users vulnerable to attack from malicious outsiders. Public outcry caused both companies to stop accessing encrypted traffic.

Government surveillance has been reported to use similar methods [24]. A report from 2011 showed that Iran monitored 300,000 citizens online using a stolen certificate from Diginotar, a company that is trusted to certify legitimate websites [7].

Two recent measurement studies show that TLS proxies account for about 1 in 250 encrypted connections on the web [10, 19]. The vast majority of these monitored connections are for benevolent purposes, but a small percentage appear to be adware, grayware, and otherwise suspicious activity.

The TLS proxy capability is essentially a backdoor into the current web authentication system. This backdoor has benevolent uses to strengthen the security of users and organizations, and a majority of users support their use. As with any backdoor, it's very existence increases the attack surface that can be exploited by attackers. For example, a recent study of client-side TLS proxies used in personal firewalls and parental filters discovered implementation flaws in a number of products that open the user to attack and weaken their security [5].

2.3 Related Work

There have been prior studies that survey user's attitudes about their online security and privacy. Still, no prior study has looked specifically at user attitudes toward the inspection of encrypted traffic.

McDonald and Cranor [16] used interviews and a survey to explore user's knowledge and perception of online behavioral advertising practices. They discuss the potential chilling effect of these practices based on 40% of the users that self-reported they would change their behavior if they learned advertisers were collecting data. Similarly, users reported in our survey that they would change their behavior if they learned that their encrypted data was being inspected.

Ur et al. [27] also studied user opinions about online behavioral advertising by conducting 48 semi-structured interviews with non-technical users. Similar to our work, they found users had nuanced opinions about the trade-offs for a technology that was both useful and privacy invasive. They determined that users were not receiving effective notice and choice mechanisms. Our surveys reveal a strong desire for notification and choice regarding the inspection of encrypted traffic.

Shay et al. [22] surveyed users via Amazon Mechanical Turk about their attitudes and experiences with compromised email or social networking sites. They found that many respondents gave high quality responses to open response questions and discussed implications for security mechanism designers. Likewise, our work has significance for the designers of mechanisms to inspect encrypted traffic.

Anton et al. [1] surveyed users in 2008 to see if their attitudes on privacy concerns had changed from the same survey administered in 2002. They found that the top three concerns of U.S. users were information transfer, notice/awareness, and information storage. While the top three concerns had not changed, their level of concern had risen. The top three concerns for European users were the same but in a different order; notice/awareness came in third place. Concerns for notice/awareness are important to both groups, and was a prominent factor in our surveys.

Woodruff et al. [29] examined how well users' classification by the Westin Privacy Segmentation Index predicted their actual behavior. They found that although many participants were classified as privacy fundamentalists, their actions in hypothetical situations were not consistent with this classification. Similarly, while we group participants into personas with names similar to the Westin categories, we do so by looking at how participants indicate they would react to hypothetical situations and not using any of Westin's several privacy indexes.

3. FIRST SURVEY – METHODOLOGY

In February 2014, we conducted the first online survey using the Amazon Mechanical Turk (MTurk) crowdsourcing service. We gathered responses on Wednesday, February 12, 2014 between 7:50 AM and 5:22 PM (PST). Each participant could take the survey once and received $1 USD as compensation upon completing the survey. In total 1,262 people completed the online survey. The survey was approved by our Institutional Review Board and is contained in Appendix A.

3.1 Instructing Participants

Before conducting this survey, we felt it was unlikely that most people would be aware of TLS proxies (an assumption that was upheld by our results). This presented a dilemma: either we would need to only survey individuals who were already aware of TLS proxies or we would need to instruct participants about TLS proxies. Both of these options have significant drawbacks. Limiting the survey to individuals with pre-existing knowledge regarding TLS proxies would

likely limit us to participants with highly technical backgrounds, thus failing to gather information about broader opinions related to the inspection of encrypted traffic. On the other hand, instructing participants on TLS proxies has the risk of unintentionally biasing them one way or another, and requires them to answer questions about a subject they potentially just learned about.

Because our research goal was to survey broad opinions regarding the inspection of encrypted traffic, we preferred not to limit our population to the small fraction of users who are already aware of this issue. Instead, we chose to accept the limitations related to instructing participants about TLS proxies and survey as many participants as possible. For our goals, this was preferable to ignoring the opinions of a large portion of users.

To address the risks related to instructing participants on an issue and then surveying them, we spent considerable effort and time crafting our description of TLS proxies. Our goals were to (1) give a simple and concise overview of how TLS proxies are used to inspect encrypted traffic, and (2) present participants with a fair and unbiased description of how the inspection of encrypted traffic could be used for both benevolent and malicious purposes.

In preparation for writing the description of TLS proxies, we examined the literature and observed that existing descriptions of TLS proxies were not neutral in tone and would unduly bias participants. We talked with businesses that sell proxies (i.e., Blue Coat, Symantec) and read opinions from privacy advocates to better understand both sides' opinions. Based on the information in these sources, we composed a draft of our description of TLS proxies, focusing on using language that was informative and neutral in tone, allowing participants to form their own opinions. Our team of researchers, which included members who are fundamentally opposed to TLS proxies and members who accept their benevolent uses, iterated on this description until all members were satisfied with its wording.

We then tested this description using a convenience sample of six individuals from our university who were not a part of our research group to ensure it was balanced and understandable. Based on feedback from the convenience sample, we made minor edits to the description.

Finally, we tested this revised description using MTurk to ensure that participants felt that the description was sufficiently understandable. Of the 80 participants in this pilot survey, nearly all participants (73; 91%) indicated that the description of TLS proxies helped them understand what TLS proxies are and how they are used (2 participants indicated the description was not helpful (2; 3%), with the remainder being undecided (5; 6%)). We also examined participant responses to free response questions and found that, as reported, most participants' answers reflected an accurate understanding of TLS proxies. As such, we included this version of the description in both surveys, as shown in Figure 1.

When you connect to the Internet you do so through some organization's network. For example, at home you connect to your Internet service provider's (ISP) network, while at work you connect to your employer's network. To protect your information from others on the network you can create secure connections to the websites you use (HTTPS). This is done automatically for you when you log into a website. The secure connection encrypts your Internet traffic so that no one else can view or modify your communication with the website (see Figure A).

Figure A

The network you use to connect to the Internet can also be set up to use a system called a TLS proxy. TLS proxies sit in the middle of your secure connection to the websites you view (see Figure B). At the TLS proxy your Internet traffic is decrypted and the web proxy can view and modify it. Afterwards, the TLS proxy will then re-encrypt your traffic and forward it along. This is done silently and without the knowledge of you or the website you connect to.

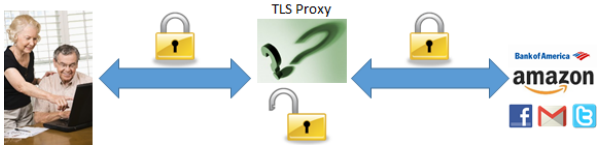


Figure B

TLS proxies can be set up by the organization that controls your Internet (for example, your ISP, school, or employer) and also by malicious attackers. TLS proxies have many different uses:

Protective	**Malicious**
Blocking malware and viruses	Stealing passwords
Protecting company secrets	Identity theft
Blocking harmful websites	Tracking government dissidents
Catching malicious individuals	Spying (for example the NSA)
	Censorship

Figure 1: TLS Proxy Description

3.2 Survey Contents

The survey begins by gathering demographic information. It then instructs participants about TLS proxies and their use in the inspection of encrypted traffic. Next, participants are asked to share their opinions regarding the use of TLS proxies and the inspection of encrypted traffic. These questions survey participant opinions as to whether TLS proxies are a breach of their privacy and whether there are acceptable uses for TLS proxies. Participants are also asked their reasoning for why TLS proxies should or should not be allowed. Also, participants are asked which

parties they are concerned about using TLS proxies and what, if any, measures should be used to regulate their use.

The survey then asks participants about how they would personally react to having a TLS proxy on a network they use to connect to the Internet. This section includes two open-ended questions, the first asking them what concerns they might have and the second asking them how it would affect their opinion of the organization running the TLS proxy. Finally, participants are given a chance to express any remaining comments they might have.[2]

3.3 Survey Development

Before running our survey, we conducted a pilot survey using MTurk to ensure that we would get meaningful and thoughtful results. This pilot survey was IRB approved and included 80 participants. Based on our analysis of participants' answers in this pilot survey, it was clear that participants generally understood the description of TLS proxies presented to them, and so we proceeded to launch the full survey. Responses from the pilot survey are not included in our results.

3.4 Qualitative Data Analysis

To better understand participants' opinions regarding TLS proxies and to avoid biasing their responses, we included several open-ended questions in the survey. For each question, we created a codebook to categorize participant responses. One researcher reviewed all the participant responses and created the initial codebooks. The codebooks were then modified through discussion with the coders.

After coding was completed, all of the coders met together to discuss the data. As part of this discussion they were encouraged to identify themes that they had seen in the data. Particular attention was paid to the themes that they felt the codebook did not adequately cover. Coders also shared responses that they felt best represented the various viewpoints expressed by participants.

In total, there were seven coders that analyzed the data. We validated the consistency of the coders using Fleiss' Kappa [6]. Coders' agreement ranged from "substantial agreement" to "almost perfect agreement" (with kappa values ranging from .687 to 1, mean of .865 and median of .833).

3.5 Amazon Mechanical Turk

We used Amazon Mechanical Turk (MTurk) to recruit survey participants. MTurk has become an increasingly popular method for gathering participant data for usability studies and user surveys. Buhrmester et al. found that MTurk participants are significantly more diverse than typical American College samples and that data obtained from MTurk studies is at least as reliable as those obtained via more traditional methods [3]. Kittur et al. used MTurk participants to classify Wikipedia entries and found that that they could produce results equivalent to expert raters [13]. While MTurk has known limitations, it is still a mostly reliable platform for rapidly obtaining results related to user sentiment [26, 12].

[2]As shown in the Appendix, questions are grouped onto several pages. After questions on one page are answered and the user continues with the survey, they are unable to return and modify their answers.

3.6 Quality Control

To ensure participants provided valid data, we accepted only participants that had previously completed 1,000 tasks on MTurk with an overall task approval rate of 95% or higher. Second, the seven coders examined participants' responses to open-ended questions in order to ensure that participants had both understood the description of TLS proxies and remained on topic. We validated the consistency of the coders' choice to exclude participants' responses using Fleiss' Kappa [6] and found that coders were in perfect agreement (kappa value of 1). During the coding process, a participant's responses were discarded if their answers were clearly spam (i.e., copying the text of a Wikipedia page), or they did not understand the questions being asked (i.e., their answers discussed HTTP proxies). In total, we excluded 153 participants' responses (12.1%) as spam and 60 participants (4.8%) as misunderstandings. The remaining 1,049 participants' responses constitute the results of our first survey.

3.7 Demographics

The demographics for the participants are shown in Table 1. Most participants were from the United States (87%), with the rest primarily from India (11.5%). Although results from a previous paper suggested that MTurk participants from India are less concerned with privacy [11], the results from our first survey found that they were more likely to report privacy concerns than their counterparts from the United States of America ($\chi^2[2, N = 1049] = 12.35, p < 0.01$).

Participants were skewed towards males (61%), and ages were centered around 25–32 (46%). Most participants were single (60%) and had no children (62%). Nearly all participants had completed high school, with the majority having completed some level of higher education (57%).

Participants were asked to self-report their level of knowledge of Internet security, with most rating somewhere between somewhat knowledgeable and mildly knowledgeable (78%).

After reading the description of TLS proxies, participants were asked whether they had prior knowledge of TLS proxies. Most participants reported having little to no awareness of TLS proxies before the survey: unaware (66.5%), unsure (8.1%), aware (25.4%). We speculate that due to the effects of illusory superiority, the number of participants that were unaware of TLS proxies before the survey was even higher than reported [8, 9]. Additionally, participants may have conflated knowledge of traditional web proxies with knowledge of TLS proxies.

3.8 Limitations

In our survey, participant demographics were slightly skewed towards a younger male population and nearly all participants were from the US and India. Additional work could be done to replicate our results with different populations. Cross-cultural, international surveys would be especially interesting, but these should be conducted by researchers that can engage participants in their native language and have an understanding of participants' cultural perceptions.

	Survey 1 (N=1,049)	Survey 2 (N=927)
Country		
United States	86.9%	94.3%
India	11.5%	5.7%
Other	0.3%	N/A
Gender		
Male	61.1%	60.6%
Female	38.6%	38.9%
Prefer not to answer	0.3%	0.4%
Age		
18–24 years old	18.7%	17.8%
25–34 years old	47.0%	45.8%
35–44 years old	19.6%	21.8%
45–54 years old	8.6%	7.9%
55+ years old	5.8%	6.3%
Prefer not to answer	0.3%	0.4%
Relationship		
Single	59.5%	60.9%
Married	35.5%	35.6%
Other	4.7%	2.7%
Prefer not to answer	0.6%	0.8%
Children		
Yes	36.6%	32.5%
No	62.3%	67.2%
Prefer not to answer	0.9%	0.3%
Education		
No diploma	1.0%	0.6%
High school	12.4%	11.0%
Some college or university credit	28.9%	29.3%
College or university degree	49.9%	50.5%
Post-Secondary Education	7.6%	8.4%
Prefer Not To Answer	0.2%	0.1%
Knowledge		
No Knowledge	4.6%	2.6%
Somewhat Knowledgeable	35.7%	32.4%
Mildly Knowledgeable	42.4%	47.8%
Highly Knowledgeable	14.4%	15.2%
Expert	2.4%	1.8%
Prefer Not To Answer	0.2%	0.2%

Table 1: Participant Demographics

As shown in prior work, participants' reported security preferences and desires do not always align with their actual behaviors [29]. Often users will report being more privacy minded than they are in practice. Interestingly, in our survey participants indicated a high level of acceptance for TLS proxies, which could suggest that real-world acceptance of TLS proxies is even higher than we measured. On the other hand, many participants reported wanting to have their consent obtained, or at least be notified of, the inspection of encrypted traffic; in practice, it is possible that fewer participants would actually be interested in being notified.

Finally, while we spent considerable effort to craft a fair and unbiased description of TLS proxies and the inspection of encrypted traffic, there is still the possibility that it had a significant effect on some participants' responses. For example, in the real world, users often learn about security issues from the news, which is often sensational and biased. In contrast, our description strove for neutrality, and as such may have led to users taking a more rational view of the inspection of encrypted traffic than would occur in the wild. While we chose to accept these limitations in order to obtain opinions from as many participants as possible, an open avenue for future research is to find a way to gather equally widespread opinions in a way that has fewer limitations.

4. FIRST SURVEY – RESULTS

In this section we discuss the results of our survey in three areas: acceptable uses for TLS proxies, general concerns toward their use, and the reaction participants would have if they discovered a network they use employed a TLS proxy.

4.1 Acceptable Uses of TLS Proxies

Figure 2 shows participant attitudes toward proxies. A somewhat surprising result is that participants largely (752; 71.7%) felt that there were acceptable uses for TLS proxies. This feeling prevailed even though nearly half of the participants (522; 49.8%) indicated that TLS proxies are an invasion of privacy, and only one-eighth of participants (185; 17.6%) felt they presented no invasion of privacy. There is a strong correlation between thinking TLS proxies were an invasion of privacy and believing that there were not acceptable uses for them ($\chi^2[4, N = 1049] = 141.50, p < 0.001$). Nevertheless, over a quarter of participants (297; 28.0%) felt that TLS proxies were an invasion of privacy, but still had acceptable uses.

To better understand what uses might be acceptable, we asked participants who felt there were acceptable uses to enumerate those uses in an open-ended question. The results from our coded responses are shown in the top part of Table 2. The acceptable uses are largely concentrated on three use cases:

1. **Protecting organizations (493; 65.6%).** Many participants felt that organizations (e.g., businesses, government agencies, schools, libraries) had a right to protect their own intellectual property and security. This included protecting the company from viruses and hackers, filtering inappropriate or potentially malicious websites, and preventing the leak of sensitive information. Participants mentioned that since these organizations provide the Internet for their employees or constituents they had a right to use TLS proxies on their own networks.
2. **Protecting individuals (339; 45.1%).** Participants saw value in businesses using TLS proxies to protect their customers. This protection came in one of two forms:
   - **Direct.** Antivirus applications and firewalls could use TLS proxies to filter malware and

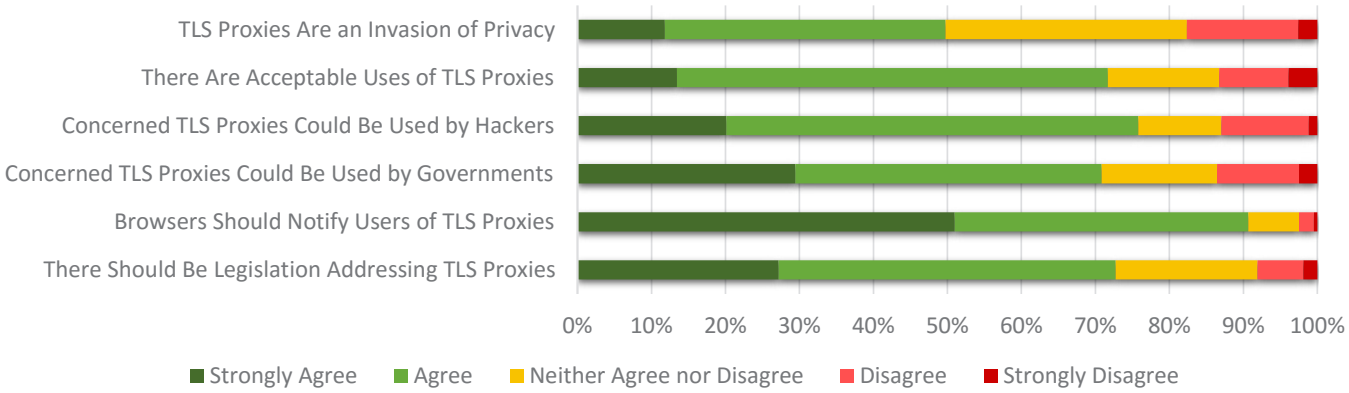


Figure 2: Participant Attitudes Toward TLS Proxies (N=1,049)

Opinion	Participants
Acceptable Uses	
Protect organizations	51.4% (n=539)
Protect individuals	34.8% (n=365)
Law enforcement and surveillance	8.9% (n=93)
Censor content	7.1% (n=75)
Never censor content	3.1% (n=32)
Acceptable at work, not at home	2.9% (n=30)
Concerns	
Hackers and spying	60.5% (n=635)
Privacy and identity theft	55.4% (n=581)
Done without knowledge or consent	13.2% (n=138)
Reactions	
Negative	60.8% (n=638)
Positive	5.0% (n=52)
Depends	34.2% (n=359)
Suspicious	25.8% (n=271)
Discontinue use	17.2% (n=180)
Change behavior (besides discontinue)	6.2% (n=65)

Table 2: Qualitative Response Categorization (N=1,049)

viruses. Similarly, ISPs could use TLS proxies to detect and prevent phishing attackers and block other inappropriate or malicious websites.

- **Indirect.** Participants recognized that they have a significant amount of private information stored externally on the web (e.g., at Amazon or Google). In order to protect this data, participants hoped that the companies storing their private data would employ TLS proxies internally to ensure the safety of the customer's data.

3. **Law enforcement and surveillance (65; 8.6%).** Nearly a tenth of participants expressed that law enforcement agencies should also be allowed to use TLS proxies. This includes use by local or federal agencies to track criminal or terrorist activity. Several participants also expressed that while this was a legitimate use it should only be done with a valid warrant or if there was an imminent threat to national security.

4.2 Concerns

Even though many participants in the first survey saw acceptable uses for TLS proxies, they were not without concerns or reservations. Based on our coding, we grouped these concerns into the categories shown in the middle part of Table 2. Three-quarters of the participants (795; 75.8%) mentioned they worried about hackers and nearly as many were concerned about the possibility for governmental spying (743; 70.9%). There was also a strong correlation between the concern that hackers could use TLS proxies and that the government could use them ($\chi^2[4, N = 1049] = 194.57, p < 0.001$).

The most visceral concerns were related to the breach of privacy. One of the open response questions asked participants to list what possible concerns they had regarding the use of TLS proxies. Over half of participants (581; 55.4%) mentioned they were concerned with a loss of privacy and personal information. Nearly a tenth of participants (104; 9.91%) mentioned having their identity stolen, and even more participants had answers that addressed the issue of identity theft generally.

A non-negligible number of the participants freely shared that either they, a family member, or other acquaintance had been the victim of account compromise. Similar to the finding of Shay et al. [22] this was a traumatic experience and it left participants especially concerned that TLS proxies could be used to perpetrate identity theft. R208 shared,

> *"A major concern that I would have would be the security of my personal and financial information. I have many friends who have been victims of identity theft and fraud, and would hate to have to go through what they did."*

Participants were also concerned that TLS proxies could be used without their knowledge. One-eighth of participants (138; 13.2%) mentioned in the open response question that they were concerned with privacy. Furthermore, when directly asked about notification, an overwhelming majority of participants (951; 90.7%) asserted they wanted

to be notified by their browsers of the presence of TLS proxies. Similarly, participants largely (942; 89.8%) felt that there should be legislation concerning TLS proxies. Most (782; 74.5%) wanted legislation to require notification, and nearly as many (701; 66.8%) wanted legislation to require consent.

4.3 Reactions

Participants in the first survey had varied responses on how they would react to learning that they currently use a network that employs TLS proxies. Based on our coding, we grouped these concerns into the categories shown in the bottom part of Table 2. Over half of participants (638; 60.8%) mentioned that it would negatively affect their opinion of the owner of that network. For example, R77 stated,

> *"I would be angry and would feel that organization violated my trust. I would wonder what information that organization had been collecting on me and what they planned to do with it. If it was my employer, I also would think that organization did not trust me and would consider working somewhere else."*

Still, a third of participants (359; 34.2%) said that their reaction would depend on who the owner of the network was and how they were using the proxy. For example, if the owner of the network was their employer they would not have a negative reaction, but if it was their ISP or government they would be very unhappy. Participants also mentioned that their approval would rest on whether or not any personal information was collected and/or sold and whether their consent had first been obtained. R960 explained,

> *"It would be on a case by case basis. I can see some instances where it would be understandable, but if it was going on without my consent, I would be wary of dealing with them in the future."*

Participants also mentioned ways in which their behavior would change if they learned a network was employing a TLS proxy. A quarter of participants (271; 25.8%) said that it would make them suspicious of the owner of that network. A quarter of participants (245; 23.4%) also mentioned that they would change their behavior on that network. For some participants (180; 17.2%) this included discontinuing use of the network and its services, while others (65; 6.2%) mentioned they would change the content they looked at on the Internet or be more careful about entering personal information, including but not limited to e-commerce transactions. At the extreme, some participants mentioned they would quit their job if they found that their employer's network used a TLS proxy. For example, R127 expressed,

> *"If my employers were secretly spying on my private data, I would sue them if legally possible, and quit the job regardless."*

Persona	Number	Percent
Pragmatic majority	802	76.5%
Privacy fundamentalist	178	17.0%
Jaded	48	4.6%
Unconcerned	11	1.0%
Unclassified	10	1.0%

Table 3: Participant Persona Categorization (N=1,049)

4.4 Personas

As our research group discussed the answers to open response questions in the first survey, it became clear that the participants could generally be classified into one of four personas: *pragmatic*, *privacy fundamentalist*, *jaded*, and *unconcerned*. After recognizing this, two members of the research group re-evaluated 90 participant responses and categorized participants into one of these four personas. The Fleiss' Kappa for this classification was 1 (i.e., perfect agreement). One researcher then classified the rest of the responses. The breakdown of participants into these categories is given in Table 3.[3]

Even though three of these personas have similar names to personas formulated by Westin [28], our categories are in no way based on the research of Westin. Instead, our methodology for creating personas more closely relates to that of Woodruff et al. [29], i.e., analyzing how participants indicate they would act in various privacy-related situations in order to determine their persona. Moreover, we do not intend these personas to be a definitive list of privacy personas, but rather view them as a helpful way to identify trends within our data.

4.4.1 Pragmatic Majority, N=802

The pragmatic majority weighed consumer benefits and protections of public safety against costs of intrusive practices, believed that organizations should earn the public's trust, and wanted to have the opportunity to opt-out of intrusive practices. This group was strongly correlated with being more likely to feel that there were acceptable uses for TLS proxies ($\chi^2[4, N = 1028] = 230.48$, $p < 0.001$). R93 stated,

> *"I think it is perfectly acceptable for organizations (companies, schools, libraries, etc.) to use TLS proxies because it protects their computers. It keeps hackers from getting to sensitive or confidential information of the organization. In addition, it blocks harmful viruses that can cause a lot of damage and expense in repair. It can also keep individuals from accessing websites (employees from playing online games or minors from accessing pornography). It is perfectly reasonable for companies to employee[sic] this device for these purposes when an individual is using their computer. We should not expect privacy when we are using someone else's computer."*

[3]There were ten participants whose answers were vague enough that we did not feel comfortable classifying them as any of the personas.

Though the pragmatic majority all weighed consumer benefits versus intrusive practices, they were not uniform in their conclusions about where and how TLS proxies should be used. Some recognized the right of employers to use them, while others believed they should only be allowed in narrow cases such as with a warrant.

4.4.2 Privacy Fundamentalist, N=178

The privacy fundamentalist was generally distrustful of organizations that ask for personal information, in favor of legislation enhancing privacy, and chose privacy controls over consumer benefits when a trade-off existed between the two. These participants were strongly correlated with being more likely to feel TLS proxies were an invasion of privacy ($\chi^2[4, N = 1028] = 114.81, p < 0.001$). These participants were also more likely to support legislation of TLS proxies ($\chi^2[2, N = 1028] = 14.40, p < 0.001$).

The defining feature of the privacy fundamentalist was that they viewed privacy as so important that it could not be traded for any benefit, no matter how great. As emphatically stated by R1119, *"I believe privacy is sacrosanct and one could argue that it's a Constitutional right."*

They were also likely to relate the use of TLS proxies to more traditional methods of surveillance such as wiretapping and intercepting mail.

4.4.3 Jaded, N=48

Jaded individuals were aware that violations of privacy happen regularly, believed that governments conduct surveillance on the general public, and had lost hope that they can have privacy online. These participants felt that "the system" was rigged to remove any real chance of them having privacy. For example, R713 expressed,

> *"I know that it is my choice to use the internet; however, since I live in a remote area with no transportation to the nearest city (30 miles away) I am 'stuck' working and banking and doing business on the internet. I feel it is unfair to be made to choose between being 'safe' and having privacy freedom. I am especially disgusted by our government's spying behaviors and the rhetoric about it being necessary for national defense."*

Likewise, when asked about concerns regarding the use of TLS proxies, R831 shared,

> *"None. The government (via the NSA) is already reading everything we do and share online. So no surprises there."*

Other jaded participants felt they had no choice in the matter because in the United States Internet service providers often have a monopoly.

4.4.4 Unconcerned, N=11

Unconcerned participants were generally trustful of organizations that ask for personal information, willing to sacrifice personal privacy to obtain consumer benefits, and not in favor of legislation to protect or enhance privacy. In our survey, we found very few unconcerned participants (1%). It is possible that the recent news regarding widespread government surveillance caused participants to be more privacy aware and sensitive. In addition, our use of qualitative data to classify participants allowed us to recognize that participants were part of the pragmatic majority even when their Likert responses might seem to indicate otherwise.

5. SECOND SURVEY – METHODOLOGY

Our first survey revealed that participants' opinions related to TLS proxies were closely tied to the situation in which TLS proxies were being used. To better clarify user feelings in this area, we formulated a second survey in which we ask participants about a series of specific scenarios where inspection of encrypted traffic could be used. This second survey serves to give quantitative backing to the qualitative data gathered in the first survey.

We collected data for our second survey on Tuesday, February 24, 2015 between 11:02 AM and 1:06 PM (PST). Each participant could take the survey once and received $1 USD as compensation upon completing the survey. The survey begins exactly as the first survey by gathering demographic information and then instructing participants about TLS proxies and their uses, both benevolent and malicious. Participants are then asked their opinions regarding the use of TLS proxies in various circumstances. In total 1,005 people completed the online survey. The survey was also approved by our Institutional Review Board and is contained in Appendix B.

5.1 Survey Description

The first portion of the second survey includes the same description of TLS proxies as the first one. It then asks several questions repeated from the first survey: whether TLS proxies are an invasion of privacy and whether there are acceptable uses for TLS proxies.

The main portion of this survey asks participants their opinion regarding different situations where TLS proxies may be used to inspect encrypted traffic, such as by an employer, at a school, or a café with free WiFi. The full list of scenarios is given in Figure 4. For each situation, participants are asked whether the organization should be allowed to run a TLS proxy, with responses taken from (1) *No*, (2) *Only if I consent*, (3) *Only if I am notified (consent not required)*, (4) *Yes (neither notification nor consent required)*, or (5) *Unsure*. To choose the situations, we used responses from open-ended questions in the first survey, along with suggestions from our research team to fill out the list. Finally, we had a single open-ended question where participants could share any opinions they still had remaining at the end of the survey.

We note that this survey had the same limitations as our first survey.

5.2 Quality Control

To ensure participants provided valid data, we accepted only participants that had previously completed 1,000 tasks on MTurk with an overall task approval rate of 95% or higher. Second, we limited participants to the United States and India. This was done because with the first survey coders struggled to understand answers to free

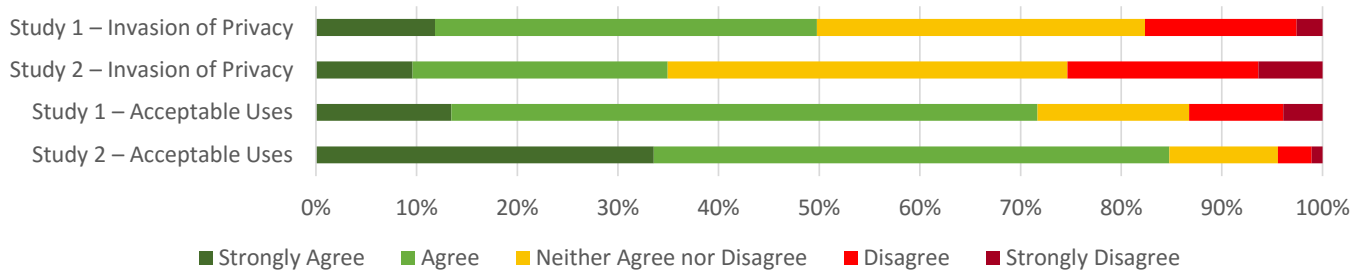


Figure 3: Participant Attitudes Toward TLS Proxies (Survey 1 – N=1,049, Survey 2 – N=927)

	Survey 1 (N=1,049)	Survey 2 (N=927)
Prior Knowledge of TLS Proxies		
Strongly Agree	4.1%	8.4%
Agree	21.3%	27.9%
Neither Agree nor Disagree	8.1%	13.2%
Disagree	48.1%	34.3%
Strongly Disagree	18.4%	16.2%

Table 4: Participants' Knowledge of TLS Proxies

response questions from outside the United States and India.[4] Third, we looked at the single open-ended question to determine if participants had entered spam (e.g., copied an answer from Wikipedia). Finally, we used two validation questions in the second survey because there were not enough open responses to always distinguish spam entries.

In total, we excluded 78 participant's responses (7.8%). The remaining 927 participant's responses constitute the results of our second survey.

5.3 Demographics

The demographics for the participants were summarized earlier in Table 1. There were no significant differences in the demographics of the first and second surveys.

6. SECOND SURVEY – RESULTS

In this section we discuss results from our second survey. First we compare results from the three questions that were the same between both surveys. We then discuss the quantitative data regarding participants' opinions regarding different deployment scenarios for TLS proxies.

6.1 Comparison

In both surveys, after reading the description of TLS proxies, participants were asked whether they had prior knowledge of TLS proxies. These are shown in Table 4. In the first survey, most participants reported having little to no awareness of TLS proxies before the survey: aware (25.4%), unsure (8.1%), unaware (66.5%). In the second survey, more participants reported being aware of proxies beforehand (the difference is statistically significant, $\chi^2[4, N = 1976] = 60.003, p < 0.001$), though over half still reported having little to no awareness of TLS proxies before the survey: unaware (50.5%), unsure (13.2%), aware (36.3%).[5]

We also compared responses relating to whether participants in both surveys felt that TLS proxies were an invasion of privacy, and whether TLS proxies had acceptable uses (see Figure 3). Participants in the second survey were less likely to view TLS proxies as an invasion of privacy (first survey – 50%, second survey – 35%), with the difference being statistically significant ($\chi^2[4, N = 1976] = 54.228, p < 0.001$). Similarly, participants in the second survey were also more likely to feel that there were acceptable uses for TLS proxies (first survey – 72%, second survey – 85%), with this difference also being statistically significant ($\chi^2[4, N = 1976] = 140.654, p < 0.001$).

It is important to note that in both surveys, after participants answered each group of questions (see Appendix) participants were unable to return to earlier groups of questions and alter their answers. As such, the above reported differences are not due to differences in the survey, as up to this point the surveys were identical.

6.2 Scenarios

We asked participants regarding their opinions towards the inspection of encrypted traffic in specific scenarios. For each scenario, participants indicate whether they were comfortable with the traffic being intercepted ("Yes"), whether they wanted to be notified ("Notified"), whether they wanted their consent to be obtained ("Consent"), or whether they were uncomfortable with it. The results for these questions are summarized in Figure 4.

Participants in our second survey are generally willing to accept the use of TLS proxies in most situations, with acceptance ranging from 65% to 90% of participants, when summing together those who accept it, those who desire notification, and those who desire both notification and consent. For both employers (when you use your own computer) and elementary schools, the support for using TLS proxies without notification or consent from users is surprisingly strong (455; 49.1% and 434; 46.8%). This may be due to a belief in employer rights in the first case and a desire to protect children in the second case. In both cases there is still strong support for either notification or

[4]Moreover, these represent a small enough portion of participants that their responses had no significant effect on the data.

[5]As before, we speculate that due to the effects of illusory superiority, the number of participants that were unaware of TLS proxies before the survey was even higher than reported [8, 9].

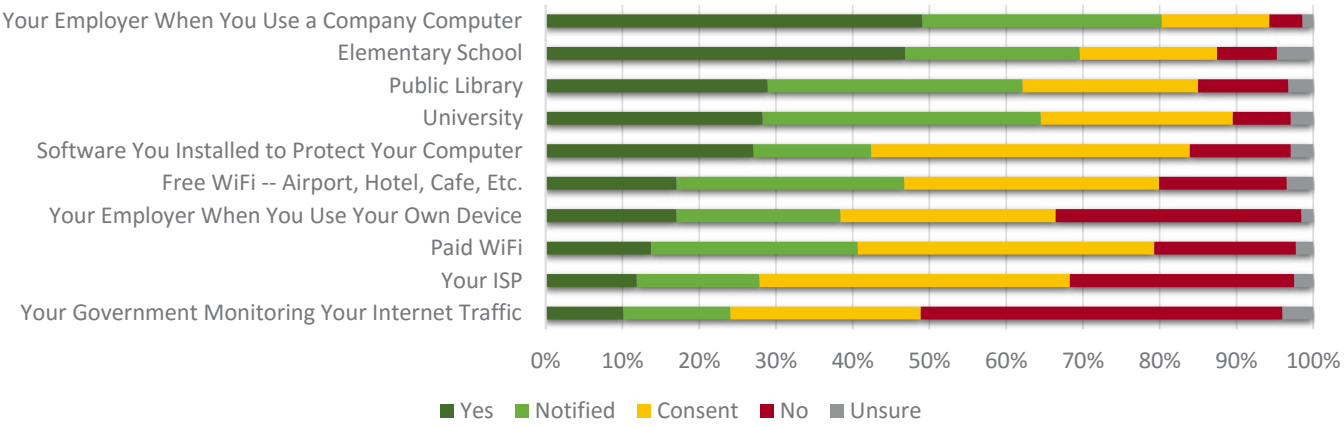


Figure 4: Participant Responses on Scenarios—Should the Organization Be Allowed To Run a TLS Proxy? (N=927)

consent (419; 45.2% and 377; 40.7%).

The strongest objections to any kind of TLS proxy are for government monitoring (437; 47.1%), using your own device at work (297; 32.0%), or using your own ISP (271; 29.2%). Note these latter two map to situations where the user has paid for the device or for network access. Users have stronger objections to TLS proxies when they pay for network access through a home ISP than when they pay for WiFi when they are away from home.

When examining the differences among opinions for notification versus consent, we see that the preference for consent is higher for personal firewalls (software you installed to protect your computer), your ISP, free WiFi, paid WiFi, and using your own device at work. The preference is higher for notification for a public library, university, elementary school, and using a company computer at work. This seems to be a clear split that favors consent in cases where the user feels in control versus notification when an organization is in control. The strongest support for consent is with a personal firewall (385; 41.5%), your ISP (375; 40.5%), and paid WiFi (358; 38.6%).

7. DISCUSSION

In this section we discuss interesting themes we saw as we analyzed participants' responses to the open-ended questions.

7.1 Informed Participants

Most of the participants showed a high level of engagement in the survey. At the end of the survey when asked if they had any additional comments, a large number of participants mentioned that they were thankful that we had informed them of this information. Some even asked where they could get more information on the topic of TLS proxies. Additionally, we were impressed with the in-depth analysis of trade-offs that many users shared, which often went far beyond the scope of any information provided to them in the survey.

Participants clearly understood that there were trade-offs involved with the use of TLS proxies and the inspection of encrypted traffic, weighing the benevolent uses for schools or workplaces and the danger of misuse by insiders or by hackers. As they struggled with this trade-off, participant responses indicated confusion, doubt, worry, equivocation, and reasoned conclusions. Confusion regarding how to resolve the conflict was evident when participants labeled it a "grey area." R988 considered both good and bad uses and worried, *"How are you supposed to know which is happening?"*

Some participants weighed the trade-offs and resolved the dilemma by deciding that proxies should only be used by consent. For example, R827 expressed:

> *"I believe that TLS proxies are an invasion of privacy, as is anything that monitors my internet usage without my permission. However if you are using someone else's (like a company's) network, they have every right to make the rules of use... This is one of those doubled-edged swords – it can be used for your good and security and it can be used to harm and spy on you. Because of the distinct possibility of lost privacy, this type of proxy should [not be] used, except by your agreement, not by anyone else."*

Others wanted companies or schools to be able to use TLS proxies for security purposes, but also wanted to prevent them from being used for government surveillance or by hackers. Still others felt TLS proxies should *only* be used by the government to catch terrorists or criminals.

Similarly, of the participants who were against the use of TLS proxies, the reasons for opposing TLS proxies were not amorphous, but concrete and rational. For example, R666 stated:

> *"I think TLS proxies don't sound very safe because it sounds like an invasion of privacy. I don't think organizations should be able to decrypt your internet traffic and modify it and re-encrypt it. Perhaps they are just trying to protect against viruses and the like but it doesn't sound safe for the person using the internet. What if this technology was misused? Someone could get [h]old of your financial information*

for example. It sounds to[o] risky. I wouldn't want to buy something online and risk someone having access to my credit card number."

7.2 Notification and Consent

Numerous participants expressed a desire for notification and consent when TLS proxies were being used on a network. A typical response as given by R413 was,

> *"Well for some things it would be understandable, I'd just like to be informed so I know the risk I'm taking."*

R313 expressed,

> *"If I encrypt something no one has the right to unencrypt it unless I give them the right to - simple as that."*

Participants expressed extreme distrust for those who would use TLS proxies without informing users, going so far as to say they *"would hate them," "would wonder what they are looking for,"* and *"would assume they were up to no good."*

Others stated they would change their behavior if notified about a proxy, such as avoiding commercial transactions, using a VPN to circumvent a proxy, or self-censorship of their Google searches and other online communication.

7.3 Jaded Participants

We were surprised to find that 4.5% of participants were "jaded" towards the current state of privacy online. They felt that currently it is largely impossible to have any expectation of privacy or security. Many felt that the government was already spying on the population at large, and that even without TLS proxies the government could find a way to gain access to their private information. Others felt that even if they discovered that their traffic was being intercepted, they would have no recourse as their access to the Internet is controlled by a monopoly.

We find this group concerning, as this is not a group of individuals unconcerned with security and privacy. Rather they are a group that still cares about privacy, but has lost all hope that they can actually achieve digital privacy. This is a troubling trend, as such individuals are unlikely to adopt solutions that could actually benefit them. As such, work needs to be done to determine how this type of user's trust can be regained.

7.4 Changing Opinions

Between our two surveys, we noticed differences in the way participants viewed TLS proxies. This demonstrates that users' perceptions towards security and privacy are not static. As such, it is important that work such as this be done on a regular basis, helping the security community stay abreast of current opinions and attitudes.

One interesting difference is that in the second survey fewer participants viewed inspection of encrypted traffic as an invasion of privacy, and more participants felt that there were acceptable uses for this practice. One possible explanation for this difference is that news stories have been discussing how encryption and other privacy preserving technologies could be used by terrorist organizations. Still, additional research is needed to better understand this shift in attitudes towards security and privacy.

8. CONCLUSION

This paper presents the first survey of general (i.e., non-expert) user attitudes toward TLS proxies. Responses indicate that participants hold nuanced opinions on security and privacy trade-offs, with most recognizing legitimate uses for the proxies, but concerned about threats from hackers or government surveillance. A significant concern about malicious uses of TLS inspection is identity theft, and many would react negatively and some would change their behavior if they discovered inspection occurring without their knowledge. We also find that a small but significant number of participants are jaded by the current state of affairs and have lost any expectation of privacy.

User attitudes toward TLS proxies provide an important data point along the spectrum of discussion that is currently taking place regarding who should have access to encrypted information. The results of our survey demonstrate that participants were generally aware of the trade-offs between privacy and security, and that most participants were willing to sacrifice some privacy for additional security. Nevertheless, participants strongly supported notification and consent for when encrypted traffic is being inspected.

9. ACKNOWLEDGMENT

The authors thank Rich Shay for providing feedback on the wording of questions in our first survey. We also thank Alexander Lemon, JJ Lowe, Brent Roberts, and Justin Wu for help with coding the data. Finally, we thank the anonymous reviewers for their helpful comments.

10. REFERENCES

[1] A. I. Antón, J. B. Earp, and J. D. Young. How Internet users' privacy concerns have evolved since 2002. *IEEE Security & Privacy*, 8(1):21–27, 2010.

[2] Blue Coat. Proxysg. `http://www.bluecoat.com/products/proxysg`. Accessed: 9 January, 2014.

[3] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5, 2011.

[4] J. Clark and P. C. van Oorschot. SoK: SSL and HTTPS: Revisiting past challenges and evaluating certificate trust model enhancements. In *IEEE Symposium on Security and Privacy (SP)*, pages 511–525. IEEE, 2013.

[5] X. d. C. de Carnavalet and M. Mannan. Killed by proxy: Analyzing client-end TLS interception software. In *Network and Distributed System Security Symposium (NDSS 2016)*. Internet Society, 2016.

[6] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378, 1971.

[7] E. Galperin, S. Schoen, and P. Eckersley. A post mortem on the iranian diginotar attack. `https://www.eff.org/deeplinks/2011/09/post-mortem-iranian-diginotar-attack`, 2015.

[8] A. M. Glenberg, A. C. Wilkinson, and W. Epstein. The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition*, 10(6):597–602, 1982.

[9] V. Hoorens. Self-favoring biases, self-presentation, and the self-other asymmetry in social comparison. *Journal of Personality*, 63(4):793–817, 1995.

[10] L.-S. Huang, A. Rice, E. Ellingsen, and C. Jackson. Analyzing forged SSL certificates in the wild. In *IEEE Symposium on Security and Privacy*, 2014.

[11] R. Kang, S. Brown, L. Dabbish, and S. Kiesler. Privacy attitudes of Mechanical Turk workers and the US public. In *Symposium on Usable Privacy and Security (SOUPS)*, 2014.

[12] P. G. Kelley. Conducting usable privacy & security studies with Amazon's Mechanical Turk. In *Proceedings of the USER Workshop at the Symposium on Usable Privacy and Security*, 2010.

[13] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 453–456. ACM, 2008.

[14] S. Loreto, J. Mattsson, R. Skog, H. Spaak, G. Gus, and M. Hafeez. Explicit trusted proxy in HTTP/2.0, Internet Draft. `http://tools.ietf.org/html/draft-loreto-httpbis-trusted-proxy20-01`, February 2014.

[15] M. Marlinspike. SSL and the future of authenticity. *Black Hat USA*, 2011.

[16] A. M. McDonald and L. F. Cranor. Americans' attitudes about internet behavioral advertising practices. In *Proceedings of the 9th Annual ACM Workshop on Privacy in the Electronic Society*, WPES '10, pages 63–72, New York, NY, USA, 2010. ACM.

[17] D. McGrew, D. Wing, Y. Nir, and P. Gladstone. TLS proxy server extension, Internet-Draft, TLS Working Group. `http://tools.ietf.org/html/draft-mcgrew-tls-proxy-server-01`, July 2012.

[18] D. Meyer. Nokia: Yes, we decrypt your HTTPS data, but don't worry about it. `http://gigaom.com/2013/01/10/nokia-yes-we-decrypt-your-https-data-but-dont-worry-about-it/`, 2013.

[19] M. O'Neill, S. Ruoti, K. Seamons, and D. Zappala. TLS proxies: Friend or foe? *arXiv preprint arXiv:1407.7146*, 2014.

[20] Palo Alto Networks. Decryption. `https://www.paloaltonetworks.com/products/features/decryption.html`. Accessed: 27 February, 2014.

[21] T. J. Seppala. New Lenovo PCs shipped with factory-installed adware. `http://www.engadget.com/2015/02/19/lenovo-superfish-adware-preinstalled/?ncid=rss_truncated`, 2015.

[22] R. Shay, I. Ion, R. W. Reeder, and S. Consolvo. My religious aunt asked why I was trying to sell her Viagra: experiences with account hijacking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2657–2666. ACM, 2014.

[23] C. Soghoian and S. Stamm. Certified lies: Detecting and defeating government interception attacks against ssl. `http://cryptome.org/ssl-mitm.pdf`.

[24] C. Soghoian and S. Stamm. Certified lies: Detecting and defeating government interception attacks against SSL (short paper). In *Financial Cryptography and Data Security*, pages 250–259. Springer, 2012.

[25] Symantec. Web gateway. `http://www.symantec.com/web-gateway`. Accessed: 9 January, 2014.

[26] B. Ur, P. G. Kelley, S. Komanduri, J. Lee, M. Maass, M. L. Mazurek, T. Passaro, R. Shay, T. Vidas, L. Bauer, et al. How does your password measure up? The effect of strength meters on password creation. In *USENIX Security Symposium*, 2012.

[27] B. Ur, P. G. Leon, L. F. Cranor, R. Shay, and Y. Wang. Smart, useful, scary, creepy: Perceptions of online behavioral advertising. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, SOUPS '12, pages 4:1–4:15, New York, NY, USA, 2012. ACM.

[28] A. F. Westin. Harris-Equifax consumer privacy survey. *Atlanta, GA: Equifax Inc*, 1991.

[29] A. Woodruff, V. Pihur, S. Consolvo, L. Schmidt, L. Brandimarte, and A. Acquisti. Would a privacy fundamentalist sell their DNA for $1000... if nothing bad happened as a result? The Westin categories, behavioral intentions, and consequences. In *Symposium on Usable Privacy and Security (SOUPS)*, 2014.

APPENDIX

A. FIRST SURVEY

A.1 Page 1

We are conducting an academic research survey about public opinions on Internet security. The survey will take approximately 5 minutes.

We will not collect any personally identifying information. If you do not complete the survey we will not store any of your responses. If you have any questions or concerns about the information collected, please contact us at [email redacted].

A.2 Page 2

What is your gender?

- *Male*
- *Female*
- *I prefer not to answer*

What is your age?

- *18 – 24 years old*
- *25 – 34 years old*
- *35 – 44 years old*
- *45 – 54 years old*
- *55 years or older*
- *I prefer not to answer*

What is the highest degree or level of school you have completed?

- *Some school, no high school diploma*
- *High school graduate, diploma or the equivalent (for example: GED)*
- *Some college or university credit, no degree*
- *College or university degree*
- *Post-secondary education*
- *I prefer not to answer*

What is your marital status?

- *Married*
- *Single*
- *Other*
- *I prefer not to answer*

Do you have children?

- *Yes*
- *No*
- *I prefer not to answer*

In which country do you reside?

A.3 Page 3

Where are taking this survey?

- *Home*
- *Work*
- *School*
- *Library*
- *Retail (coffee shop, internet cafe, etc.)*
- *Other*
- *I prefer not to answer*

What type of Internet connection are you using?

- *Wired*
- *WiFi*
- *Cellular (3G, 4G, etc.)*
- *Other*
- *I don't know*
- *I prefer not to answer*

How knowledgeable are you about Internet security?

- *Expert*
- *Highly knowledgeable*
- *Mildly knowledgeable*
- *Somewhat knowledgeable*
- *No Knowledge*
- *I prefer not to answer*

When connecting to a website securely, for example when doing online shopping or banking, who should be able to see the contents of your Internet traffic? (Choose all that apply)

- *Me*
- *My Internet provider*
- *The website*
- *Malicious individuals*
- *Everyone*

A.4 Page 4

When you connect to the Internet you do so through some organization's network. For example, at home you connect to your Internet service provider's (ISP) network, while at work you connect to your employer's network. To protect your information from others on the network you can create secure connections to the websites you use (HTTPS). This is done automatically for you when you log into a website. The secure connection encrypts your Internet traffic so that no one else can view or modify your communication with the website (see Figure A).

Figure A

The network you use to connect to the Internet can also be set up to use a system called a TLS proxy. TLS proxies sit in the middle of your secure connection to the websites you view (see Figure B). At the TLS proxy your Internet traffic is decrypted and the web proxy can view and modify it. Afterwards, the TLS proxy will then re-encrypt your traffic and forward it along. This is done silently and without the

knowledge of you or the website you connect to.

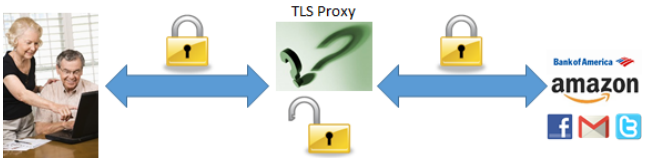


Figure B

TLS proxies can be set up by the organization that controls your Internet (for example, your ISP, school, or employer) and also by malicious attackers. TLS proxies have many different uses:

Protective	**Malicious**
Blocking malware and viruses	Stealing passwords
Protecting company secrets	Identity theft
Blocking harmful websites	Tracking government dissidents
Catching malicious individuals	Spying (for example the NSA)
	Censorship

`Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, Strongly Agree`

- ○ *The above description of TLS proxies helped me to clearly understand what TLS proxies are and how they are used.*

A.5 Page 5

`Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, Strongly Agree`

- ○ *Prior to taking this survey, I was aware that organizations were using TLS proxies.*
- ○ *TLS proxies are an invasion of privacy.*
- ○ *There are acceptable uses for TLS proxies.*

`Only seen if selected "Agree" or "Strongly Agree" to acceptable uses for TLS proxies.`
Please explain which organizations should be allowed to use TLS proxies and for what purpose. (only shown on an Agree or Strongly Agree answer from above)

`Only seen if selected "Disagree" or "Strongly Disagree" to acceptable uses for TLS proxies.`
Please explain why TLS proxies should never be allowed.

A.6 Page 6

`Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, Strongly Agree`

- ○ *I am concerned that TLS proxies could be used by hackers to compromise my Internet security.*
- ○ *I am concerned that TLS proxies could be used by the government to collect my personal information.*
- ○ *Browsers should notify users if there is a TLS proxy intercepting and decrypting their Internet traffic.*
- ○ *There should be legislation that addresses TLS proxies.*

`Only seen if selected "Agree" or "Strongly Agree" to legislation that addresses proxies.`

What should legislation that addresses TLS proxies do? (Choose all that apply)

- ○ *Prevent their use*
- ○ *Require organizations to obtain consent before using a TLS proxy*
- ○ *Require organizations to inform users when a TLS proxy is being used*
- ○ *I don't believe that legislation is required*
- ○ *Other*

A.7 Page 7

The following statements and questions are about how you would personally react to having a TLS proxy on a network you use to connect to the Internet.

`Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, Strongly Agree`

- ○ *I believe TLS proxies are in use on a network I use to connect to the Internet.*

Please explain what concerns you have about a TLS proxy being used on a network you personally use to connect to the Internet.

Please explain how it would change your opinion of an organization if you discovered that they were using a TLS proxy.

If you have any other thoughts, please share them with us below:

B. SECOND SURVEY

B.1 Page One

What is your gender?

- ○ *Male*
- ○ *Female*
- ○ *I prefer not to answer*

What is your age?

- ○ *18 – 24 years old*
- ○ *25 – 34 years old*
- ○ *35 – 44 years old*
- ○ *45 – 54 years old*
- ○ *55 years or older*
- ○ *I prefer not to answer*

What is the highest degree or level of school you have completed?

- ○ *Some school, no high school diploma*
- ○ *High school graduate, diploma or the equivalent (for example: GED)*
- ○ *Some college or university credit, no degree*
- ○ *College or university degree*
- ○ *Post-secondary education*
- ○ *I prefer not to answer*

What is your marital status?

- ○ *Married*
- ○ *Single*
- ○ *Other*
- ○ *I prefer not to answer*

Do you have children?

- ○ *Yes*
- ○ *No*
- ○ *I prefer not to answer*

In which country do you reside?

- ○ *United States*
- ○ *India*
- ○ *Other*

How knowledgeable are you about Internet security?

- ○ *Expert*
- ○ *Highly knowledgeable*
- ○ *Mildly knowledgeable*
- ○ *Somewhat knowledgeable*
- ○ *No Knowledge*
- ○ *I prefer not to answer*

B.2 Page 2

When you connect to the Internet you do so through some organization's network. For example, at home you connect to your Internet service provider's (ISP) network, while at work you connect to your employer's network. To protect your information from others on the network you can create secure connections to the websites you use (HTTPS). This is done automatically for you when you log into a website. The secure connection encrypts your Internet traffic so that no one else can view or modify your communication with the website (see Figure A).

Figure A

The network you use to connect to the Internet can also be set up to use a system called a TLS proxy. TLS proxies sit in the middle of your secure connection to the websites you view (see Figure B). At the TLS proxy your Internet traffic is decrypted and the web proxy can view and modify it. Afterwards, the TLS proxy will then re-encrypt your traffic and forward it along. This is done silently and without the knowledge of you or the website you connect to.

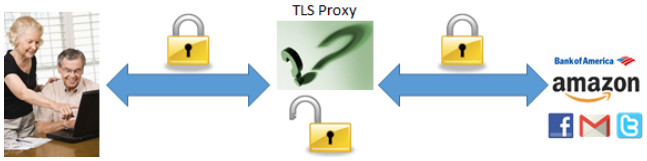


Figure B

TLS proxies can be set up by the organization that controls your Internet (for example, your ISP, school, or employer) and also by malicious attackers. TLS proxies have many different uses:

Protective	**Malicious**
Blocking malware and viruses	Stealing passwords
Protecting company secrets	Identity theft
Blocking harmful websites	Tracking government dissidents
Catching malicious individuals	Spying (for example the NSA)
	Censorship

Ordering of questions randomized.
`Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, Strongly Agree`

- ○ *The above description of TLS proxies helped me to clearly understand what TLS proxies are and how they are used.*
- ○ *Stealing passwords and identity theft are in the list of malicious uses shown above.*
- ○ *Blocking malware and viruses are in the list of malicious uses shown above.*
- ○ *Prior to taking this survey, I was aware that organizations were using TLS proxies.*
- ○ *TLS proxies are an invasion of privacy.*
- ○ *There are acceptable uses for TLS proxies.*

B.3 Page 3

For each scenario listed below, provide your opinion on whether or not the organization should be allowed to run a TLS proxy.

Ordering of questions randomized.
`No, Only if I consent, Only if I am notified (consent not required), Yes (Neither notification nor consent required), Unsure`

- ○ *Your employer when you use a company computer*
- ○ *Your employer when using your own device (cell phone, tablet, laptop)*
- ○ *Elementary school*
- ○ *Public Library*
- ○ *University*
- ○ *Paid WiFi – Airport, Hotel, Cafe, etc.*
- ○ *Free WiFi – Airport, Hotel, Cafe, etc.*
- ○ *The company that provides Internet access at your home*
- ○ *Personal firewall – software that you have installed to protect your computer*
- ○ *Your government monitoring your Internet traffic*

B.4 Page 4

Please feel free to write any thoughts you have on the subject of TLS proxies. We will use this information to help guide future research. (Optional)

Expert and Non-Expert Attitudes towards (Secure) Instant Messaging

Alexander De Luca[1], Sauvik Das[2], Martin Ortlieb[1], Iulia Ion[1], Ben Laurie[1]
[1]Google; [2]Carnegie Mellon University, Pittsburgh, United States
{adeluca,mortlieb,iuliaion,benl}@google.com,sauvik@cmu.edu

ABSTRACT

In this paper, we present results from an online survey with 1,510 participants and an interview study with 31 participants on (secure) mobile instant messaging. Our goal was to uncover how much of a role security and privacy played in people's decisions to use a mobile instant messenger. In the interview study, we recruited a balanced sample of IT-security experts and non-experts, as well as an equal split of users of mobile instant messengers that are advertised as being more secure and/or private (e.g., Threema) than traditional mobile IMs. Our results suggest that peer influence is what primarily drives people to use a particular mobile IM, even for secure/private IMs, and that security and privacy play minor roles.

1. INTRODUCTION

Due to increasing processing power, modern smartphones offer access to manifold services like games, navigation and even office applications. Despite this multi-faceted functionality, communication is still one of the most important reasons why people use smartphones [1, 10] and, as opposed to most other smartphone activities, is in constant use throughout the whole day [1].

Mobile instant messaging (MIM), a highly popular form of communication, is steadily growing with service providers such as WhatsApp[1] broaching more than 800 million active users.[2] With their expansive feature sets, current mobile instant messengers (mobile IMs) have manifold uses, including group chats [22], sharing media files [5], dwelling with friends [13], and even fleeting encounters with strangers [25].

As these applications see more use, the privacy and security problems associated with their use become increasingly important. More and more apps that promise advanced security/privacy over traditional mobile IMs have entered the app market. However, there are, as yet, few insights about how and why users do or do not use these messengers.

To bridge this gap in the literature, we performed two studies – an online survey with 1,510 participants and a set of in-person interviews with 31 participants. For the interviews, we recruited a balanced sample of people from the general public and IT security experts. Furthermore, to better represent arguments both for and against using secure or private mobile messengers, we recruited a balanced sample of people who either used or did not use mobile IMs advertised as secure or private. Our primary goal was to understand the reasons why people use mobile IMs in general, as well as whether and how privacy and security influenced people's decisions to use particular mobile IMs. Furthermore, we also wanted to explore the differences between IT security experts, i.e., people who have the knowledge to make informed privacy and security decisions, and non-experts and whether they behaved differently in their use of mobile IMs (e.g., more or less secure).

The results of our study show that privacy and security play a minor role in people's decisions to use a mobile IM. Security-optimized mobile IMs are not widely adopted and participants who use them have them for a variety of reasons, such as for communicating with a person who is important to them. We also show that while experts are more aware of possible risks, they do not necessarily behave more securely than non-experts.

While some of our work extends existing insights, such as the importance of peer influence on technology adoption, into the context of secure/private IMs, our work offers several novel contributions. For instance, we offer a detailed understanding of why people choose secure IMs and how this process differs between lay users and security experts.

2. RELATED WORK

Privacy and security in mobile instant messaging has been approached from different directions. For this work, we are mainly interested in privacy and security attitudes towards common IMs as well as reasons for migrating to more secure IMs or staying with old, potentially insecure messengers.

In their work, Patil et al. [14] state that IM users (not exclusively on mobile) have three main desires for privacy: privacy from non-contacts, privacy of availability (e.g., their status) and privacy of messaging content. For instance, people are worried that they can be contacted without their explicit consent, something that most current mobile IMs rely

[1]Please note that at the time of the studies reported in this paper, WhatsApp had not yet introduced end-to-end encryption and was encrypted in transit.
[2]Announced by founder Jan Koum on Twitter on April 17, 2015.

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22–24, 2016, Denver, Colorado.

on (e.g., based on the mobile phone number). In follow-up work, the authors further found differences in privacy attitudes towards various categories of contacts [15].

Grinter et al. [9] showed that teenagers' privacy perceptions are centered around data protection. They are worried what happens to their messages after they have been received. This includes how a message is stored, whether it is ephemeral or long-lived and whether it is further shared by the receiver. They often consider possible negative outcomes of such data leaks and thus want their messages safe. This might explain the success of messaging apps promising zero data retention like Snapchat.[3] Technically, these services cannot live up to their promises, creating a potentially problematic false sense of security [17].

Related to this, proving the identity of the communication partner is a difficult task. However, identity is an important factor as users share specific data with specific people but not with others [15]. This has influenced research effort in using behavioral biometrics to ensure that two communication partners are who they claim they are [3].

Interestingly, simple features like the "last seen" indicator in WhatsApp, that are meant to positively support interaction, can be considered problematic by users [2] (despite being bad predictors for actual attentiveness and causing social pressure [16]). For instance, users turn such features off to avoid trouble with their partners.

It was also shown that for convenience reasons, some messengers like WhatsApp and Viber employ practices that might have negative consequences on privacy and security, e.g., when they upload whole address books from the smartphone to enable friend finding [21]. Thus, it is not surprising that many users consider mobile instant messaging to be less secure and less privacy-respectful than SMS[4] [6].

Reasons for using secure mobile IMs or reasons for migrating to them have rarely been explored. The most notable work is by Schreiner et al. [20] who created a model based on the Push-Pull-Mooring migration framework on privacy reasons that would make a user migrate from WhatsApp to Threema. Roughly said, this framework considers pulling factors (privacy advantages of Threema), pushing factors (privacy problems of WhatsApp) and mooring factors (reasons for staying where you are, for example, different costs). They showed that financial costs had no significant effect on the decision. Psychological and emotional switching costs had the strongest impact. In addition, peer influence (i.e., where the users' friends are) was a strong facilitator for switching, something that was identified as an important factor of using a messenger in the first place [6].

A factor not covered in their study is that the bad usability of many available solutions [23] can have a negative influence on user retention – a finding that we also identified among our interview participants.

However, participants in Schreiner et al.'s study [20] were all well-informed before they study. For instance, they received detailed privacy and security information about both messengers and how these worked, thus creating an unrealistic situation. While their study provides many useful insights, we were more interested in decision making processes and current practices based on the actual, unbiased knowledge of the users.

[3]https://www.snapchat.com/ (last access: February 8, 2016)

[4]The actual security of SMS depends on the encryption employed by the service provider.

3. MESSENGER SECURITY

This paper is not meant to provide technical details on messenger security and privacy (see Unger et al. [24] for a comprehensive list of messenger security features). However, it is important to mention two kinds of encryption in order to better understand this work: encryption in transit and end-to-end encryption (e2e).

Most modern IMs use encryption in transit. That means the messages are sent encrypted from the sender to the server and the server to the recipient. On the server, they remain in clear text or in a way that enables at least the service provider to read the information. In many cases, this fact is used to improve service quality and usability.

End-to-end encrypted IMs encrypt the message on the sender's phone and it remains in this state until it is decrypted on the recipient's phone. No third entity has access to the information, not even the service provider. End-to-end encryption comes with some challenges like how to exchange the required keys between the communication entities. For further important security attributes and a list of mobile IMs and their security properties, please refer to the EFF secure messaging scorecard.[5]

The difference between a secure and insecure IM is fuzzy. For this work, we used a rather conservative definition: IMs are secure and/or private if they are actively advertised as secure/encrypted/privacy-preserving. This advertising has to be visible on either the website or the store page without scrolling or clicking any links. In most cases, these were even part of the title, like for Threema which, at the time of the study, was titled "Threema. Seriously secure messaging." All interview participants who used a "secure IM", by the above definition, mentioned that they had seen these labels.

While the promise of security or privacy is no guarantee of actual security or privacy, we consider these promotional messages as the main source of information with which an average person decides whether a messenger is secure and/or private. This assumption is supported by "the paradox of the active user" [4], which states that users never read manuals. Thus, we did not expect that our participants used any more information than the one immediately visible during download to inform themselves. Our results show that there were in fact other sources of information (like knowledgeable peers) that non-experts used but no one mentioned further information from the official websites/manuals.

4. ONLINE SURVEY

The main goal of the online survey was to inform the design of the interview study, to ensure that the interview questions were meaningful and appropriate. In addition, the survey was used to gain first insights into current practices around reasons for choosing mobile IMs and attitudes towards secure mobile IMs.

[5]https://www.eff.org/secure-messaging-scorecard (last access: February 10, 2016)

	US	UK	DE
18-24	19.4%	27.1%	19.3%
25-34	24%	32.2%	33.6%
35-44	20%	30.4%	28%
45-54	21.2%	7.3%	11.3%
55-64	14.4%	2.4%	5.4%
65+	1.2%	0.6%	2.4%

Table 1: Online study participants: Age.

	US	UK	DE
Female	49.2%	51.5%	45.7%
Male	50.8%	48.5%	54.1%

Table 2: Online study participants: Gender.

We used Google Consumer Surveys (GCS) to run the survey.[6] We picked GCS as it is a fast and convenient tool to collect survey responses and was shown to have a user base that is close to the demographic profile of internet users of major research facilities like the Pew research center [12]. In addition, participants on GCS have similar privacy attitudes to other major online sample providers [19].

GCS enables us to target the survey to respondents from the internet or from Android phones only. As we were specifically interested in mobile IMs, we limited respondents to the latter. Android participants are compensated with play store credits, the amount of which is unknown to us.

We ran our survey in three countries, Germany, UK and USA, between March 20 and April 2, 2015. As opposed to the interview study, we did not explicitly recruit for experts or distinguish experts and non-expert users as we were interested in general insights and attitudes.

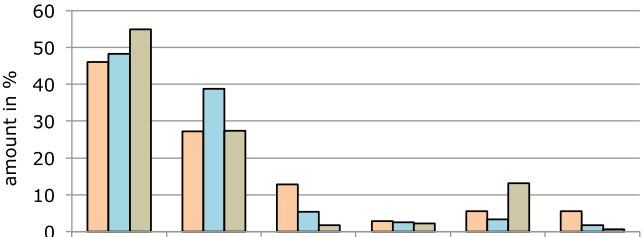
4.1 Survey Design

In general, GCS studies are kept short to avoid click-through answers, i.e., participants who click anything just to receive their incentives. Therefore, we limited the survey to the following four questions. We checked the language in several iterations with different German and English native speakers. We then pre-tested the questions in our lab for language and understanding using the think aloud methodology:

Q1: *"Which of the following mobile instant messengers are you using actively (more than once a week)?"* together with a list of some of the most common secure and standard IMs plus an "other" text field. The order of the answers to this question was randomized with "other" always being shown last.

Q2: *"What is the main reason for your decision to use an instant messenger?"* allowing only one answer (for all options see figure 1). We are aware that this way, prominent answers can mask other options. However, we only meant to collect main reasons and single-response questions work better in GCS, further reducing the chance of dishonest or random answers. In the interviews, we extended these results by exploring all possible reasons instead of focusing on main reasons.

[6] http://www.google.com/insights/consumersurveys/home (last access: February 8, 2016)



	Friends use it	It is free	Pre-installed	Required	Privacy/ Security	Others
US	46.1	27.2	12.8	2.8	5.6	5.6
UK	48.2	38.7	5.3	2.6	3.4	1.8
DE	54.9	27.4	1.8	2.2	13.1	0.6

Figure 1: GCS survey results: main reasons for using IMs.

Q3: *"Please name the mobile instant messenger you are using most frequently."* to identify which of the previously mentioned mobile IMs they are using as their main messenger (same answer options as Q1 and also randomized order).

Q4: *"Have you heard of encrypted or secure mobile instant messaging?"* on a 5-point scale: "I am currently actively using it", "I have tried, but don't use it often", "I have tried, but no longer use it", "I have heard of it, but I am not using it", "I have not heard of it". We also allowed "other" responses. This question mainly served as a baseline to judge whether they are aware of what they are using or not.

4.2 Participants

We set a quota of 500 participants per survey. As GCS slightly over-recruits to make sure you get your quota as fast as possible, we ended up having 1,510 participants (Germany (503), UK (506), USA (501)). The age and gender ratios for the three countries are listed in tables 1 and 2.

4.3 Results

4.3.1 Main Usage Reasons

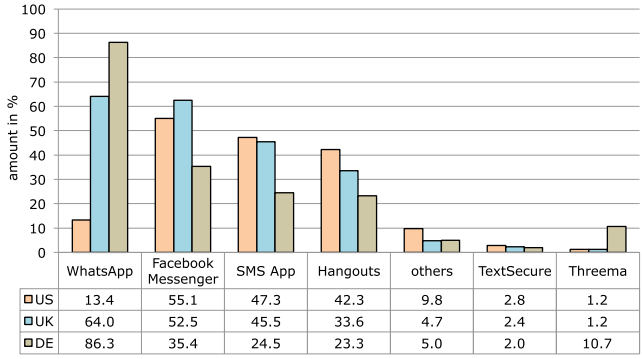
The responses to question 1 can be found in figure 1. The main factor for using a mobile IM in all countries was whether friends were using the messenger, which is in line with results from related work [6, 20]. Whether the messenger was free was another important factor in all countries.

When it comes to privacy and security, only a small fraction of participants stated this being their main factor. An exception is Germany, in which it is the third most important factor with 13.12%.

The "other" reasons include nice integration with the smartphone, being required (for instance by an employer), communication with family members, being associated with other accounts of other apps like Facebook et cetera.

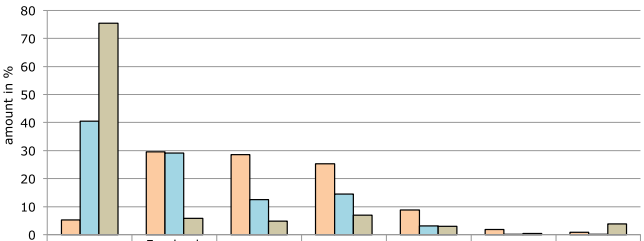
4.3.2 Messenger Use

The picture is similar when looking at the numbers of mobile IMs used by the survey participants as shown in figure 2 and their main IMs (figure 3). Please note that since participants were allowed to mention several messengers, the numbers in figure 2 do not add up to 100%. Also, to allow for comparisons with SMS use, these numbers are included in the figures as well.



	WhatsApp	Facebook Messenger	SMS App	Hangouts	others	TextSecure	Threema
US	13.4	55.1	47.3	42.3	9.8	2.8	1.2
UK	64.0	52.5	45.5	33.6	4.7	2.4	1.2
DE	86.3	35.4	24.5	23.3	5.0	2.0	10.7

Figure 2: IMs used by participants of the GCS survey. Multiple selections possible.



	WhatsApp	Facebook Messenger	SMS App	Hangouts	others	TextSecure	Threema
US	5.2	29.5	28.5	25.4	8.8	1.8	0.8
UK	40.5	29.1	12.5	14.4	3.2	0.2	0.2
DE	75.4	5.8	4.8	7.0	3.0	0.4	3.8

Figure 3: Main IMs of the GCS survey participants. Only one selection per participant.

The three main messengers in all countries are WhatsApp, Facebook Messenger and Hangouts with WhatsApp being less frequently used in the US. The only two applications in the list that are advertised as being secure (and that provide end-to-end encryption) are Threema and TextSecure. In all three countries, use of these messengers is limited with Germany leading the lists with 10.7% of participants using Threema (compared to 1.2% in the US and 1.2% in the UK). Furthermore, 3.8% of participants in Germany used Threema as their main messenger (compared to 0.8% in the US and 0.2% in the UK).

In the "other" group, we identified 9 users of secure mobile IMs (US: 0, UK: 3, DE: 6) who stated to use them as their main messengers. In all instances this was Telegram. No other messengers that are advertised as being secure or private were named.

4.3.3 Security and Messenger Use

Table 3 depicts the frequencies of the five options of question 4. It shows that with the exception of "I have heard of it but I am not using it", the German participants differ in their self-perception of security adoption (data points marked with *).

The association between country and level of secure or encrypted IM knowledge/use is significant ($\chi^2(8) = 115.6656$, $p < .001$). This means that we can reject the null hypothesis that the two variables are independent. For instance, German participants are 2.4 times more likely than UK participants and 2.2 times more likely than US participants to judge themselves as using secure instant messaging.

	not heard	heard but not using	tried but not using	not often	act. using
US	249	174	12	27	28
UK	256	180	8	23	37
DE	138	181	33	64	84

Table 3: Answers to "Have you heard of encrypted or secure mobile instant messaging?" (Q4).

As mentioned before, we also allowed "other" responses to Q4. Participants used this option in 6 instances. For instance, P284 (DE) stated "I would like to use it but too few of my friends do".

Looking further at the data, we identified a discrepancy between participants stating to actively use encryption and the fact that they had no secure or private messenger in their list of actually used messengers. This was the case for almost all participants in the US (35) and UK (31) stating to use secure instant messaging and around half of the German contingent (38). For instance, P459 (DE) mentioned to be actively using secure instant messaging but only used WhatsApp.

Contrary to this finding, some participants who used secure IMs mentioned they would not use secure/encrypted instant messaging. One of them, P407 (DE), uses Threema and stated to not having heard of secure/encrypted instant messaging (again, the Threema logo stated "seriously secure messaging" at the time of this study) and named "friends use it" as the main reason for using an IM.

Overall, there were 21 participants who used Threema or TextSecure and stated to use secure or encrypted messaging (US: 1, UK: 1, DE: 19) and 17 participants who used Threema or TextSecure but stated to not know or not use secure or encrypted messaging (US: 12, UK: 2, DE: 3).

4.4 Takeaways

There are several things we learned from the online survey that influenced the design of the interview study, as we wanted to learn more about these aspects and find out the rationale behind specific decisions:

Security is not the most dominant reason for choosing mobile IMs. Even in the German sample with a higher proportion of secure IM users, the numbers are comparably low with most participants using insecure messengers in addition to secure ones. Due to the survey setup, participants could only provide one answer. In the interviews, we wanted to explore all reasons in detail, not only the main reasons to find out what role security and privacy really play.

The survey showed **discrepancies between what participants assumed and the reality**. For instance, 104 participants stated to use secure/encrypted mobile IMs while in fact the IMs they listed were not. The main question here is whether this discrepancy is a real effect, i.e., users do not know about possible risks, or whether the security provided by their apps is enough for them and if yes, what are the reasons for it.

Our survey supports the finding that **peer influence** is one of the most important factors in the decision making process on which communication tool to use, but also left some

questions unanswered. For instance, we wanted to shed more light on the overall decision making process and the reasons why this factor is so dominant. Accordingly, our interview protocol questions reflect these goals.

Finally, the data we collected is from a general internet population [12]. Inspired by work in other areas, we were interested in whether security experts behave differently from non-expert users, which is sometimes not the case despite their advanced risk knowledge (e.g., [11]).

5. INTERVIEW STUDY

As mentioned before, we used the results of the online survey to inform the design of the interview study. Please note that none of the online study questions was directly reused for the interviews. We rather used the results of the online survey to define where to uncover additional, more granular, insights into why people decide to use certain mobile IMs. We specifically focused on comparing IT security experts with non-experts.

5.1 Study Design

We designed two sets of questions for the semi-structured interviews (one for experts and one for non-experts) based on the open questions of the online surveys. We also added questions, for instance to test for technical knowledge and security/privacy perception of the participants.

While most questions in the two sets were identical, the experts had an additional row of questions that asked them to answer specific questions from the point of view of an "average user". For instance, they were asked both *"What does the term secure instant messenger mean to you?"* as well as *"What do you think the term secure instant messenger means to a user?"*. For this second type of question, the interviewer explicitly told them to answer them from the point of view of a typical end-user.

To avoid influences of specific questions on one another, the order in which these specific questions were asked was counterbalanced (i.e. participants answered them in different orders). We identified four such questions including *"What does the term private instant messenger mean to you?"* and *"What does the term secure instant messenger mean to you?"*.

We had several rounds of language checks and pre-tested the interviews with two participants (one for each set) based on which we created the final questions.

5.2 Procedure

All interviews were conducted in-person by the same interviewer. That is, the interviewer traveled to the countries and locations where the participants lived. At the beginning of each session, each participant read and signed an NDA and consent form. The interviewer then explicitly asked for permission to record audio for the interview. It was explained that the recordings were only used for creating transcripts of the interviews that were needed for the analysis and that they were not shared with anyone outside the research team. We also de-identified recordings to protect participants' privacy. All participants agreed to this procedure.

After that, the interviewer assigned an anonymous ID to each participant and encouraged each interviewee to talk aloud everything that came to their minds. It was also highlighted that they could skip any question they did not feel comfortable answering (this possibility was not used). Participants were not interrupted until they finished answering. After all questions were answered, the participants were debriefed and were given the chance to ask questions themselves. Depending on the replies, the interviews lasted between 30 and 60 minutes.

		ID	Age	M/F	Job
Non-Experts	Normal IM	2	55	m	clerical assistant
		3	38	f	real estate agent
		5	36	m	advisor
		6	23	f	student
		11	39	f	assistant
		12	50	m	clerical assistant
		13	50	f	engineer economics
		14	20	f	student
	Secure IM	1	46	f	secretary
		4	27	m	student
		7	34	f	tanning studio manager
		8	44	m	human resources
		9	51	m	receptionist
		10	44	m	legal advisor
		15	31	f	translator

		ID	Age	M/F	Years in IT Sec.
Experts	Normal IM	1	30	m	7
		2	27	m	4
		3	38	m	7
		4	40	m	3
		5	33	m	8
		6	42	m	1
		7	37	m	16
		8	47	m	30
	Secure IM	9	27	m	4
		10	38	m	8
		11	31	m	12
		12	32	m	6
		13	38	m	15
		14	34	m	20
		15	31	m	10
		16	32	f	6

Table 4: Demographics of the interview study participants.

Participants received a compensation of around EUR 50 for their time, either cash or in the form of a voucher. As the compensation was adapted to the respective country, it slightly varied between participants. Some participants in the IT experts group did not want/take the compensation for different reasons.

5.3 Participants

We recruited 31 interview participants, 15 non-experts and 16 experts. To recruit non-experts from the general public, we worked together with an external recruiting agency providing them with a detailed screener. For instance, we provided a list of mobile IMs that fulfilled our definition of secure or privacy-respectful IMs, in order to get an equal split of secure and non-secure mobile IM users. We also targeted for gender diversity and different professions and education. Non-expert users were recruited in Germany as,

based on the online survey, we considered them more privacy and security aware when it comes to mobile instant messaging. The interviews were also conducted in German and then translated to English for coding.

Recruiting IT security experts was more complex and did not allow for equal gender splits and naturally did not allow for diversity in professions and education as well. For this work, our definition of an IT security expert was someone who had a respective education (e.g., computer science) and was currently working in IT security. We ended up recruiting IT security professionals in several EU countries. Again, we made sure that half of the experts used secure IMs while the other half did not.

Table 4 lists the demographics of all interview study participants.

5.4 Results

In order to analyze the open-ended questions, we used an inductive coding approach. At first, two researcher independently coded the transcribed answers. They then met and discussed discrepancies in their codes to create the final codebook. They then used the final codebook to do the final round of coding for each answer.

5.4.1 Messenger Use

On average, non-experts started using mobile IMs 2.8 years ago (SD=1.7; MIN=0.5; Max=6). Experts started 7.1 years ago (SD=3.3; MIN=0.5; Max=13). All non-experts stated that their first mobile IM was WhatsApp. For experts, the picture is more diverse with 11 different mobile IMs including Skype, TextSecure, iMessage and BlackBerry messenger.

Non-experts stated to have 3.3 messengers that they use more than once a week (SD=1.3; MIN=1; Max=6). Experts use 3.1 mobile IMs (SD=0.9; MIN=1; Max=4). Three non-experts reported their main messenger being a secure IM. With two, this number was even lower for experts. As in the GCS study, the only secure/private messengers named were Threema, TextSecure and Telegram. Note again that secure/private refers to whether they are being advertised as such and not to their actual technical security properties.

Out of the five participants who stated to use secure mobile IMs as their main messenger, three also frequently used other messengers, mainly to stay in contact with specific people. The two (one in each group) who do not use other IMs for this purpose still have fallback strategies in case they want to reach other people, including SMS and email.

5.4.2 Main Usage Reasons

We first asked participants which mobile IM they used as their first ever IM and why they picked that specific messenger. After that, we went through the list of all their currently used IMs and asked them to name the reasons why they used each of them. As opposed to the GCS survey, the interview participants were encouraged to name all reasons.

Table 5 lists the top reasons for messenger use in three categories: reasons for starting to use IMs, reasons for using the IMs being advertised as secure or private, and reasons for using non-secure IMs. The first category allowed us to identify drivers that made participants migrate from other forms of communication to mobile IMs.

	Non-Experts		Experts	
	Reason		**Reason**	
First IM	Everyone uses it	9	Everyone uses it	7
	Free	6	Convenient	6
	Convenient	3	Free	5
	Worldwide use	2	Worldwide use	2
Non-Secure IM	Everyone uses it	9	Specific people use it	11
	Worldwide use	6	Specific functionality	8
	Specific people use it	7	Everyone uses it	7
	Free	5	Groups	7
	Share media	5	For work	7
	Specific functionality	4	Passive use	4
	Convenient	3	Convenient	3
	Groups	3	Integrated	3
	Fast	3	Cross-device	3
	Usability	3	Share media	3
Secure IM	Specific people use it	5	Specific people use it	6
	Distrust in other IMs	3	Security/Privacy	4
	Encryption	2	Encryption	3
	Sharing secrets	2	Audited/Open Source	2
	Security/Privacy	2		

Table 5: Top reasons for mobile IM use, mentioned by the interview study participants, divided into three categories: a) First IM - reasons why they startes using mobile IMs. b) Non-secure IM - reasons they named for the IMs that are not advertised as secure/private. c) Secure IM - reasons for using IMs that are advertised as secure/private.

The data shows that security or privacy were not major considerations in the decision making process when participants started to use mobile IMs. In both groups, the main reason was other people (mainly friends) using the respective messengers and the subsequent desire to stay in contact with them. Furthermore, free conversations (as opposed to SMS), convenience and the ability to be in contact with people worldwide (again without added costs) were major reasons in both groups.

When looking at the results for messengers that are not advertised as secure or private, the main important factors, again, have to do with the participants' peers. In both groups, "everyone uses it" and "specific people use it" are within the top 3 reasons. "Specific people use it" refers to statements like *"Person x does not use my main messenger but I want to stay in contact with this person."*.

This factor is even more prominent when it comes to reasons why people in both groups chose to use secure messengers. Participants accept additional costs (financial and setup/use) even though sometimes it is only for a small group or even one important person. The following quote highlights this: *"My security junkies said they send me sensitive data and don't want to be wiretapped. So if I want to communicate with them, I have to use this messenger [Threema]. It's only around 5 people but I would have done it even for one of them who is an old friend"* (P9, non-expert).

The participants' own privacy and security considerations only play a secondary role in the decision to use a messenger advertised as being secure or private.

Non-Experts		Experts	
Difference		**Difference**	
Functionality	7	Encryption	6
Usability	7	Security/Privacy	6
Security/Privacy	4	Identification/Contacts	6
Technology	4	Technology	5
Costs	3	Costs	3
User base	1	Functionality	3
		Trust	3
		Cross-device	3
		Usability	2
		Availability	2
		User base	2

Table 6: Major mobile IM differences reported by the interview study participants.

5.4.3 Messenger Differences

All participants acknowledged differences between the different messengers they use (see table 6). Non-experts experience strong differences in usability (all non-secure IM users) and functionality. They also repeatedly mentioned that specific functionality worked better in some messengers while they then lacked other features.

Experienced differences in usability had a negative influence on whether participants used secure IMs. Five (2 experts) out of the 16 participants who did not use secure IMs explicitly mentioned that those would be more difficult to use. Usability problems included complex setup phases as well as the lack of a searchable message history.

The expert view on messenger differences was focused on technical and security properties (often related). For instance, the top three differences were: 1. whether the messengers used encryption and if yes, which kind (e.g., end-to-end or in transit); 2. general security and privacy properties (e.g. how and how long data is stored); 3. identification/contacts, i.e., how communication partners were identified and whether this process was protected or not.

5.4.4 Message Sending

To better understand where security and privacy factor into people's rationales for choosing mobile IMs, we also asked participants questions to gauge their understanding of how messengers work (and where security and privacy play a role).

One such question was focused on the mental model participants had about the process of sending mobile instant messages: *"What do you think happens between pressing send and the moment the message arrives on the recipient's phone?"*

All non-experts assumed that the messages would go through an intermediary for several reasons like storage until the message can be sent. In 11 instances, non-experts stated that this would be servers and the remaining four were not sure what the intermediary was but they were sure it existed. Five non-experts assumed (or hoped) that the data transmission would be encrypted in some way.

Seven non-experts thought that their data being read, stored or processed (e.g., for profiling or other analysis) was a normal part of the process that one simply has to accept when "sending messages over the internet". Three explicitly mentioned that this was acceptable as they had nothing to hide. While seven experts mentioned this possibility as well, the difference is that they had a clearer picture why this was done (and sometimes necessary).

In general, experts had a very thorough and technology-focused mental model of the process which was well-informed and based on the fact that they were all educated in this matter. Another major difference was that 11 experts mentioned encryption (or sometimes the lack thereof), and which encryption exactly was used as part of the sending process. They also stated that not using end-to-end encryption enabled advanced features like searching their old messages on a server for specific information. However, this requires a certain amount of trust in the respective service provider.

When asked what a non-expert knows about the sending process, 13 experts stated that they would know little to nothing and if they would think about it, they would most likely assume a direct connection between the two smartphones (8 mentions). Six of them even assumed that normal users would consider it "magic". Furthermore, only one expert thought that normal users would think about whether the communication was encrypted or not. Interestingly, the experts highly underestimated the non-experts' knowledge.

5.4.5 Message Importance

Ten participants in each group stated that they considered it important to keep old instant messages. The main reasons were for non-purposeful lookups, e.g., to re-experience old conversations for emotional reasons. Furthermore, 18 participants (12 experts) look up information, most of which is short-term (5) unimportant information like grocery shopping lists. No participants stated to have information in their instant messages that would be important in the long run.

Three participants in each group considered instant messages not important enough to keep them, not even for emotional reasons. Two participants in the non-expert group considered losing old instant messages to be a cleanup of their mobile device. Seven participants (4 experts) stated that only a few selected messages are important while the majority of them would be expendable.

As opposed to this, most participants considered emails highly important, more or much more important than instant messages (10 experts, 11 non-experts) as in many cases, emails are for non-personal (e.g., business) communication (4 experts, 5 non-experts). Additionally, 7 participants (4 experts) explicitly mentioned that the importance of email was usually long-term or permanent. In 5 cases, the importance of emails and instant messages was either similar or the same, mainly since those participants observed a slight shift from conversational use to business use of instant messages.

Related to the fact that participants consider instant messages short-term information, 5 of them (2 experts) stated that such messages were usually time-sensitive, meaning that they should be read or received as fast as possible. Therefore, 10 participants (6 experts) highlighted that message delivery for them was more important than security and if

	Non-Experts		Experts	
	Statement		**Statement**	
"Secure IM"	Confidentiality	10	Confidentiality	10
	Encryption	5	Encryption	8
	No analysis/profiling	2	Identification	3
	Secure storage	2	Secure storage	3
	Control	1	Control	2
	Authentication	1	Non-existent	2
			Audited	2
"Private IM"	Confidentiality	7	Confidentiality	10
	Encryption	4	Non-existent	5
	Visibility: hidden	3	Encryption	4
	Anonymity	2	Anonymity	4
	No data sharing	2	No leakage	3
	No leakage	2	Marketing	2

Table 7: Main themes reported by the interviewees when asked what they thought the terms "secure instant messenger" and "private instant messenger" meant.

security features would keep their messages from being delivered in a timely fashion, they would consider changing or uninstalling the respective messenger (4 experts, 5 non-experts).

5.4.6 Content Sharing

When asked whether there was specific content that participants would not share over mobile IMs, 26 of them (14 experts) agreed with that statement. In the expert group, all non-secure IM users were among those who agreed. Two of the experts added that this was content that they would not share over any channel.

Examples of sensitive content they would not share include banking information, sexual content or sensitive content that could be used for blackmailing them in case it was leaked. Leakage to people who know the participant was often considered more problematic than leakage to unknown entities as highlighted by the following statement: *"[Leakage to] state agencies is not as bad as they are not interested in what I do and write. People who know me should not be able to get access though."*

In order to avoid such problems and still be able to transmit the information (if necessary), participants named several alternative strategies. These included SMS, telephone, fax, writing letters and even email (PGP or encryption not mentioned) which was mentioned by 7 non-experts and 1 expert.

Out of the remaining 5 participants (2 experts) who answered "no", 3 were secure IM users (2 experts). That means that 2 of the non-experts who would share anything over instant messaging used IMs of which they didn't know whether they were secure or not.

5.4.7 Privacy/Security

One part of each interview focused on the participants' mental models about the terms "secure instant messenger" and "private instant messenger". Table 7 shows the main themes that we identified during coding. For both terms and in both groups, confidentiality was the most important property. This meant that the communication should be protected against any third party but the sender and the recipient. Encryption was another important factor in all groups and for both terms.

A theme that only popped up in the experts group was disbelieve (coded as "non-existent"). For the term "private instant messenger", this was the second most prominent statement. Experts referred to perfect privacy as something technically extremely difficult or even impossible. For instance, control over how recipients handle messages they receive was hard to achieve. Thus, two experts explicitly stated that whenever they read the term, they thought it was simple marketing and they would not trust those promises. This was similar for the term "secure instant messenger", which one expert referred to as "snake oil".

When asked whether these terms influence their impression of an IM (e.g., if the terms are shown as part of the description), 7 experts stated they would check the technical details of the messenger. Furthermore, 6 experts said that messengers need to be audited in order to verify such claims. In the non-expert group, only one participant mentioned audits as a necessary feature. All other participants in the non-expert group would trust the service providers to use the terms correctly or would base their decisions on recommendations by tech-savvy peers they trusted or information they got from the news.

Participants also compared the terms with each other. Eight (2 experts) said the terms referred to the same or highly similar things. Seven participants (2 experts) defined the terms as referring to different parts of the overall process, e.g., *"Security refers to the messages and how they are sent and privacy is the way my data is treated."* (P3, non-expert). The remaining participants either stated that privacy meant more and encapsulated security (2 non-experts, 6 experts) or the other way round (2 non-experts, 8 experts).

6. DISCUSSION

6.1 Peer Influence No. 1 Usage Reason

In both studies, we identified peer influence as the number one reason for choosing and using an IM. In the interview study, this was consistent across both groups, experts and non-experts.

Overall, there were two main types of peer influence. The first has to do with the largest group of people using the same messenger and the subsequent desire to use this messenger as well, irrespective of whether the messenger provides adequate security or privacy. The second type of peer influence was specific (important) people using the service. Often, the groups of people the participants used the respective messenger with was very small, in some cases only a single friend or partner.

This effect works in both directions. If a messenger is not used by a critical mass of contacts (or important people) it will not be used or users will decide to abandon it. For instance, one of the non-expert participants in the interviews mentioned having switched to a secure messenger after a privacy incident with another messenger was reported in the media but then switch back due to peer influence: *"... but it [the privacy incident] had no long-term consequences because [old messenger] is too dominant. At some point, you need to come back."*

Related to literature on social influence in the adoption of technology (e.g., [7, 18]), our participants' did consider their peers' opinions in their decision process. However, the simple desire to stay connected with their friends or specific (important) people was their main consideration which even made them use IMs that they considered inferior to other IMs they had.

6.2 Bad Usability Leads to Abandoning IMs

While usability was only in three instances mentioned as an important factor for choosing an IM, it was mentioned as an important factor in which mobile IMs differed from one another. Specifically secure IMs were repeatedly attributed with having worse usability properties. Consequently, bad usability was a major factor when it came to dropping messengers or not using them at all.

Many of the reported usability problems had to do with security-related properties. For instance, when P15 (non-expert) discussed why he/she had few contacts in his/her main messenger Threema, the explanation was that *"People often don't understand why specific things don't work in Threema"*.

Participants in both groups also repeatedly reported that in order for secure messengers to be better accepted, they should have feature and usability parity with existing major players like WhatsApp.

6.3 Unclear Terms: Privacy vs. Security

The results of the interviews showed that the definitions of the two terms "security" and "privacy" are very fuzzy. This is highlighted by the fact that almost no two definitions given by our participants were identical. Neither the experts nor the non-experts gave identical descriptions.

Furthermore, some participants defined privacy as being a subset of security while others defined security as a subset of privacy. Others, in turn, thought the terms were synonyms or at least highly similar.

Being such highly overloaded terms had two main implications in our study: 1) They have no or only limited effect on experts. They do not trust the terms and require additional (technical) details and information. 2) For non-experts, the fuzziness of the terms had no negative influence as they lack understanding of the technical details anyway. The terms gave them a positive and reassuring feeling.

As a consequence, using both terms in addition to optional details (most likely ignored by non-experts) could be a possible conclusion from these insights.

6.4 (In)Secure Behaviour

Despite experts showing a much higher level of understanding of technical details and possible threats, (voluntarily conducted) insecure behaviour exhibited by the participants in our study was roughly identical across both groups.

For instance, experts are aware of the fact that keeping an (unencrypted) history can be a security/privacy problem but they are willing to accept this for improved service quality like easier backup/recovery and search functionality. In general, participants were mostly happy with the level of security and privacy their messengers provided, even if they did not know the real security properties.

Several interview participants mentioned a trade-off between connectivity and security. Delayed or impossible message delivery due to problems with encryption was unacceptable for them. In such cases, they would prefer unencrypted message transfer in order to keep message delivery timely. They would even go as far as deleting and changing IMs if this remained a problem. In many cases, participants even reported to use "backup messengers" in order to stay connected with certain people, even if they did not trust the security properties of these messengers.

We assume that this behaviour strongly relates to the fact that instant messages are considered short-term information that is only useful for a limited period of time (e.g. shopping list sent by spouse). They are mainly thought of as being of conversational nature and most participants mentioned not sending sensitive information (or information they considered sensitive) through IMs, even with those that are advertised as secure or private.

It has to be noted here that we did not check whether the reported sensitivity of the data participants sent over IMs and the real sensitivity of this data matched. As shown by Egelman et al. [8], self-reported data sensitivity of smartphone data often highly underestimates real risks. Thus, we assume this could be similar in the case of instant messaging.

Our data nevertheless suggests that participants (voluntarily) behave insecurely. This is in line with related work. For instance, Kang et al. [11] found that technical people know more about security risks on the internet but do not spend more effort on protecting their systems.

6.5 Security vs. Reality

When further looking at results related to message importance and security of a user's data, we found that security properties as imagined by the study participants and reality, that is how secure a software really is, often conflicted. This effect was almost exclusively found in the non-experts group.

For instance, 7 non-experts thought that email was a much more secure medium, simply based on the fact that emails contain more important information. For instance, booking information, invoices, bank statements and other (possibly sensitive) information is sent to them through this channel. That is, their rating of email security was based on information independent of actual security.

However, the truth is that most IMs are as technically secure as email, or even more secure. Most mobile IMs (almost all that participants mentioned to use in our study) are in fact at least encrypted in transit, while for email, it depends on both, the provider of the sender and the provider of the recipient. Often, users have no way of knowing whether their emails will be encrypted in transit or not.

This unclarity can lead to users choosing a less secure channel due to confusing cues provided to them in their everyday lives. One way to solve this issue would be to advertise this information better. For instance, even if a messenger was not designed to provide advanced security like end-to-end encryption but provides encryption in transit, this should be highlighted both on the user interface level and in the information available through the service's media channels like its websites.

A current example of more clearly highlighting the security status of a message is the introduction of end-to-end encryption in WhatsApp, which was accompanied by a UI change to display this new property to users[7].

6.6 Sources of Security Information

The security (and to a smaller extent the privacy) properties of mobile IMs were very difficult to judge for the non-experts and multiple participants in the GCS survey. For instance, five interview participants hoped but were unsure that their data was transmitted in an encrypted form.

Non-expert interview participants often referred to security and privacy related terms such as encryption, without actually knowing what they exactly meant. For instance, one user (P4, non-expert) explained encryption in the following way: *"It's when, for instance, my password is changed to those stars [asterisks] so no one can read it."*

Overall, we identified two main forms of sources of information for non-experts: 1) peers like "security junkies" who not only recommend software to them but also help them with other computer related problems like which updates to do and which not; 2) incidents reported in the news.

In the security experts group, the main source of information (other than trusting what they know and checked themselves) were security audits, mainly performed by famous people or organizations they trusted.

7. LIMITATIONS

Using an online survey like GCS does not allow to directly control whether participants correctly fill out the questionnaire or if they just click through it without reading the questions. To avoid such problems, we employed several methods to mitigate the presence of careless answers.

As mentioned in the survey design section, we kept the questionnaire short and the single questions easy to answer (e.g., avoiding multiple choice where possible). GCS also comes with precise timing information about how long it took a participant to answer the whole questionnaire and a specific question. We used this information to eliminate all answers that came to fast to actually read the question. Overall, 4 responses were removed.

Limiting the recipients of the GCS surveys to Android allowed for better control of the fact whether participants were indeed smartphone users or not. On the downside, this meant that we excluded users of other platforms from this sample. For instance, iMessage users are thus not represented in the results.

It also has to be mentioned that both studies were conducted in western democratic countries and thus, the results have to be interpreted with this limitation in mind. For instance, participants living with oppressive regimes or people from specific concerned populations (e.g., journalists working with people who need protection) are likely to respond completely different to our questions.

However, in this study, we were interested in reasoning of the wider general public and thus, decided to recruit for this population rather than the extreme ends of the spectrum. Nonetheless, this is important and we argue that such populations should be investigated in future work.

[7]Again, this feature was not implemented by the time of the studies.

8. CONCLUSION

Our results provide insights into the decision making process of whether, how and why people choose and use certain mobile IMs. Most importantly, despite security and privacy playing a role in the decision making process for some people, they were only seldom the primary factor, while peer influence, i.e., who and how many people use the IM, was identified as the most important factor. We also found that while, not surprisingly, experts had advanced knowledge about possible privacy and security risks related to using mobile IMs, their behaviour did not notably differ from how non-experts used mobile IMs.

The main factor pulling people away from using secure IMs in our study was usability. That is, if a secure IM does not provide the features desired by the users or if it is more difficult to use than a common IM, it drives people away from it. Both, the general usability and the feature set provided by an IM need to be comparable to major players in order to avoid this effect.

Some study participants, specifically the ones who reported to sometimes use IMs for work, noted a trend that in their opinion, IM use slightly shifted to becoming a replacement for email and other communication channels. Based on the fact that some participants (mainly non-experts) thought of email as a secure communication channel just because important information was sent through it, this leads to the question of whether attitudes towards mobile IM will change once it is more integrated into our everyday work lives.

Future work should focus on groups who already made this transition and find out whether privacy and security requirements are different for those groups. Another highly important group for future research is at-risk users. That is, populations who for different reasons require enhanced security and privacy in their communication (like journalists who work in risky parts of the world). Their attitudes towards communicating with mobile IMs are likely to be quite different from the ones of the general population.

9. REFERENCES

[1] M. Böhmer, B. Hecht, J. Schöning, A. Krüger, and G. Bauer. Falling asleep with angry birds, facebook and kindle: A large scale study on mobile application usage. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '11, pages 47–56, New York, NY, USA, 2011. ACM.

[2] A. Buchenscheit, B. Könings, A. Neubert, F. Schaub, M. Schneider, and F. Kargl. Privacy implications of presence sharing in mobile messaging applications. In *Proceedings of the 13th International Conference on Mobile and Ubiquitous Multimedia*, MUM '14, pages 20–29, New York, NY, USA, 2014. ACM.

[3] U. Burgbacher and K. Hinrichs. An implicit author verification system for text messages based on gesture typing biometrics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 2951–2954, New York, NY, USA, 2014. ACM.

[4] J. M. Carroll and M. B. Rosson. *Paradox of the active user*. The MIT Press, 1987.

[5] Y.-Y. Chen, F. Bentley, C. Holz, and C. Xu. Sharing (and discussing) the moment: The conversations that occur around shared mobile media. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '15, pages 264–273, New York, NY, USA, 2015. ACM.

[6] K. Church and R. de Oliveira. What's up with whatsapp?: Comparing mobile instant messaging behaviors with traditional sms. In *Proceedings of the 15th International Conference on Human-computer Interaction with Mobile Devices and Services*, MobileHCI '13, pages 352–361, New York, NY, USA, 2013. ACM.

[7] R. B. Cialdini. *Influence: Science and practice*, volume 4. Pearson Education Boston, 2009.

[8] S. Egelman, S. Jain, R. S. Portnoff, K. Liao, S. Consolvo, and D. Wagner. Are you ready to lock? In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 750–761, New York, NY, USA, 2014. ACM.

[9] R. E. Grinter and L. Palen. Instant messaging in teen life. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, CSCW '02, pages 21–30, New York, NY, USA, 2002. ACM.

[10] A. Hang, E. von Zezschwitz, A. De Luca, and H. Hussmann. Too much information!: User attitudes towards smartphone sharing. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*, NordiCHI '12, pages 284–287, New York, NY, USA, 2012. ACM.

[11] R. Kang, L. Dabbish, N. Fruchter, and S. Kiesler. "my data just goes everywhere:" user mental models of the internet and implications for privacy and security. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 39–52, Ottawa, July 2015. USENIX Association.

[12] S. Keeter and L. Christian. A comparison of results from surveys by the pew research center and google consumer surveys, 2012.

[13] K. P. O'Hara, M. Massimi, R. Harper, S. Rubens, and J. Morris. Everyday dwelling with whatsapp. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, pages 1131–1143, New York, NY, USA, 2014. ACM.

[14] S. Patil and A. Kobsa. Instant messaging and privacy. In *Proceedings of HCI*, pages 85–88, 2004.

[15] S. Patil and A. Kobsa. Uncovering privacy attitudes and practices in instant messaging. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work*, GROUP '05, pages 109–112, New York, NY, USA, 2005. ACM.

[16] M. Pielot, R. de Oliveira, H. Kwak, and N. Oliver. Didn't you see my message?: Predicting attentiveness to mobile instant messages. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 3319–3328, New York, NY, USA, 2014. ACM.

[17] N. A. Poltash. Snapchat and sexting: A snapshot of baring your bare essentials. *Rich. JL & Tech.*, 19:1, 2012.

[18] E. M. Rogers. *Diffusion of innovations*. Simon and Schuster, 2010.

[19] S. Schnorf, A. Sedley, M. Ortlieb, and A. Woodruff. A comparison of six sample providers regarding online privacy benchmarks. In *SOUPS Workshop on Privacy Personas and Segmentation*, 2014.

[20] M. Schreiner and T. Hess. Examining the role of privacy in virtual migration: The case of whatsapp and threema. In *Proceedings of the 21st Americas Conference on Information Systems*, AMCIS '15, 2015.

[21] S. Schrittwieser, P. Frühwirt, P. Kieseberg, M. Leithner, M. Mulazzani, M. Huber, and E. R. Weippl. Guess who's texting you? evaluating the security of smartphone messaging applications. In *NDSS*, 2012.

[22] M. E. Smith and J. C. Tang. "they're blowing up my phone": Group messaging practices among adolescents. 2015.

[23] R. Stedman, K. Yoshida, and I. Goldberg. A user study of off-the-record messaging. In *Proceedings of the 4th Symposium on Usable Privacy and Security*, SOUPS '08, pages 95–104, New York, NY, USA, 2008. ACM.

[24] N. Unger, S. Dechand, J. Bonneau, S. Fahl, H. Perl, I. Goldberg, and M. Smith. Sok: Secure messaging. In *2015 IEEE Symposium on Security and Privacy*, pages 232–249, May 2015.

[25] Y. Wang, Y. Li, and J. Tang. Dwelling and fleeting encounters: Exploring why people use wechat - a mobile instant messenger. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15, pages 1543–1548, New York, NY, USA, 2015. ACM.

Snooping on Mobile Phones: Prevalence and Trends

Diogo Marques,[1] Ildar Muslukhov,[2] Tiago Guerreiro,[1] Konstantin Beznosov[2] and Luís Carriço[1]

[1] LaSIGE, Faculdade de Ciências
Universidade de Lisboa, Lisbon, Portugal
[dmarques, tjvg, lmc]@di.fc.ul.pt

[2] Department of Electrical and Computer Engineering
University of British Columbia, Vancouver, Canada
[ildarm, beznosov]@ece.ubc.ca

ABSTRACT

Personal mobile devices keep private information which people other than the owner may try to access. Thus far, it has been unclear how common it is for people to snoop on one another's devices. Through an anonymity-preserving survey experiment, we quantify the pervasiveness of *snooping attacks*, defined as "looking through someone else's phone without their permission." We estimated the 1-year prevalence to be 31% in an online participant pool. Weighted to the U.S. population, the data indicates that 1 in 5 adults snooped on at least one other person's phone, just in the year before the survey was conducted. We found snooping attacks to be especially prevalent among young people, and among those who are themselves smartphone users. In a follow-up study, we found that, among smartphone users, depth of adoption, like age, also predicts the probability of engaging in snooping attacks. In particular, the more people use their devices for personal purposes, the more likely they are to snoop on others, possibly because they become aware of the sensitive information that is kept, and how to access it. These findings suggest that, all else remaining equal, the prevalence of snooping attacks may grow, as more people adopt smartphones, and motivate further effort into improving defenses.

1. INTRODUCTION

Mobile phones are not just phones anymore, they are interfaces to much of users' social lives, and keep records which, in all likelihood, include intimate, sensitive, or confidential information. As long as those records are interesting to anyone, there is a risk that they might try to obtain them.

The speed and extent to which mobile devices are being adopted has created new opportunities for remote, sophisticated adversaries. Phenomena like mobile malware, surveillance by state-sponsored actors, and personal data tracking for commercial purposes, have entered into public discourse, and became, reasonably so, a point of concern [38]. However, in their daily lives, users face a more immediate threat: people with whom they have close social ties can infringe on their privacy just by picking up their devices and browsing through their data. Those social *insiders* [29] can act opportunistically, without having any special skills or abilities. Such may happen when devices are left unattended, or handed over with the expectation of limited use. Often, social insiders can achieve their objectives just by undertaking what we will refer to as a *snooping attack*, that is, by looking at information that was not intended for them, without a primary intent to extract data or make changes. If we conceive of privacy as the ability to have control over the ways others know us [33], being snooped on by people whose opinion we care about is a violation of privacy in its most fundamental sense.

There are technological defenses against snooping attacks, most notably authentication mechanisms. However, it has become clear that people very often do not use them [12, 17, 23]. While there is debate over why people make such a choice, and over if and how they could be encouraged to choose differently, users remain in a situation where there are more opportunities for snooping than there could otherwise be. More opportunities, however, do not necessarily translate into more actual offenses. This uncertainty about whether people's phones are commonly, or only rarely, being snooped on, casts doubt over the importance and/or urgency of securing their devices against third parties that are, at first sight, trusted.

In this paper, we bring new evidence into this conversation, by measuring actual successes in conducting snooping attacks, from the attacker's perspective. From a security standpoint, it is of special importance to know how successful snooping attacks are, because high degrees of success indicate that existing defenses, both behavioral, like keeping the device on oneself at all times, and technological, like device locking, are inadequate. We thus aimed to measure the proportion of people, in population with a large degree of mobile device adoption, that successfully snooped on someone else's device, and to explore the pervasiveness of the phenomenon, or lack thereof, across population groups. We selected the U.S. adult population as a target, because it is easily accessible and well characterized in terms of mobile device adoption.

The main challenge with obtaining such data is methodological. If we were to field a survey asking people whether they had snooped on someone else's device, we could not reasonably expect honest responses, because such behavior is commonly deemed to be censurable. Thus, we employed the list experiment (e.g., [27]), a technique in which participants are asked to look at a list of items, and indicate how many (not which) they identify with. In list experiments, one group of participants receives a list of control items, and another group a list of the same control items plus an item of interest. An aggregate estimate of positive response to the item of interest can be calculated by the difference between groups, without knowing the true answer for each respondent. A more detailed description of the technique, and the rationale for its selection over other techniques, is provided in Section 3.

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22–24, 2016, Denver, Colorado.

List experiments are understood to provide less biased estimates of response to sensitive questions, in comparison with direct self-reporting, but require careful design. The way in which list questions are worded, and the way in which surveys are administered, can have significant impact on measurement error [27]. We conducted two empirical studies to address these issues.

In a first study, conducted with Google Consumer Surveys (GCS), we selected the control and sensitive items to include in the list. For the control items, we measured, with direct questioning, the prevalence of previously reported behaviors that relate to privacy and security. Based on 1,140 responses, we selected a mix of items that prevents ceiling and floor effects. For the sensitive item, special consideration was given to how it was framed, because the specific wording would be the operational definition of the construct that we wanted to measure – in this case, successful snooping attacks. We tested 4 alternative ways of wording the concept such that it was easy to understand and mapped to the security issue at stake. Based on 1,086 responses, we concluded that the most adequate wording, among the alternatives, was "looked through someone else's phone without their permission". This study is reported in Section 4.

A second methodological challenge arose from a decision, made at the outset, to field the survey in Amazon Mechanical Turk (MTurk). MTurk is commonly used to target large participant pools [34], but doubts have been raised about its appropriateness for survey research [14], since participants, and especially those with low reputation, may engage in satisficing [32, 35]. To validate that list experiments on MTurk produce reliable measurements, we ran a list experiment with one control group and two treatment groups (n = 434), who received extra items with known prevalence of ~100% (having opened eyes in the morning) and ~0% (having travelled in interplanetary space). We were then able to compare the known prevalence to the one estimated by the list experiment, across 3 groups of MTurk participants, with distinct reputation levels. We concluded that list experiments appropriately estimated expected proportions, without the need to control for participant reputation. This finding, which is reported in Section 5, is a secondary generalizable contribution of this work.

Taking these findings into consideration, and making conservative design choices, we deployed a list experiment to MTurk to measure the prevalence of snooping attacks (n = 1,381). In Section 6, we describe the final design, the data collection process, and report on the proportion of people who, in 1 year, successfully engaged in snooping attacks on others' mobile phones, offering both a point estimate of prevalence, and predictors of such behavior. We provide estimates for the MTurk sample, which is often taken as being representative of the Internet population, and further project it into the U.S. adult population, by post-stratification weighting. The main findings are as follows:

- An estimated 31% of participants had "looked through someone else's phone without permission," in the 12-month period before the survey was conducted.
- Adjusting the younger and more male MTurk sample to the U.S. adult population, the 1-year prevalence was estimated at 20%.
- Engaging in snooping attacks does not seem to be strongly related to gender, level of education, or geographical region.
- Younger participants were notably more likely to have engaged in snooping attacks, to the extent that the behaviour was estimated to be prevalent (52%) among those between 18 and 24 years of age.
- Those who own smartphones are much more likely to snoop on others.

Although this study could not establish mechanisms by which the observed trends emerged, the fact that the youngest participants and those who used smartphones were more likely to snoop on others suggested a common cause. It has been noted that smartphone users often engage in a pattern of adoption in which the phone mediates important aspects of their private social life [9, 39]. In a follow-up study (n = 653), with a similar design to the previous, we examined whether, among smartphone users, depth of adoption predicted the prevalence of snooping attacks. We confirmed that the more people use their smartphones in ways that generate privacy-sensitive data, the more likely they are to snoop on others, even when controlling of age. A compelling explanation for these findings is that, as people learn by their own usage what kinds of sensitive information is kept on smartphones, they gain a better sense of what they could have access to if they were to snoop. This final study is reported in Section 7.

Overall, these results indicate that snooping on other people's devices must be relatively easy, to be so common. Furthermore, the population trends that we found suggest possible growth of the phenomena. If it is the youngest, and those who adopt smartphones to a larger extent, that are more likely to snoop on others, then growth may come from aging of the cohort, or from more people adopting smartphones in ways that make them aware of the private data that is kept. The situation calls for additional efforts in providing adequate defenses against socially-close adversaries, and for a re-examination of assumptions of trust in mobile security threat models.

2. RELATED WORK

It has been widely documented that smartphones are used very differently than either regular phones or computers, and, as a result, store a great deal of sensitive information, including access codes, personal communication, call and text logs, contacts, pictures, videos, and location records (e.g., [2, 13, 28]). Users have been found to be concerned about the risks to their privacy that have therefore emerged [9, 36]. Events have not proved them wrong.

In the last few years, there has been much discussion about phenomena like mobile malware, government surveillance, and personal data gathering for commercial purposes (e.g., [13, 37, 38]). Threats such as these, in which adversaries are technologically sophisticated, and act remotely, have traditionally been seen as the potentially most damaging. However, end-users are very rarely affected in a practical sense, and, when they are, the impact on their lives has been somewhat limited, mainly taking the form of unsolicited advertising [13].

Recently, as spearfishing and insider threats have gained more attention in the computer security community, so have socially-close adversaries been recognized as a threat to personal mobile computing [29]. Younger users, the so-called digital natives, are indeed more concerned about insiders: they are more aware of threats with a social context (like those arising from loss, theft, snooping or shoulder-surfing) than of threats with a technical connotation (like those arising from malware or network attacks) [23].

In a recent Pew survey [36], 12% of US mobile phone owners reported having had another person access the contents of their phone

in a way that made them feel their privacy was invaded. This statistic can be seen as an indirect measure of snooping attack success, but one that is likely inaccurate. For instance, many people may have had their smartphones snooped on but not know about it. Conversely, the fact that someone felt that their privacy was invaded does not mean that there was an explicit intention by the person accessing the device.

Corroborating that finding, in a recent survey with an MTurk sample, 14% of participants reported being targets of snooping ("Someone used my mobile phone without my permission with intention to look at some of my data"), and 9% reported being attackers ("I used someone's mobile phone without owner's permission to look into his/her data") [29]. This is, as far as we know, the first measurement of successful snooping attacks from the attacker's perspective. This measurement, however, is not generalizable, for two reasons. First, because it was meant to be a sample summary, not a population estimate, as part of a study with a broader objective. Second, because the questions were asked directly, and thus the number of people willing to identify with behaviour that can be seen as offensive is expected to be biased by the social desirability effect, as suggested by a 6 percentage point mismatch between reported targets and attackers.

We aimed to measure how often people actually succeed in conducting snooping attacks, taking into consideration that they might not be willing to admit it. Furthermore, we were interested in an estimate bounded in time, namely one year, to allow periodical comparisons. By measuring 1-year prevalence periodically, it is possible to discern any changes, which could, for instance, indicate adoption of new defenses. In contrast, if participants are periodically asked if they *ever* snooped, changes might not be observable until there is a sufficiently large proportion of new entrants into the population.

Comparing our results to previous statistics, the problem does seem to have been underestimated. We found that 20% of U.S. adults engaged and succeeded in snooping attacks in a year, while only 12%, over their lifetime, report having had the contents of their devices accessed [36]; and, for a comparable MTurk population, using the list experiment procedure, we estimated 1-year prevalence of snooping attacks (31%) to be approximately 3 times as high as the previous lifetime prevalence estimate obtained with direct questioning (9%) [29]. Unless there was a very large upward shift in prevalence that would explain these differences, it seems that indeed many people never come to learn that they were snooped on, and that when asked directly, people who have snooped on others often do not admit to it.

3. ASKING SENSITIVE QUESTIONS

Studies of attitudes, opinions and behaviors run into measurement error whenever self-reports can not be trusted. One classic example is that men consistently report having had a far greater number of sexual intercourse partners than women, which, if true, would defy logic [42].

One source of measurement error is social desirability bias [41]. When questions are sensitive, respondents tend to give answers that they understand to be the right ones, and not necessarily the truth. Questions that pertain to protecting one's privacy are known to be subject to that bias. It has been shown that the mere addition of privacy wording in surveys makes respondents much more likely to give socially desirable responses [6].

Indirect survey techniques to reduce social desirability bias have emerged in the last few decades. Their main principle is assurance of response confidentiality by design, not policy. Respondents have strict guarantees that their individual answer will not be revealed, and are therefore more likely to answer truthfully. The cost to researchers is that they will not know the response of each individual, only aggregate estimates.

Two main types of such survey instruments have received attention. One is the *randomized response technique* (RRT) [5]. In its simplest form, respondents are shown a sensitive question and asked to privately flip a coin. If it lands on one side, participants must answer "yes", regardless of truthfulness, and if it lands on the other side, they must answer truthfully, "yes" or "no". Each individual respondent is thus assured that answering "yes" does not reveal their true response, as long as no one else knows on which side the coin landed. But knowing that the probability of a coin landing heads or tails is equal, the total proportion of positive responses can be calculated by assuming that half the positive responses are a consequence of the coin toss, and the remaining are truthful.

The other technique is the *list experiment* (sometimes called unmatched count technique, or item count technique, or unmatched block design), which we have employed. List experiments are a kind of survey experiment [30], which involve dividing a sample into two groups, the control and the treatment. As an example, in a recent study [40], where researchers addressed the puzzle of why a particular ballot initiative failed to pass when opinion polls indicated otherwise, the control group was asked the following question:

Here is a list of four things that some people have done and some people have not. [. . .] Do not tell me which you have and have not done. Just tell me how many:
- Discussed politics with family or friends;
- Cast a ballot for Governor Phil Bryant;
- Paid dues to a union;
- Given money to a Tea Party candidate or organization.
How many of these things have you done in the past two years?

The treatment group saw the question with the following extra item:

- Voted 'YES' on the 'Personhood' Initiative on the November 2011 Mississippi General Election ballot

With this technique, participants do not have to reveal their truthful answer to the extra item, which is the one actually being measured. Yet, the proportion can be estimated by comparing the mean number of items selected by respondents in control and treatment groups. All the rest being equal, a difference in means can be attributed to the presence of the extra item. The difference in means is thus the estimate of proportion of positive responses to the sensitive item.

It has been shown that both the list experiment and the RRT reduce response bias. In the mentioned validation study [40], which tested both approaches, it was found that an RRT survey predicted almost exactly the actual vote. A list experiment survey considerably reduced the bias, but still underestimated the actual vote share.

For online surveys, however, application of the RRT is problematic. Since the procedure is complex, respondents have to expend considerable time to understand it, and often they have trouble believing their true answers are not revealed [11]. As we intended to deploy the survey on MTurk, where participant attention is already scarce (e.g., [35]), and extra time is costly, we opted for a list experiment. Even if list experiments provided estimates that were overly conservative, on the issue of snooping, it was best to err on the side

of caution. If even a conservative estimate was relevant, than surely a higher estimate would have at least the same consequence.

The list experiment procedure seldom appears in HCI research (with one exception that we know of [1]). With this paper we also wanted to call attention to the growing tool belt of survey research methods for sensitive topics, which can help untangle the often found discrepancy between self-reports and actual behaviour in privacy-related studies.

4. STUDY 1: ITEM SELECTION

List experiments aim to reduce the measurement error that would occur if sensitive questions were asked directly. For them to be effective, careful consideration has to be given to the composition of the list. The perception of confidentiality can be jeopardized when lists are not credible, or when truthful answers would reveal that respondents had answered positively to the sensitive item. With this first empirical study, we aimed to compose a list of items that would minimize the chances of obtaining unreliable measurements from a full-scale survey experiment.

The danger of unreliable measurement can be mitigated by following common advice on designing list experiments (e.g., [4, 11, 15, 27]), which includes:

1. **Avoid ceiling effects** A ceiling effect happens when all the control items are so common that many participants would, if answering truthfully, identify with all items, thus revealing their positive answer to the sensitive one.
2. **Avoid floor effects** A floor effect occurs when the control items are so uncommon that, for many participants, the only item they could credibly report as identifying with would be the sensitive one.
3. **Avoid lists that are too short** Short lists increase the likelihood of a ceiling or floor effect.
4. **Avoid lists that are too long** Long lists increase variance and demand more attention from participants.
5. **Avoid contrast effects** If the sensitive item is too salient, respondents might worry that any non-zero answer to the list is indicative of identification with it. The list should therefore include control items that are on the same topic as the sensitive item, which itself should be worded in neutral language.

Taking this advice into account, we decided to run surveys on individual behaviors to obtain prevalence estimates, so we could select a combination of control items, and a wording for the item pertaining to snooping attacks, that would make confidentiality plausible.

4.1 Procedure

To build the list of items, we ran direct question surveys on several candidate items using Google Consumer Surveys (GCS).

For each candidate control item, we aimed at a target sample of 100 participants. For candidate sensitive items, we targeted a sample of 250 participants, as we expected lower sensitivity, due to social desirability bias. The actual number of participants is often different than the target, because of the particular way in which GCS samples [26].

For the control items, to avoid contrast effects with the sensitive item, we selected candidates among previously documented behaviors or situations related to mobile privacy [13] and online privacy [38], shown in Table 1, rows 1 to 8.

For the sensitive item, that pertains to snooping attacks, we tested four ways of wording the behavior, shown in Table 1, rows 9 to 12. The formulations avoid the word "snooping", which we deemed to have a too-negative connotation, and instead test a maliciousness dimension, with "used" vs. "looked through" wording, and an egregiousness dimension, with "without knowledge" vs. "without permission" wording.

4.2 Results

4.2.1 Control item selection

Our surveys did not find privacy-relevant behaviors or situations that can be said to be of high prevalence, but items of low prevalence were abundant. In part, such could be explained by the existence of social desirability bias for some of the controls.

Nevertheless, taking the measured prevalences for candidate items as indicative of true differences in the population, results indicated it would be trivial to avoid ceiling effects (advice 1) even with a short list, by selecting among the items with very low prevalence.

Avoiding floor effects (advice 2) was more challenging, as we did not find highly prevalent items. We decided to include 4 control items in the final list, at the cost of possible lower precision in estimates (advice 4). With 4 control items rather than 2 or 3, there were, we reasoned, enough guarantees of confidentiality. Even if respondents answered "1" it would be plausible enough that they were referring to one of the controls that is not abundantly privacy-sensitive, such as receiving spam.

We finally selected the items from surveys 1, 2, 4 and 5, which are the ones with the highest and lowest prevalence, that still pertain to mobile security, and thus generate less contrast (advice 5) with the sensitive item.

4.2.2 Sensitive item selection

For the item conveying the "snooping attack" construct, the surveys we conducted did not show any appreciable differences as a result of different wording. A Chi-squared test did not provide evidence that the wording had an overall effect on the rate of positive answers ($\chi^2(3) = 5.36$, $p = 0.1471$, Cramer's V = 0.07), nor that wording conveying either egregiousness or maliciousness had significant effects in isolation ($\chi^2(1) = 2.610$, $p = 0.1062$, Cramer's V = 0.05, and $\chi^2(1) = 1.192$, $p = 0.2749$, Cramer's V = 0.04, respectively). In a logistic regression model of positive or negative answer as a function of egregiousness or maliciousness wording, we also did not find either factor to be a significant predictor at the 0.05 significance level, and the model accounted for very little of the deviance (null deviance 751 on 1085 d.f. vs. residual deviance 746 on 1083 d.f.).

We could have expanded the sample to get more precise estimates and possibly establish minute differences between wording choices, but given the observed effect sizes, and the likelihood that social desirability bias was already introducing measurement error, any differences, even if statically significant, were unlikely to be of practical importance. We thus concluded that, for the purpose of our main survey, we should use the wording that, on its face, represented an egregious violation of an access policy with malicious intent: having *looked through* someone else's cell phone without their *permission*.

4.3 Discussion

Based on the results of direct question surveys, we composed a list of items that included a mix of controls which were low to medium

Table 1: Results of single question surveys conducted in Google Consumer Surveys: 1 to 8 for candidate control items for the list experiment question (1-5 behaviors related to mobile security, 6-8 behaviors related to online privacy), 9 to 12 for different ways to word the item conveying the concept of snooping attacks. Items selected for the list in bold.

Survey	Question	Yes	No	Participants
1	**In the past 12 months, have you purposefully made phone calls or sent text messages to 1-900 numbers from your cell?**	6%	94%	183
2	**In the past 12 months, have you lost or had your cell phone stolen from you?**	11%	89%	191
3	In the past 12 months, have you sent a text message to the wrong person by mistake?	17%	83%	155
4	**In the past 12 months, have you shared pictures from your cell phone on social media (for instance, Facebook or Twitter)?**	27%	73%	108
5	**In the past 12 months, have you received at least one text message/IM with unsolicited advertising (spam) on your cell?**	42%	58%	173
6	In the past 12 months, have you been asked to create a new password for an online service?	37%	63%	110
7	In the past 12 months, have you at least once cleared your cookies or browsing history?	54%	46%	113
8	In the past 12 months, have you at least once deleted / edited something you posted online?	26%	74%	107
9	In the past 12 months, have you used someone else's cell phone without their knowledge?	9%	91%	250
10	In the past 12 months, have you used someone else's cell phone without their permission?	11%	89%	335
11	In the past 12 months, have you looked through someone else's cell phone without their knowledge?	10%	90%	250
12	**In the past 12 months, have you looked through someone else's cell phone without their permission?**	15%	85%	251

prevalence, and an item of interest that referred to a "snooping attack" with mild language.

One limitation of this study is that an entire baseline list wasn't tested, but only individual items. Although unlikely, interactions may exist between items that increase likelihood of ceiling or floor effects. Another limitation is that this study is not informative as to the possibility of contrast effects between the selected controls and the candidate items of interest, which could hurt credibility.

Either of these two limitations could potentially result in estimates of prevalence that were lower than the true population proportion. This was deemed acceptable, as it represented a conservative design choice for the main survey.

5. STUDY 2: CONDUCTING LIST EXPERIMENTS ON MTURK

Once we had selected items for the list question, we considered the problem of satisficing in MTurk list experiments.

In surveys, respondents sometimes select answers that are acceptable and easy to give, regardless of their truthfulness [31, p. 244]. Previous research suggests that satisficing is indeed a concern for survey research with MTurk samples [14, 22, 32].

There was reason to suspect that this concern extended to list experiments. List questions are cognitively more demanding than short, direct ones [11], taking more time and effort to answer thoughtfully. Yet, MTurk workers have incentives to maximize compensation per time unit [34]. For studies in which groups of observational units are compared, as is the case of list experiments, there are concerns that MTurk samples, especially those with non-naive participants, may provide measurements with greater error, leading to underestimation of effect sizes [8] and, at worst, to not finding effects when they are present (type II error).

One popular way to counteract satisficing is using attention check question (ACQs) [32, 35]. ACQs are questions whose right answer is known in advance, such as logic puzzles, trick questions, and direct instructions to answer a certain way. Although their use is well accepted and built on evidence (e.g., [35]), MTurk workers are now very much aware of this practice, and may have therefore adjusted. It has been suggested that some workers may scan for ACQs, answer them attentively, and rush through the remaining questions [18].

Another way to mitigate satisficing is restricting participation to high-reputation workers. When posting a task to MTurk, it is possible to restrict participation on a set of criteria. Two such criteria are commonly used as proxies for reputation: the total number of tasks that participants have completed in the past, and the proportion of their submitted work that was accepted by requesters. Previous research indicates that filtering participation to workers with at least 95% acceptance rate is sufficient to obtain good quality data [35]. But, based on our own experience conducting studies on MTurk, and expert opinion we had solicited, we came to believe that a 95% acceptance rate was now relatively easier to attain than at the time in which that research was conducted. There's indication that requesters have grown weary of refusing work, as it might affect their own reputations, which are disseminated in platforms like Turkopticon [21].

Since satisficing, and the measurement error associated with it, would affect the reliability of the estimates we were to obtain in our main study, we aimed to understand if list experiments in MTurk could be made trustworthy by restricting participation based on reputation and using ACQs. We devised a between-subjects experiment where surveys were administered to MTurk workers with distinct degrees of reputation (3 levels). Participants in each reputation group would be randomly assigned to receive a question with only the control items, or with the control items plus an item with ~0% expected prevalence, or with the control items plus an item with ~100% expected prevalence. Thus, we could compare the expected prevalence to the one estimated by the difference-in-means between groups.

5.1 Procedure

We configured an online questionnaire to randomly assign participants to receive a list question with one of the following lists:

Control The 4 control items derived from Study 1 (Table 1, items in bold).

Treatment-0 Control items, plus: "In the past 12 months, I've been to space, aboard an interplanetary vessel that I built myself" (~0% true prevalence).

Table 2: Number of participants, and mean items selected, by level of reputation and question version.

	Control		Treatment-0		Treatment-1	
	n_c	Mean	n_{t0}	Mean	n_{t1}	Mean
Low	51	1.71	54	1.61	44	2.59
Medium	46	1.13	47	1.51	42	2.43
High	57	1.46	33	1.45	60	2.50
Overall	154	1.44	134	1.54	146	2.51

Table 3: Prevalence estimated by the difference-in-means between groups, by level of reputation and question version.

	Treatment-0 - Control		Treatment-1 - Control		Treatment-1 - Treatment-0	
	Estimate	SE	Estimate	SE	Estimate	SE
Low	-9 %	0.190	88 %	0.186	98 %	0.189
Medium	38 %	0.195	130 %	0.201	92 %	0.206
High	-0.2 %	0.182	104 %	0.177	105 %	0.201
Overall	10 %	0.110	107 %	0.110	97 %	0.115

Treatment-1 Control items, plus: "In the past 12 months, I've opened my eyes in the morning at least once (for instance, after waking up)" (~100% true prevalence).

The attention check items were created by us, and, as far as we know, not previously used in MTurk surveys. In this way, we intended to minimize the effect of respondents detecting them without expending much mental effort, or using automated tools.

The rest of the questionnaire had the same structure and questions as the one to be used in the main survey. We posted it as a task on MTurk 3 times, assuring no repeated participation by the custom qualifications method [25]. Each time we posted it, we enforced system-level qualifications that limited participation to workers in the US, and created the following three reputation groups:

High Approval rate of 98% or higher, and at least 10,000 completed tasks.

Medium Approval rate of 95% or higher; at least 5,000, and no more than 10,000 completed tasks.

Low No minimum approval rate, and at most 5,000 completed tasks.

We targeted 150 participants by reputation group, with randomization expected to assign approximately 50 to each version of the list questions.

5.2 Results

5.2.1 Effect of reputation

Table 2 shows the average number of items that participants selected, discriminated by levels of reputation and version of questionnaire.

We found no evidence that the mean number of selected items was different depending on reputation, when the list question was either the Treatment-0 or Treatment-1 versions (columns 5 and 7; one-way ANOVA for Treatment-0: $F(2) = 0.305$, $p = 0.737$; for Treatment-1: $F(2) = 0.292$, $p = 0.747$). Only those that received the Control version, which had no attention check items, were found to have answered differently according to reputation level (column 3, $F(2) = 5.053$, $p = 0.00751$). Particularly, those in the (*Medium reputation x Control version*) condition selected, on average, 1.13 items, which was the lowest among those that received either the Control version or the Treatment-0 version.

5.2.2 Comparison to ground truth

Table 3 shows the estimates, by the difference-in-means, of positive answers to "been to space" (Treatment-0) and "opened eyes in the morning" (Treatment-1) items.

The difference between the means of the Treatment-0 group and the Control group was expected to be 0 if participants were answering attentively, since they had the same number of items they could identify with. If, on the other hand, participants were choosing at random, those that received the Treatment-0 version would have selected, on average, more items, because there is one more option – a truly random response pattern in both groups would yield a difference-in-means of 0.5. The difference we actually found, not taking into account level of reputation, was 0.1, which is non-negligible, as it would mean that 10% of our sample had travelled in space. We also observed an inconsistent pattern across reputation groups, with the abnormally low mean in the (*Medium reputation x Control version*) condition inducing a difference-in-means of 0.38, thus closer to 0.5 than the expected 0.

For differences between Treatment-1 and the two possible baselines, Control and Treatment-0, the same principle applies: attentive participation should yield a difference-in-means of 1.0, and random response 0.5. Either the Control or Treatment-0 can be baselines because one item in the Treatment-0 version has true prevalence of 0%. What we found was that when the baseline was Control, the overall difference-in-means, regardless of reputation, was 1.07, and when the baseline was Treatment-0, it was 0.97. The comparison between the groups that received attention checks, Treatment-0 and Treatment-1, was the closest to yield the expected proportion of 1.0. Furthermore, that comparison did not overestimate the true proportion, as did the comparison between Treatment-1 and Control.

Thus, the attention checks we had crated seemed to elicit enough attention from participants as to prevent degrees of satisficing that would jeopardize the validity of difference-in-means estimates. The feedback form that we included in the task provided some anecdotal indication that they generated goodwill among workers. As an example, participant 208 (low reputation group, Treatment-0 questionnaire version) commented: *"That was a funny attention check. I wish I could have answered as having done that."*

5.3 Discussion

Although we could not exclude that there were workers who engaged in satisficing, we did not uncover evidence of a pattern of misreporting that could be attributed to reputation, as measured by work history. The estimates by difference-in-means generally approached the expected 0% and 100% proportions. However, the Control group, which did not receive attention check items in their questions, was seemingly less consistent.

The differences-in-means between Treatment-1 and Treatment-0, both of which contained attention checks, were very close to the expected 100%, suggesting that the attention checks indeed mitigated the effect of satisficing.

We thus decided not to use reputation criteria to exclude partici-

pants in the main survey, as well as to add both the attention checks items. Inclusion of attention checks in both conditions of the main survey was the conservative design choice, as we had observed that their absence had, in this experiment, led to overestimation.

6. STUDY 3: MEASURING SNOOPING ATTACKS

6.1 Design

Having selected the list of items, and validated that a deployment to MTurk could provide good quality data, we proceeded to design and deploy the main survey.

We opted to create a very short questionnaire, with only the list question, and six other questions on personal characteristics, none of them open-ended. The questions are shown in Appendix B. The decision to not include more questions was made for two reasons. First, we had started with very concise research question, and broadening the scope before that question was answered could be a waste of time. Second, with more questions, or questions that were more probing, there was a risk that participants might feel that anonymity was reduced. For instance, they could reasonably suspect that their identity could be triangulated with responses to other surveys.

For that reason, we chose questions on personal characteristics carefully, for instance not including questions about level of income or race, which are very common in surveys, but that participants may feel to be very personal. We also asked for state of residency, but not city; and asked for level of education in broad categories.

Another design choice was the ordering of questions. We chose to show the list question at the beginning of the survey, to maximize attention and decrease incomplete responses. Since the question is cognitively heavy, it would be more frustrating to answer it after having cruised through simple demographics questions. We also inquired about personal characteristics in what we reasoned to be an increasing level of identifiability, to keep the sense of anonymity strong, as long as possible.

The list question included the control items and the item of interest selected in Study 1, and the two attention checks used as treatment manipulations in Study 2. The main purpose of including the attention checks was not to "catch" inattentive participants but to engage participants when thinking of the answer.

6.2 Fielding

We put the questionnaire online on a private web server, and configured it to randomly assign participants to either the treatment or the control group, each receiving the corresponding version of the list question. The survey proper was preceded by an informed consent form. We posted the survey several times as a task in MTurk, so that it would re-appear on the front page. Repeated participation was prevented by the custom qualification method [25]. MTurk qualifications were also used to restrict participation to residents in the United States. No other restrictions regarding past performance were enforced, as we found them to be superfluous in Study 2. Participants were paid $0.20, regardless of them giving valid responses. The survey took 1 to 2 minutes to complete attentively.

6.3 Data cleanup

We received a total of 1,481 responses to the survey. Of those, 84 (6%) were incomplete, and were removed from the dataset. Additionally, 16 responses (1%) were eliminated for being obviously invalid: 8 for responding “none” to the list question, and 8 for responding “all”. The following analysis is based on the remaining 1,381 responses.

Table 4: Summary of participant demographics, overall and by group, in the survey containing the list experiment question.

	Control (n_c = 688)	Treatment (n_t = 693)	Total (n = 1381)
By gender			
Female	43.2 %	42.3 %	42.7 %
Male	56.4 %	57.6 %	57 %
Other	0.4 %	0.1 %	0.3 %
By age group			
18-24	26 %	26 %	26 %
25-34	46.2 %	47.3 %	46.8 %
35-44	15.4 %	14.6 %	15 %
45-54	6.8 %	8.5 %	7.7 %
55-64	5.4 %	3 %	4.2 %
65 +	0.1 %	0.6 %	0.4 %
By level of education			
Less than high school	0.6 %	0.9 %	0.7 %
High school	28.3 %	27.4 %	27.9 %
Other college degree	18.8 %	19.9 %	19.3 %
Bachelor’s degree	41.4 %	39 %	40.2 %
Masters or PhD	9.6 %	11.4 %	10.5 %
Other	1.3 %	1.4 %	1.4 %
By region			
Midwest	23 %	21.1 %	22 %
Northeast	19.5 %	21.2 %	20.3 %
South	35.2 %	33.8 %	34.5 %
West	22.4 %	24 %	23.2 %
By ownership status			
Doesn’t own smartphone	12.4 %	10.1 %	11.2 %
Owns smartphone	87.6 %	89.9 %	88.8 %

Following Pew’s approach [39], we computed smartphone ownership status combining responses from two questions on ownership, SMART1 and SMART2. Whenever the response to the question “Is your cell phone, if you have one, a smartphone?” was “Not sure”, or “No, it is not a smartphone”, we referred to the next question, “Which of the following best describes the type of cell phone you have”, and assumed participants to be smartphone users if they selected either “iPhone”, “Android”, “Windows Phone” or “Blackberry”. There were 12 (1%) such cases.

Responses to the question about state of residency were binned into the 4 statistic regions defined by the US Census Bureau: Northeast, Midwest, South and West. For some of the analysis, ages were binned into commonly used age groups.

6.4 Dataset

6.4.1 Demographics

Table 4 summarizes the personal characteristics of the sample, segregated by control and treatment groups. A logistic regression of characteristics as predictors, and membership to either control or treatment group as outcome, did not reveal any significant differences between groups. Applying stepwise elimination of variables, starting with a model with AIC = 1926.1 and no significant predictors, the final model marginally improved AIC to 1916.45, with the elimination of all variables. In the final model, the remaining term was not a significant predictor ($Z = 0.135$, $p = 0.893$).

Therefore, as expected from randomized assignment, there was no evidence to suggest existence of a priori differences between the control and treatment groups, which would hurt the validity of the

Table 5: Number and proportion of respondents who selected each option in the list experiment item (adjusted for 4 control items).

	Control	Treatment
0	88 (12.8%)	76 (11%)
1	258 (37.5%)	204 (29.4%)
2	249 (36.2%)	239 (34.5%)
3	84 (12.2%)	122 (17.6%)
4	9 (1.3%)	43 (6.2%)
5	-	9 (1.3%)

prevalence estimates obtained through this list experiment. The demographics were similar across experimental groups, and any possible confounds could reasonably be expected to be equally distributed among them.

6.4.2 Attentive participation

We investigated if there were any indications that answers were inattentive. For that we looked at the relationship between how much time it took to answer the list question, and the actual response. If participants were rushing through the question, it would be expected that they had selected one of the first options, and hence that there would be a negative correlation between the time to complete the task and the number of behaviors that participants reported as having engaged in.

The correlations for either group were close to 0 (treatment: r = -0.0015 with 95% CI -0.0760 to 0.0730; control: r = 0.0185 with 95% CI: -0.0563 to 0.0931), and, for both, the hypothesis of the true correlation being 0 could not be excluded (treatment: t (691) = -0.402, p = 0.968; control: t (686) = 0.484, p = 0.6284). We therefore found no evidence that participants chose one of the first options that were available.

The possibility remains that participants chose an answer at random. Given the random assignment to groups, the noise created by responses at random should be equally distributed among groups, thus affecting the error, but not the difference-in-means.

6.4.3 Response to list experiment question

Table 5 shows the raw distribution of responses to the list experiment question for both groups. The vast majority of participants selected an answer between 1 and 3 (85.9% in the control group, 81.5% in the treatment group). Thus, the presence of appreciable ceiling or floor effects was unlikely.

We then investigated the possibility that the sensitive item changed how participants in the treatment group identified with the control items. For instance, participants could be more willing to identify with having called a 1-900 number because it appeared to be less censurable when compared to snooping. Blair and Imai [4] describe a statistical procedure to check for such an effect. Constructing the prescribed tabulation of estimated proportions of types of responses, we found no negative estimate. We therefore concluded that there wasn't evidence of a design effect.

Taking all this evidence together, we concluded that the design of the study and its deployment yielded a sound dataset.

6.5 Prevalence estimate

We defined (1-year) prevalence as the proportion of people in the population who internally identified as having had looked through someone else's cell phone without their permission. Prevalence was estimated by the difference-in-means between groups in a list experiment.

Table 6 summarizes the estimated 1-year prevalence for the sample and further breaks it down by segments of personal characteristics. For the overall sample (line 1), the 12-month estimate of prevalence was 31%. Our sample was not, however, a fair reflection of the U.S. population. Participants, on average, were younger, attained a higher level of education, and predominately identified as being male, which is expected in MTurk convenience samples [7]. We adjusted the data to the U.S. population estimates from the 2010 Census, and obtained an estimate of 20% for the U.S. adult population (see Table 7).

The data was adjusted with cell-based post-stratification weighting. We created weights for strata which, from the sample subset summaries, we found to have appreciably different prevalence estimates between levels. Using every possible demographics criteria to stratify would create cells with two few observations. Even the combination of gender, age group and region yielded marginal frequencies of 0. Moreover, using demographics criteria for which there weren't diverging differences between strata would have little impact on the overall prevalence estimate. We therefore decided to use weights based on the cross-tabulation of only age group and gender. At that granularity, the number of observations for some (AGE * GENDER) subsets was still too low to obtain reasonable weights. Recoding the 3 older age groups into one (45+), we were able to obtain more adequate weights, shown in Appendix C. As with any adjustment of this type, we obtained a more representative estimate, at the cost of increasing standard error. The national population statistics and diagnostics are shown in Table 7, and were computed with the R "survey" package, which implements Lumley's [24] weighted analysis instruments.

6.6 Trends

Although the overall 1-year estimates are informative by themselves, having a large sample allows us to look at differences between cohorts that can help explain the phenomenon. Table 6 suggests that in all demographic criteria, except for level of education, the estimates of prevalence are considerably different between subsets, but more detailed analysis is required to discern if demographic criteria can *predict* lower or higher prevalence.

It is, however, impractical and uninformative to try to understand the underlying demographics of snooping behavior based on all possible criteria. We therefore sought to find the demographic variables that better explained the list experiment outcomes, and only then to model the prevalence according to those variables.

6.6.1 Variable selection

To find relationships between demographic criteria and prevalence, we first constructed linear regression models of the number of items participants selected as a function of each available variable (gender, age, level of educations, region, and ownership), controlling for assignment to control or treatment group. Table 8 summarizes those models with the R-squared and F statistic, and shows comparisons to a smaller model in which the group assignment is the only predictor. Coefficients of each model are reproduced in Appendix D.

Regarding gender, for respondents who identified as being female, the prevalence estimate in the sample was 38%, whereas for the ones who identifies as male, it was 26% – a difference of more than 10 percentage points (Table 6, lines 2 and 3). However, the model with the both gender and experimental group as predictors, indicated that the gender variable explained very little of the vari-

Table 6: Estimated 1-year prevalence in the sample, as estimated by the difference in means between experimental groups. The table shows estimates for overall sample and for subsets based on personal characteristics. No estimations were made for subsets in which there were less than 20 observations in either experimental group, except for the age 65+ subset, which was binned with the 54-65 subset into the 55+ level. *P*-values from a t-test with the null hypothesis that there was no difference between experimental groups, with alpha set at 0.05. Bonferroni-adjusted significant differences in bold.

	Control group mean (SE)	Treatment group mean (SE)	Prevalence (SE)	*P*-value
Overall	2.517 (0.035)	2.825 (0.042)	30.8 % (0.055)	**<0.00001**
By gender				
Male	2.500 (0.046)	2.759 (0.057)	25.9 % (0.073)	**0.00043**
Female	2.542 (0.053)	2.918 (0.063)	37.6 % (0.083)	**0.00001**
By age group				
18-24	2.631 (0.067)	3.156 (0.086)	52.4 % (0.109)	**<0.00001**
25-34	2.522 (0.051)	2.820 (0.062)	29.8 % (0.080)	**0.00023**
35-44	2.509 (0.089)	2.644 (0.096)	13.4 % (0.131)	0.30730
45-54	2.362 (0.116)	2.407 (0.124)	4.5 % (0.169)	0.79038
55+	2.158 (0.158)	2.240 (0.202)	8.2 % (0.257)	0.75036
By level of education				
High school	2.482 (0.061)	2.789 (0.087)	30.7 % (0.106)	**0.00396**
Other college degree	2.667 (0.085)	2.949 (0.096)	28.3 % (0.129)	0.02889
Bachelor's degree	2.526 (0.054)	2.826 (0.067)	30.0 % (0.086)	**0.00053**
Masters or PhD	2.318 (0.110)	2.633 (0.105)	31.5 % (0.153)	0.04102
By region				
Midwest	2.494 (0.071)	2.699 (0.092)	20.5 % (0.117)	0.07989
Northeast	2.515 (0.078)	2.776 (0.093)	26.1 % (0.122)	0.03290
South	2.566 (0.060)	2.915 (0.072)	34.8 % (0.094)	**0.00024**
West	2.468 (0.073)	2.855 (0.086)	38.8 % (0.113)	**0.00067**
By ownership status				
Doesn't own smartphone	1.800 (0.093)	1.914 (0.093)	11.4 % (0.131)	0.38513
Owns smartphone	2.619 (0.036)	2.928 (0.044)	30.9 % (0.057)	**<0.00001**

Table 7: Proportion of U.S. adults who snooped on mobile phones in a 12 month period, as estimated by the difference in means between groups in a list experiment. Sample adjusted by cell-based post-stratification weighting to the 2010 Census by age and gender. *P*-value from a design-based t-test of the difference in means.

	Control group	Treatment group	Prevalence	*P*-value
Adjusted mean	2.41	2.61	20%	0.01515
SE	0.055	0.061	0.081	

Table 8: Linear regression models of number of items selected in the list experiment question. The first row indicates the proportion of variance explained by being in the treatment or control group. In the remaining rows, a variable is added to that model. F statistic from an ANOVA of the smaller and larger models.

Predictor variables	R^2	ΔR^2	*F*	D.f.	*P*-value
GROUP	0.022				
GROUP + GENDER	0.025	0.003	1.87	2	0.1542
GROUP + AGE	0.053	0.031	44.78	1	**<0.0001**
GROUP + EDUCATION	0.031	0.009	2.47	5	**0.0306**
GROUP + REGION	0.025	0.003	1.32	3	0.2671
GROUP + OWNER	0.100	0.077	118.38	1	**<0.0001**

ance in either group. This model did not significantly improve on the smaller model, with just the experimental group as predictor, explaining only an additional 0.003 of the variance (Table 8, line 2). Gender, therefore, did not seem to have strong relationship with snooping behavior, or at least not strong enough to justify including it in a model with other predictors.

Age (modelled as continuous variable, not by age group), on the contrary, significantly contributed to selecting more items. Looking at the details of the model, each additional 10 years predicted selecting, on average, less 0.18 items ($p < 0.0001$), in addition to the effect of group membership. Age, was therefore, considered a good candidate variable for a larger model.

The results of the model of level of education were mixed. Level of education can be thought of as an ordered variable, raising the question of whether more education could predict selecting a greater or lower number of items. Looking into the estimates of that regression, we found no clear evidence. Taking post-graduate education as a baseline, the model indicated that those with a college or Bachelor's degree selected a higher number of items (+ 0.33 with $p = 0.0016$, and + 0.20 with $p = 0.0347$, respectively), but there wasn't evidence of an effect for other levels of education. We expected to find that greater predicted difference in number of selected items would be associated with the greater differences in level of education, but that was not the case. Without an interpretation for that pattern, we concluded that this variable was not a good candidate for a larger model, despite the fact that adding it modestly improved the smaller model.

Region, like gender, did not seem to have a relationship with prevalence, on the basis that the model including it as a predictor did not significantly improve on the smaller model. We found it, therefore, to not be a good candidate.

Finally, regarding ownership status, the model suggested that those who owned smartphones selected more items from the list, even when controlling for membership in either control or treatment group. Adding ownership status to a model of only group membership explained 7.7% more of the variance, the greatest difference we found. Looking at the estimates of the model, we found the additional effect of owning a smartphone to be selecting 0.91 more items ($p < 0.0001$). Thus, ownership was clearly judged as candidate variable for a larger model.

6.6.2 Model

Having identified gender and smartphone ownership status as variables of interest, we finally aimed to understand how they predicted the probability of engaging in snooping attacks. For variable selection, we had used number of items selected, controlled by group membership, as an indicator of higher probability. For the final model, we wanted to look at actual predicted probability, while using both variables as predictors, and accounting for possible non-linear relationships.

Recently, it has been noted that although list experiments cannot reveal what each participant responded to the sensitive item, it is still possible to estimate conditional and joint proportions [10, 15], and thus model the joint probability distribution [4, 20]. Using the R "list" package [3] to that end, we created a model of the proportion of respondents identifying with the sensitive item, as a function of age and ownership status.

Appendix E.1 shows the coefficinet of that model, and Figure 1 depicts it graphically. It shows two clear trends:

- There is a sharp, concave decline in likelihood of snooping as people get older. Each additional year of age disproportionately decreases the likelihood of snooping on others.
- Those that own smartphones are more likely to engage in snooping. The difference is attenuated, and eventually disappears, as people get older.

The model also suggests that the youngest participants who are smartphone owners, are more likely to have snooped on others than to have abstained from it. Thus, for some groups, conducting snooping attacks, as we have defined them, may be the norm, not the exception.

6.7 Discussion

Summarizing, through a list experiment, we estimated the 1-year prevalence of successful snooping attacks to be 30.8% in an online sample. With post-stratification weighting, we generalised that finding to a national population, estimating that 20% of US adults had engaged in snooping in a 1-year period. Looking at specific subsets of the sample, some apparent trends emerged, but, due to the nature of list experiment data, comparisons between raw subsets can be misleading. Expanding our analysis, we did not find gender, level of education, or geographical sub-region to be strongly related to snooping behavior. We did however find that being young, and owning a smartphone, was independently linked to the likelihood of engaging in snooping. In the sample, those that did not own smartphones were, indeed, much less likely to have engaged in snooping attacks (11% 1-year prevalence), while those that were younger were more likely (52% 1-year prevalence in the 18-24 age group).

It should be noted, however, that being young and owning a smartphone is very much related: in the US, 85% of those between 18 and 29 own a smartphone, whereas for those that are 65 or older, the proportion is 27% [39]. In our sample, there is also a notable relationship between the two variables ($r_{\text{point-biserial}} = 0.28$). This fact suggests that there other variables, which we did not examine, relating to both age and ownership.

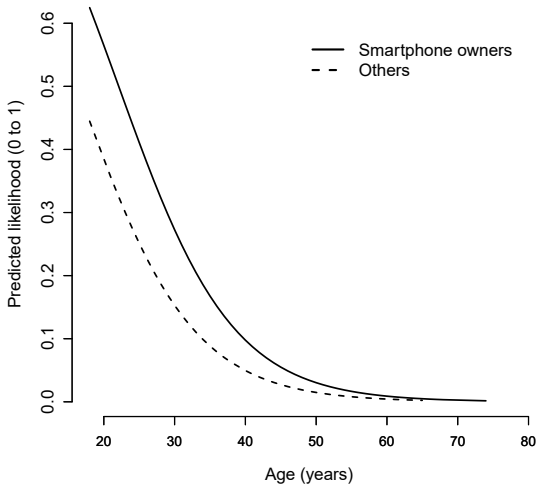


Figure 1: Predicted 1-year likelihood of having engaged in snooping attacks, by age and smartphone ownership status. Predictions from a list experiment regression model, shown in Appendix E.1.

7. STUDY 4: SNOOPING ATTACKS AND DEPTH OF ADOPTION

Being young and owning a smartphone, variables which the model suggests to be indicative of higher likelihood of engaging in snooping attacks, are also the typical characteristics of "digital natives." This population is known to be much more aware and concerned about threats within social context, such as snooping [23]. Where does that concern stem from? We hypothesize that those who use smartphones intensively as gateway to their social lives, thus producing privacy-sensitive information, become, by their own experiences, more aware of what they would have to gain, or loose, with a snooping attack. Thus, they would be more concerned about others snooping on their devices, and they would also be more likely to snoop on others.

In a final list experiment, we examined the likelihood of engaging in snooping attacks among smartphone users. Specifically, we explored how that likelihood is influenced by age, and by the degree to which people use their devices for personal purposes, in ways that may leave a trace of potentially privacy-sensitive data.

7.1 Procedure

We created a new online survey, similar to the one used in Study 3. The questions about gender, level of education, geographical region, and smartphone ownership were removed. The list experiment question, the question about age, and the question about the kind of smartphone the participant had were kept (the latter without the option "I do not have a cell phone").

An additional question group, shown in a second page, was added. This question group was a Likert scale of depth of adoption for

privacy-sensitive purposes, with 10 questions. For each, participants rated their perceived degree of frequency of use, from "Never" (1) to "All the time" (7). As an example, one item was "I use my smartphone to to look up information about health conditions". Items were based on behaviors of smartphone users that were reported in a Pew survey [39]. The scale is reproduced in Appendix F.

The survey was fielded in MTurk, following the same procedure as Study 3. The advertisement (HIT) asked specifically for smartphone users, both in the title ("Survey of smartphone users") and the description ("[...] Do not accept this HIT if you do not regularly use a smartphone"). Data cleanup was done also as described in Study 3, resulting in the exclusion of 7 responses (1%). All participants were paid $0.25.

There were 653 valid responses, 314 of which in the control group, and 339 in the treatment group. The majority of participants (56%) reported having an Android smartphone, followed by an iPhone (41%), Windows Phone (3%) and Blackberry (<1%). No participants selected the option "I do not have a smartphone", that was kept to exclude responses in case of inattentive reading of the advertisement.

7.2 Results

7.2.1 Depth of adoption and age

Responses to the depth of adoption scale, whose possible values are between 10 and 70, ranged from 16 to 70, and where somewhat skewed toward the higher end. The middle point of the scale is 40, and the mean response was 44.66 (SD = 10.6). Details about the distribution of responses, for the scale and individual questions, can be found in Appendix F.2.

Responses to the depth of adoption scale were, as expected, negatively correlated with age (r = -0.18, $t(651)$ = -4.78, $p < 0.00001$). This correlation, however, was not strong (according to Cohen's effect size criteria, it falls between small, 0.1, and medium, 0.3). Because depth of adoption, as it was measured, was relatively independent of age, it could more easily be interpreted as a predictor of likelihood of engaging in snooping attacks.

7.2.2 Depth of adoption as predictor

Using the same procedure as in Study 3, we created a model of likelihood of having engaged in a snooping attack, based on age and depth of adoption. The model predictions are depicted in Figure 2, and coefficients shown in Appendix E.2. In the left panel, the predictions are shown as a function of age, with a trend line representing a reduced model, with only age as predictor. In the right panel, the predictions are shown as a function of depth of adoption, with the corresponding reduced model line.

If there were noticeable differences in the pattern of dispersion in relation to the lines, such could be interpreted as one variable being a stronger predictor than the other (the stronger predictor should show less dispersion, or none at all). What is observable, however, is that neither the age or depth of adoption variables explain the other away.

The model with both variables has Log-likelihood of -868.786, which is higher than either the reduced models for age (-880.458) and depth of adoption (-873.834), indicating that it's a better fit. Predictions of both reduced models are strongly correlated to the ones of the larger model (age: r = 0.75, $t(651)$ = 28.6, $p < 0.00001$; depth of adoption: r = 0.71, $t(651)$ = 25.7, $p < 0.00001$). They are also correlated amongst themselves, as would be expected from the correlation of the variables, but no strongly (r = 0.14, $t(651)$ = 3.7, p = 0.00022). Again, these correlations indicate that neither variable explains the other away, and both contribute independently to the larger model.

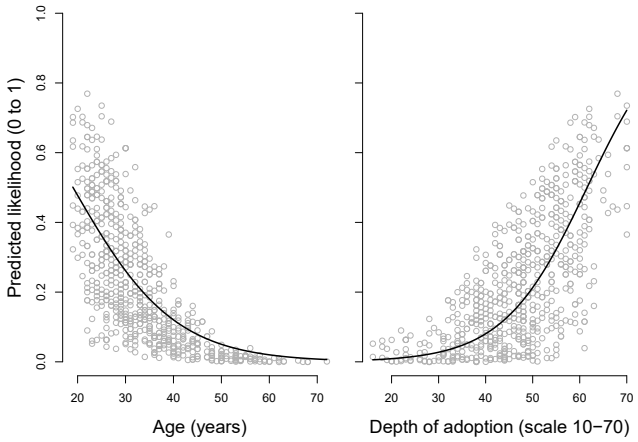


Figure 2: Predicted 1-year likelihood of having engaged in snooping attacks, by age (left panel) and depth of privacy-sensitive adoption (right panel). Dots represent per-participant predicted likelihood based on a list experiment regression model, with both age and depth of adoption as predictors. Trend lines represent the respective single predictor regression model. Regression coefficients are shown in Appendix E.2.

7.3 Discussion

We find evidence supporting the theory that people that use their smartphones in ways that may lead to privacy-sensitive information being kept, are more likely to snoop on others. Higher depth of adoption, as measured by a short scale we developed, predicts higher likelihood of identifying with the list experiment item indicating having "looked through someone else's cell phone without their permission" in the last 12 months, even when controlling for age.

However, depth of adoption does not explain away the effect of age that we had found in Study 3. Our scale, which was not thoroughly validated, may not have captured the factor it attempted to measure correctly. In fact, the scale does not accurately measure the frequency of certain behaviors, but how people *feel* about the frequency, which may be a weaker proxy for the construct of depth of privacy-sensitive adoption. Alternatively, there may also be, and we believe there are, other factors, related to age, which weren't measured but also play a role in predicting higher likelihood, like tech-savvy, or degree of volatility of social relationships.

8. CONCLUSIONS

8.1 Summary of findings

In this paper, we shown that the prevalence of snooping attacks on mobile devices is considerably higher than previously estimated. We found new evidence supporting that the problem is related to depth of adoption of mobile technology, and thus, that it is the youngest, those who use smartphone, and particularly those that use smartphones in ways that it stores privacy-sensitive data, that are more likely to snoop on others. In some segments of the population, people were more likely to "have gone through someone else's phone without permission", than not, in a period of one year.

To obtain these findings, we conducted a series of empirical studies. In the first two studies, we designed items for a list experiment, and validated the use of that methodological approach in MTurk. Our finding that list experiments in MTurk produce reliable data, as long as there are appropriate attention checks, is a secondary contribution of this work.

In the latter two studies, we conducted list experiments that inform on the prevalence of snooping attacks. Employing conservative design choices, that may have had the effect of underestimating prevalence, we were still able to estimate 1-year prevalence rates for the MTurk population, and, by weighting, for the U.S. adult population, that are much higher than previous lifetime prevalence indicators. Furthermore, we uncovered predictors of the likelihood of engaging in snooping attacks, and discerned independent population trends related to age and adoption of smartphones. We hypothesize that one mechanism for the observed trends is that users learn by their own experiences the kinds of valuable information kept on smartphones, which makes them more capable of engaging in snooping attacks.

8.2 Implications

This state-of-affairs can and should be addressed. There is room to improve privacy-preserving technologies that still impose too much effort on users, like mobile authentication. In recent year, biometric authentication on mobile devices, especially fingerprint authentication, has become more available and usable. There have also been extensive research efforts in making secret-based authentication more usable. Trends such as these indicate that defenses may be catching up.

However, two considerations should be given to the authentication approach of defense. First, as usable as authentication is made to be, it is not unreasonable to think that, for many people, it will never be attractive. Potential users of secret-based authentication may continue to think that it's a hassle. Potential users of biometric authentication may have privacy concerns. Defenses against snooping attacks for those people are few, if any.

A second consideration it that innovations in authentication should include snooping attacks in their threat models, because snooping attacks are likely to be attempted. Some adaptive authentication methods that have been proposed can reduce authentication requirements when devices are in "trusted places", like at home or at work (for instance, Android's Smart Lock [16]). It should now be clear that, in face of the pervasiveness of snooping attacks, that increase in usability will likely come at the cost of increased security risk.

Another possible road to improve the current situation is education and awareness-building. In that respect, however, it should be noted that in the realm of security, there has been little success in getting expert's messages across to users [19]. Specifically in the case of snooping, the reality if that many people are already aware of the risk, and want to secure against it, but fail to find practical ways to do it [28].

We hope this work plays a role in helping builders of interactive systems, educators, and policy-makers, to consider, when reasoning about mobile security, how prevalent it is for users' privacy to be violated by people they know.

8.3 Snooping as an attack

We have abstained throughout this paper from making judgements on whether snooping on others is justified. The use of the word *attack*, common in security lingo, should not be taken as having legal or moral connotations. It is an *attack* in the sense that actions were taken by an agent to circumvent an access policy; as much as one would call a *brute-force attack* to a situation where a mobile device owner who, upon forgetting their own PIN, ran a script to try out all possible combinations. We are aware that some people think it is acceptable for parents to go through their children's devices, or for romantic partners to go through one another's devices, and we do not dispute those opinions.

We note, however, that people who hold the opinion that their unauthorized access is acceptable, should also not be greatly impacted by social desirability bias. Thus, they should be expected to trend towards answering truthfully to a direct question on the topic. In the first study here reported (Section 4), and in previous studies [29], between 9% and 15% of respondents admit to having had snooped when asked directly. However, we found, for a comparable sample, that 2 to 3 times more people (~31%) self-identify with the behaviour when asked indirectly. The gap can be explained by participants themselves finding their actions censurable. We must conclude that a large portion of the population engages in a behavior that they know to be, from their own personal perspective, an attack, in the common sense of the word.

8.4 Future work

Security risks are often seen as being a function of the probability that they materialize and the severity of their consequences. This series of studies is informative as to the first factor, probability. We have, in this paper, focused on an overall measure of probability, and its relationship to demographic and usage factors. It would now be important to find other factors, especially ones related to the relationship between the attacker and the attacked (like social distance and motivation), and factors related to the context that creates the opportunity for the attack (like physical environment and circumstance). Both would be important for evaluating the effectiveness of new or existing defenses.

The other factor of which risk is a function is the severity of the consequences. We did not explore severity in this paper, but note that theory (e.g., [33]) predicts that the loss of control over what people that matter to us know about us, is likely to have considerable impact. We also note that one practical challenge in assessing severity is that people may not associate negative outcomes in their lives with someone having had snooped through their device, because, as our data suggests, they may never find out that it happened. Still, it is possible to gage how people *think* they would feel, or how they felt in the instances they know about, and find distinctions related, again, to context or social relationship between parties.

Both a fine-grained understanding of probability and of severity requires additional research, which we leave for future work. The quantitative approach we have employed here is not appropriate for a wide exploration of possible explanations, and possible outcomes, of snooping attacks. Finding factors requires breadth, and calls for a more qualitative approach. We believe that the fact that snooping attacks are much more common than previously thought justifies such an effort.

9. ACKNOWLEDGMENTS

This work was partially supported by FCT through funding of a PhD studentship, ref. SFRH/BD/98527/2013, and of the LaSIGE Research Unit, ref. UID/CEC/00408/2013. Special thanks to Serge Egelman, Kristy Milland, to several anonymous members of TurkerNation.com, and to the MTurk workers who participated in our surveys.

10. REFERENCES

[1] J. Antin and A. Shaw. Social desirability bias and self-reports of motivation. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, page 2925, New York, New York, USA, May 2012. ACM Press.

[2] N. Ben-Asher, N. Kirschnick, H. Sieger, J. Meyer, A. Ben-Oved, and S. Möller. On the need for different security methods on mobile phones. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services - MobileHCI '11*, page 465, New York, New York, USA, Aug. 2011. ACM Press.

[3] G. Blair and K. Imai. list: Statistical methods for the item count technique and list experiment. Available at The Comprehensive R Archive Network (CRAN) `http://CRAN.R-project.org/package=list`.

[4] G. Blair and K. Imai. Statistical Analysis of List Experiments. *Political Analysis*, 20(1):47–77, Jan. 2012.

[5] G. Blair, K. Imai, and Y.-Y. Zhou. Design and analysis of the randomized response technique. *Journal of the American Statistical Association*, 110(511):1304–1319, 2015.

[6] A. Braunstein, L. Granka, and J. Staddon. Indirect content privacy surveys. In *Proceedings of the Seventh Symposium on Usable Privacy and Security - SOUPS '11*, page 1. ACM Press, 2011.

[7] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, 6(1):3–5, Feb. 2011.

[8] J. Chandler, G. Paolacci, E. Peer, P. Mueller, and K. A. Ratliff. Using nonnaive participants can reduce effect sizes. *Psychological Science*, 26(7):1131–1139, 2015.

[9] E. Chin, A. P. Felt, V. Sekar, and D. Wagner. Measuring user confidence in smartphone security and privacy. *Proceedings of the Eighth Symposium on Usable Privacy and Security - SOUPS '12*, July 2012.

[10] D. Corstange. Sensitive Questions, Truthful Answers? Modeling the List Experiment with LISTIT. *Political Analysis*, 17(1):45–63, Feb. 2008.

[11] E. Coutts and B. Jann. Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT). *Sociological Methods & Research*, 40(1):169–193, Feb. 2008.

[12] S. Egelman, S. Jain, R. S. Portnoff, K. Liao, S. Consolvo, and D. Wagner. Are You Ready to Lock? Understanding User Motivations for Smartphone Locking Behaviors . In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security - CCS '14*, pages 750–761, 2014.

[13] A. P. Felt, S. Egelman, and D. Wagner. I've got 99 problems, but vibration ain't one: A survey of smartphone users' concerns. In *Proceedings of the 2nd ACM Workshop on Security and Privacy in Smartphones and Mobile Devices*, pages 33–44, New York, New York, USA, Oct. 2012. ACM Press.

[14] U. Gadiraju, R. Kawase, S. Dietze, and G. Demartini. Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, pages 1631–1640, New York, New York, USA, 2015. ACM Press.

[15] A. N. Glynn. What Can We Learn with Statistical Truth Serum?: Design and Analysis of the List Experiment. *Public Opinion Quarterly*, 77(S1):159–172, Feb. 2013.

[16] Google. Google Smart Lock. Online, Retrieved Jan 19, 2016. `https://get.google.com/smartlock/`.

[17] M. Harbach, E. V. Zezschwitz, A. Fichtner, A. D. Luca, and M. Smith. It's a Hard Lock Life: A Field Study of Smartphone (Un) Locking Behavior and Risk Perception. In *Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 213–230, Menlo Park, CA, July 2014. USENIX Association.

[18] D. J. Hauser and N. Schwarz. Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, 48(1):400–407, 2016.

[19] C. Herley. More is not the answer. *IEEE Security & Privacy*, 12(1):14–19, Jan.-Feb. 2014.

[20] K. Imai. Multivariate Regression Analysis for the Item Count Technique. *Journal of the American Statistical Association*, 106(494):407–416, June 2011.

[21] L. C. Irani and M. S. Silberman. Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 611–620, New York, NY, USA, 2013. ACM.

[22] A. Kapelner and D. Chandler. Preventing Satisficing in Online Surveys : A " Kapcha " to Ensure Higher Quality Data. In *Proceedings of the 2010 CrowdConf*, 2010.

[23] S. Kurkovsky and E. Syta. Digital natives and mobile phones: A survey of practices and attitudes about privacy and security. In *International Symposium on Technology and Society, Proceedings*, pages 441–449. IEEE, June 2010.

[24] T. Lumley. Analysis of complex survey samples. *Journal of Statistical Software*, 9(8):1–19, 2004.

[25] S. Maldonado. Using mTurk Qualifications to prevent workers from participating in an experiment multiple times. Online, Retrieved Jan 19, 2016 `http://sgmaldonado.com/main/content/using-mturk-qualifications-prevent-workers-participating-experiment-multiple-times`.

[26] P. McDonald, M. Mohebbi, and B. Slatkin. Comparing Google Consumer Surveys to existing probability and non-probability based internet surveys. Google Whitepaper, Retrieved Jan 19, 2016. `https://www.google.com/insights/consumersurveys/static/consumer_surveys_whitepaper_v2.pdf`.

[27] S. McNeeley. Sensitive Issues in Surveys: Reducing Refusals While Increasing Reliability and Quality of Responses to Sensitive Survey Items. *Handbook of Survey Methodology for the Social Sciences*, pages 377–396, 2012.

[28] I. Muslukhov, Y. Boshmaf, C. Kuo, J. Lester, and K. Beznosov. Understanding users' requirements for data protection in smartphones. In *Proceedings - 2012 IEEE 28th International Conference on Data Engineering Workshops, ICDEW 2012*, pages 228–235. IEEE, Apr. 2012.

[29] I. Muslukhov, Y. Boshmaf, C. Kuo, J. Lester, and K. Beznosov. Know your enemy: the risk of unauthorized access in smartphones by insiders. In *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services - MobileHCI '13*, page 271, New York, New York, USA, Aug. 2013. ACM Press.

[30] D. C. Mutz. *Population-Based Survey Experiments*. Princeton University Press, 2011.

[31] H. Müller, A. Sedley, and E. Ferrall-Nunge. Survey Research in HCI. In J. S. Olson and W. A. Kellogg, editors, *Ways of Knowing in HCI*, pages 229–266. Springer New York, 2014.

[32] D. M. Oppenheimer, T. Meyvis, and N. Davidenko. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4):867–872, 2009.

[33] L. Palen and P. Dourish. Unpacking "privacy" for a networked world. In *Proceedings of the conference on Human factors in computing systems - CHI '03*, number 5, page 129, New York, New York, USA, 2003. ACM Press.

[34] G. Paolacci and J. Chandler. Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science*, 23(3):184–188, 2014.

[35] E. Peer, J. Vosgerau, and A. Acquisti. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, 46(4):1023–1031, Dec. 2014.

[36] Pew Research Center. Privacy and Data Management on Mobile Devices. Report. Retrieved Jan 19, 2016 `http://www.pewinternet.org/2012/09/05/privacy-and-data-management-on-mobile-devices/`, 2012.

[37] Pew Research Center. Anonymity, Privacy, and Security Online. Report. Retrieved Jan 19, 2016 `http://pewinternet.org/Reports/2013/Anonymity-online.aspx`, 2013.

[38] Pew Research Center. The Future of Privacy. Report. Retrieved Jan 19, 2016 `http://www.pewinternet.org/2014/12/18/future-of-privacy/`, 2014.

[39] Pew Research Center. The Smartphone Difference. Report. Retrieved Jan 19, 2016 `http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/`, 2015.

[40] B. Rosenfeld, K. Imai, and J. N. Shapiro. An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions. *American Journal of Political Science*, 2015.

[41] R. Tourangeau and T. Yan. Sensitive questions in surveys. *Psychological Bulletin*, 133(5):859–883, Sept. 2007.

[42] M. W. Wiederman. The truth must be in here somewhere: Examining the gender discrepancy in self-reported lifetime number of sex partners. *The Journal of Sex Research*, 34(4):375–386, 1997.

APPENDIX

A. REPRODUCTION MATERIALS

Minimized datasets and R analysis code can be found at `https://github.com/diogomarques/snooping-paper`.

B. SURVEY QUESTIONS

List of questions in online survey reported in Section 6 (Study 3).

The first is a list experiment question, here shown in the version distributed to participants in the treatment group. Participants in the control group received the same question without sensitive item, in bold. The second and sixth items are attention checks.

LIST EXPERIMENT Below is a list of experiences you might have had in the past 12 months. To preserve your anonymity, we ask you only to indicate HOW MANY, not WHICH ONES, apply to you.

- In the past 12 months, I've shared pictures from my cell phone on social media.
- In the past 12 months, I've opened my eyes in the morning at least once (for instance, after waking up).
- In the past 12 months, I've purposefully made phone calls or sent text messages to 1-900 numbers.
- In the past 12 months, I've received at least one text message with unsolicited advertising (spam) on my cell phone.
- **In the past 12 months, I've looked through someone else's cell phone without their permission.**
- In the past 12 months, I've been to space, aboard and interplanetary vessel that I built myself.
- In the past 12 months, I've lost or had my cell phone stolen from me.

Please count how many you have had and indicate below.

◯ 0 (None) ◯ 1 ◯ 2 ◯ 3 ◯ 4 ◯ 5 ◯ 6 ◯ 7 (All)

AGE How old are you (years)?

␣␣␣

GENDER What is your gender?

◯ Male ◯ Female ◯ Other

EDUCATION What is your highest level completed education?

◯ Less than High School
◯ High School
◯ Community College or Professional School (College degree)
◯ University (Bachelor's)
◯ Graduate School (Master or PhD)
◯ Other: _____

STATE In which state do you reside?

◯ Alabama ◯ Alaska ◯ Arizona ◯ Arkansas [...]

SMART1 Some cell phones are called "smartphones" because of certain features they have. Is your cell phone, if you have one, a smartphone?

◯ Yes, it is a smartphone.
◯ No, it is not a smartphone.
◯ Not sure if it is a smartphone or not.
◯ I do not have a cell phone.

SMART2 Which of the following best describes the type of cell phone you have, if you have one?

◯ iPhone
◯ Android
◯ Windows Phone
◯ Blackberry
◯ Something else
◯ I do not have a cell phone

C. WEIGHTS

Weights used in post-stratification adjustment, based on the difference between Study 3's (Section 6) sample and the U.S. adult population, as measured by the 2010 Census.

Weights reveal that the sample was younger and had a greater proportion of males than the general population.

Gender	Age group	Proportion of US population	Proportion of respondents	Weight
Female	18-24	6.4%	10.4%	0.6162
Female	25-34	8.7%	19.0%	0.4596
Female	35-44	8.8%	6.5%	1.3459
Female	45+	27.6%	7.0%	3.9534
Male	18-24	6.7%	15.6%	0.4276
Male	25-34	8.8%	27.7%	0.3171
Male	35-44	8.7%	8.5%	1.0254
Male	45+	24.3%	5.3%	4.5923

D. VARIABLE SELECTION MODELS

Coefficients of linear regression models of number of items selected in the list experiment question, in Study 3 (Section 6). Models used for identifying candidate predictors of likelihood of having had engaged in snooping attacks.

The first model has a single predictor: assignment to either treatment or control group.

The remaining models add each of the other variables (gender, age, level of education, region, and smartphone ownership), controlling for assignment to control or treatment group.

Differences between models reported in Table 8.

Variables	Estimate	SE	*t*	*p*
Intercept	2.51744	0.03885	64.806	<0.00001
GROUP	0.30795	0.05484	5.616	<0.00001
RSE(1379) = 1.109; R^2 = 0.02236				

Variables	Estimate	SE	*t*	*p*
Intercept	2.57587	0.04999	51.531	<0.00001
GROUP	0.30797	0.05483	5.617	<0.00001
GENDER: Male	-0.10050	0.05545	-1.812	0.0702
GENDER: Other	-0.40287	0.51104	-0.788	0.4306
RSE(1377) = 1.018; R^2 = 0.02501				

Variables	Estimate	SE	*t*	*p*
Intercept	3.08289	0.09275	33.24	<0.00001
GROUP	0.30305	0.05399	5.613	<0.00001
AGE	-0.01784	0.00267	-6.692	<0.00001
RSE(1378) = 1.003; R^2 = 0.05313				

Variables	Estimate	SE	*t*	*p*
Intercept	2.32081	0.08951	25.929	<0.00001
GROUP	0.30991	0.05474	5.662	<0.00001
EDU.: Bachelor's	0.20050	0.09483	2.114	0.0347
EDU.: Some coll.	0.33175	0.10484	3.164	0.0016
EDU.: H. School	0.16002	0.09906	1.615	0.1065
EDU.: Less H.S.	-0.00675	0.33226	-0.02	0.9838
EDU.: Other	0.46345	0.24794	1.869	0.0618
RSE(1374) = 1.016; R^2 = 0.03108				

Variables	Estimate	SE	*t*	*p*
Intercept	2.44412	0.06408	38.14	<0.00001
GROUP	0.30814	0.05485	5.618	<0.00001
REGION: NE	0.08432	0.545	0.586	
REGION: S	0.1418	0.07478	1.896	0.0582
REGION: W	0.06479	0.0816	0.794	0.4274
RSE(1376) = 1.019; R^2 = 0.02516				

Variables	Estimate	SE	*t*	*p*
Intercept	1.72178	0.08209	20.975	<0.00001
GROUP	0.2875	0.05268	5.458	<0.00001
OWNER: Yes	0.90782	0.08344	10.88	<0.00001
RSE(1378) = 0.9781; R^2 = 0.0997				

E. LIST EXPERIMENT REGRESSIONS

E.1 By age and ownership status

Coefficients from a list experiment regression model where the sensitive item is whether someone "looked through someone else's cell phone without their permission" in the last 12 months. Data from Study 3 (Section 6).

Regression using Maximum Likelihood (ML) estimation with the Expectation-Maximization algorithm [4]. Control group parameters not constrained to be equal.

Variables	Sensitive item		Control items $h_0(y; x, \psi_0)$		Control items $h_1(y; x, \psi_1)$	
	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	2.014	1.714	-1.167	0.194	-3.529	4.567
Age	-0.124	0.057	-0.002	0.004	-0.024	0.018
Owner	0.732	0.953	0.832	0.122	3.824	4.542

E.2 By age, depth of adoption and both

Coefficients from list experiment regression models where the sensitive item is whether someone "looked through someone else's cell phone without their permission" in the last 12 months. Data from Study 4 (Section 7).

Regression using Maximum Likelihood (ML) estimation with the Expectation-Maximization algorithm [4].

Variables	Sensitive item		Control items	
	Estimate	SE	Estimate	SE
Intercept	1.80821	1.48669	-0.17064	0.17876
Age	-0.09492	0.05080	-0.00728	0.00474

Variables	Sensitive item		Control items	
	Estimate	SE	Estimate	SE
Intercept	-6.95467	4.36000	-1.00872	0.21807
Depth adop.	0.11296	0.07714	0.01315	0.00446

Variables	Sensitive item		Control items	
	Estimate	SE	Estimate	SE
Intercept	-1.48857	4.23927	-0.88936	0.37492
Age	-0.11248	0.06047	-0.00457	0.00536
Depth adop.	0.07617	0.06999	0.01360	0.00505

F. PRIVACY-SENSITIVE ADOPTION

F.1 Scale

Scale used in Study 4. Each item indicates the perceived frequency of a type of smartphone use that can leave potentially sensitive information on the device. It attempts to measure, in a range from 7 to 70, the depth of privacy-sensitive adoption of smartphones.

PROMPT Here are some statements about smartphone usage for personal purposes.

Please answer on a scale from 1 to 7, where a 1 means that the statement indicates something you *feel like* you never do, and a 7 means that the statement indicates something you *feel like* you do all the time.

You can also use the values in-between to indicate where you fall on the scale.

[RANDOMIZE]

Item-1 I use my smartphone to check my personal email account.

Item-2 I use my smartphone to take pictures of myself or of people close to me.

Item-3 I use my smartphone to go on social networks (like Facebook, Twitter, Snapchat) with my personal account.

Item-4 I use my smartphone to exchange instant messages with people that are close to me.

Item-5 I use my smartphone to to look up information about health conditions.

Item-6 I use my smartphone to do online banking on my personal accounts.

Item-7 I use my smartphone to look up jobs or submit job applications.

Item-8 I use my smartphone to look up government services or information.

Item-9 I use my smartphone to look up directions to places, or to get turn-by-turn navigation.

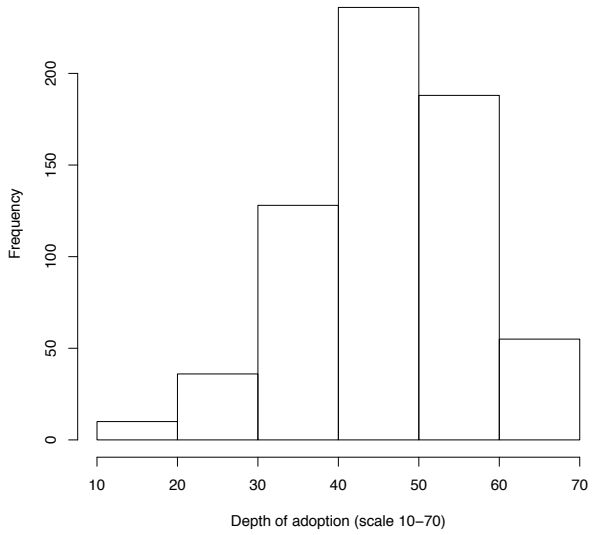
Item-10 I use my smartphone to organize personal affairs (for instance, access personal notes, calendar or shopping list).

F.2 Responses

Distribution of responses to scale and individual items in Study 4 (Section 7).

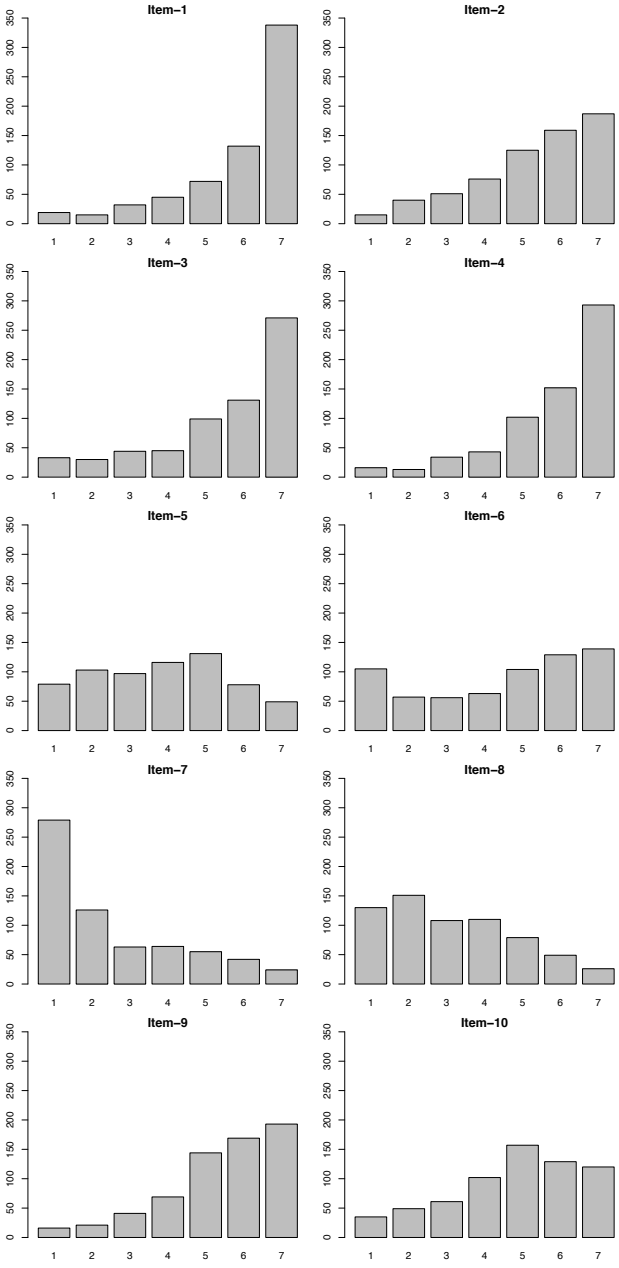
F.2.1 Scale

Sum of ratings to individual items.



F.2.2 Items

Frequency of response to scale items, each rated 1 (Never) to 7 (All the time).



Understanding Password Choices: How Frequently Entered Passwords are Re-used Across Websites

Rick Wash
School of Journalism
Michigan State University
wash@msu.edu

Emilee Rader
Media and Information
Michigan State University
emilee@msu.edu

Ruthie Berman
Macalester College
rberman@macalester.edu

Zac Wellmer
Michigan State University
wellmerz@msu.edu

ABSTRACT

From email to online banking, passwords are an essential component of modern internet use. Yet, users do not always have good password security practices, leaving their accounts vulnerable to attack. We conducted a study which combines self-report survey responses with measures of actual online behavior gathered from 134 participants over the course of six weeks. We find that people do tend to re-use each password on 1.7–3.4 different websites, they reuse passwords that are more complex, and mostly they tend to re-use passwords that they have to enter frequently. We also investigated whether self-report measures are accurate indicators of actual behavior, finding that though people understand password security, their self-reported intentions have only a weak correlation with reality. These findings suggest that users manage the challenge of having many passwords by choosing a complex password on a website where they have to enter it frequently in order to memorize that password, and then re-using that strong password across other websites.

1. INTRODUCTION

Passwords are a key part of many security technologies; they are the most commonly used authentication method. For a password system to be secure, users must make good choices about what password to use, and where to re-use passwords. Advice from security experts directs people to create, remember, and use passwords that are long, random, and unique to each account [21]. However, evidence from prior research suggests that people struggle to comply with this advice. For example, Das et al. [7] estimated that 43-51% of users re-use passwords across accounts, and Ur et al. [36] found that people feel like re-using passwords is not a problem, because they have never personally experienced negative consequences stemming from re-use. In reality, password re-use can introduce a serious security vulnerability which is difficult for any individual service operator to protect against [7].

People self-report that they re-use passwords to cope with the difficulty of remembering too many passwords, and that they believe they are not at risk because they re-use mainly passwords they believe are strong [36]. It isn't clear whether these self-reports represent wishful thinking by the users or whether they accurately reflect actual behavior. Few studies have been able to connect users' password-related attitudes and intentions with their own real-world password behavior, across accounts and over time.

It is especially important to be able to draw these connections between self-report and actual behaviors regarding password re-use, because re-use is a coping mechanism that occurs as a result of the demands and constraints users face when authenticating. Re-use is a user response to the burden of allocating limited memory capacity across the accounts and systems people use on a daily basis [15]. Despite many attempts to design more secure and usable systems, passwords remain one of the most widely deployed security systems in use today. The majority of people who use computers enter a password at least once a day; prior estimates [12, 30] suggest that computer users undertake between 8 and 23 password entry events every day![1]

We analyze a dataset that measures actual use and re-use of real-world passwords for web accounts. We captured password entry events that occurred in 134 subjects' web browsers over approximately six weeks. We also surveyed those same subjects immediately before and after the study period to collect self-reported demographics, attitudes, and intentions related to passwords. This allows us to examine not only how people think about passwords, but how that thinking translates into real-world password creation and re-use.

We found that people are re-using passwords across multiple websites. Our subjects primarily re-used passwords that are more complex, and re-used passwords that they entered frequently, such as the password for their university's website. We suspect that frequently entering a password is a way to memorize strong passwords, which are then re-used because

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22–24, 2016, Denver, Colorado.

[1]Our users entered an average of 3.8 passwords per day that they were active on their computer, or 3.2 passwords per day overall.

they are already very familiar. We also found that when asked about password use, subjects' responses were correlated with their actual password behaviors, but the correlation is relatively weak. This suggests that password choices are intentionally made, but that there are influences on password behavior other than password intentions.

Our results illustrate an important constraint on users' behavior that impacts password choice: how often a user is forced to authenticate with a particular password is related to how much they re-use that password on other accounts. This presents an opportunity for organizations to encourage the memorization of objectively strong passwords. However, it also results in greater potential for cross-site vulnerabilities as users prioritize using that stronger password in more places.

2. RELATED WORK

2.1 Password Creation and Management

People use passwords to authenticate on many different systems and servers on a daily basis. Estimates of the number of accounts that users maintain range from an average of 7-8 per person reported in a 2006 paper using data from a self-report user study with 58 subjects [16], to around 25 per user measured in a large-scale data collection including data from 544,960 browsers that was conducted around the same time [12]. In 2013, an online survey of 583 subjects found an average of about 18 accounts per user (median 14). And in a recent interview study, Stobert and Biddle [33] found that subjects reported having between 9 and 51 accounts, with a median of 27. People log in to a significant fraction of these accounts daily; Florêncio and Herley [12] found that people enter 8.11 passwords per day, and Hayashi and Hong [19] estimated that people use around 12 accounts per day.

Common password advice directs users to create passwords that are unique to each account, and random. However, for most people, authentication is a secondary task that presents a hurdle they must overcome in order to accomplish their primary task [2, 8]. So when people create passwords, their main goal is to make them easy to remember so entering a password does not impede their progress. When creating passwords, people often use information that is meaningful and important to them [8, 33], or has some connection to the service for which they are creating the password. For example, Inglesant and Sasse [20] reported that a subject described creating a password based on an item on his or her desk. People often use common names, words, and phrases in their passwords [29]; Shay et al. [31] found that about 80% of subjects reported they based their passwords on a word or a name. People also use rules, or an 'algorithm' [36], to compose new passwords. These strategies allow people to more easily recall their passwords when they are needed [6].

Creating easy-to-remember passwords is especially important for people as the number of passwords they must remember increases. The more passwords one has, the harder it is to remember all of them [38]. Infrequently used passwords are also harder to remember, as are passwords that people are forced to change on a regular basis [30]. Despite these difficulties, memorization is still a common strategy for managing passwords. Several studies have found that relying on one's memory is more common than other mechanisms of storing passwords such as saving passwords in one's browser, using password manager software, or writing down passwords in an electronic file or on paper [16, 19]. Only two out of 49 subjects in a recent think-aloud lab study conducted by Ur et al. [36] reported that they use a password manager; 17 said that they "simply memorize their passwords without writing them down or storing them anywhere". Another common strategy is relying on automatic software mechanisms to store passwords. For example, 81% of subjects in an interview study conducted by Stobert and Biddle [33] said that their passwords are stored in their browsers or in the Apple Keychain. People write down or store hard-to-remember passwords even when they recognize that this is a "bad" password management strategy [34].

2.2 Password Strength and Guess Resistance

Because people report that they rely on their memories for password management and feel like they're protecting against attacks by other human beings [33, 35], even when they create passwords that meet or exceed forced constraints imposed by password composition policies [36], their passwords are still not very complicated. In a lab experiment that asked subjects to create passwords for 8 different kinds of websites, passwords for sites that subjects rated as less important were shorter, and for the least important sites the passwords tended to be lowercase only [18]. However, people do self-report that they try to use stronger passwords for more sensitive accounts [34, 17, 8]. Ur et al. [36] found that subjects believed adding a digit or a symbol to a password they were already using elsewhere would make it stronger and more secure.

In a survey of people affiliated with Carnegie Mellon University who updated their password as a result of a changed password composition policy, only 24% of respondents reported creating a password of length 8 (the minimum length to meet the new requirements); the rest of the passwords were longer [31]. The average length of the passwords of the subset of subjects who answered questions about length and the types and positions of classes of characters in their passwords was 10.1 characters, with estimated entropy of 31 bits. In contrast, Bonneau [3] found in a dataset of 70 million passwords (69.3 million users) for Yahoo! sites collected in May 2011 that passwords were in the 10-20 bit range, and Florêncio and Herley [12] found in their dataset from 544,960 browsers that had the Windows Live Toolbar installed (between 7/24/06 and 10/1/06) that average entropy was 40.54 bits.

Traditionally, password strength has been measured using Hartley Entropy [4]: the log base two of the size of the set of possible passwords. This corresponds to Shannon's definition of entropy only when all passwords are equally likely. However, Hartley Entropy mostly measures complexity, and is not a good measure of objective password strength when it comes to offline guessing attacks. While there is a relationship between entropy and guess resistance [22], Bonneau found that entropy doesn't measure the same thing as guess resistance [3]. It does not take into account that there are non-random patterns in users' password creation choices that make guessing easier than if all passwords were random sequences of characters. For example, a longer password made up of a dictionary word is easier to guess but can have a higher entropy score than a shorter, random password [14]. People who engage in the common practice of adding num-

bers to the ends and capitalizing the beginnings of passwords expect that this makes their passwords stronger. However, this is not the case; passwords with these patterns are likely to be less guess resistant in an offline attack because they are non-random. Password composition policies may be able to increase entropy, but adhering to a composition policy is not a guarantee of guess resistance [36].

2.3 Password Re-Use

In addition to creating passwords that are easy to remember, people cope with the cognitive demands of authenticating on many different systems by re-using passwords. This is a very common practice; for example, 50% of subjects in an interview study conducted by von Zezschwitz, De Luca, and Hussman [37] reported that they re-used passwords, and explained that if they did not re-use passwords it would be too hard for them to remember them all. In that same study, 45% of subjects said they were still using the very first password they had ever created, and most of them were still using it to create new accounts! Florêncio and Herley [12] collected "re-use events" where a password was re-used across different websites, and found that in 2006 an average user had 6.5 passwords, and each was used on 3.9 different accounts. Komanduri et al. [23] found that even when subjects in their online experiment did not reuse exact passwords in their entirety, they created new passwords by modifying existing passwords. Less than 30% of subjects in Shay et al.'s survey [31] said they had created an entirely new password to meet the new password requirements—most said they modified a password they were already using. Only three subjects in Ur et al.'s 49-subject think aloud study [36] said they would never re-use passwords; most said they had not experienced any problems stemming from password re-use on any of their accounts.

Analyses of leaked password datasets also show that people re-use passwords on multiple different accounts. For example, Das et al. [7] identified 6077 usernames that appeared in two or more leaked password datasets; for 43% of these usernames the passwords on the different sites were identical, and for 19% they were similar. Bailey, Dürmuth and Paar [1] obtained access to a dataset containing usernames and the associated passwords that had been collected by a malware trojan, and calculated a metric they called the "re-use rate": for two randomly chosen accounts of a random user, how likely is it that the two passwords for the accounts are identical? In their dataset, the re-use rate for identical passwords was 14%, and for similar passwords it was 19%. Most of the password re-use in their dataset was "exact" reuse of an entire password on another site. One subject in Sasse et al.'s interview study [30] said that they have one "central" password that they use for everything, which they make as strong as possible.

The more accounts people have, the more they report that they re-use passwords across accounts [27]. One finding from Inglesant and Sasse's diary study [20] was that people use "good" passwords—ones that are memorable and conform to password composition policy—as a "resource" they return to again and again when creating passwords for new accounts. Subjects in Stobert and Biddle's interview study [33] spoke about re-using passwords on infrequently used accounts because those accounts had less "need for security". Many other self-report studies have found that people categorize accounts and re-use the same password for accounts that are similar to each other. People say that they re-use passwords more on low-importance accounts, and avoid password reuse for high-importance accounts that have a greater need for security [16, 27, 33]. However, in a lab study, Haque, Wright and Scielzo [18] found that it was possible to use a common password list and knowledge of a subject's password created in the "lower-level" account condition to successfully guess their "higher-level" account condition passwords 33% of the time. This indicates that people's beliefs and intentions may be inconsistent with their actual re-use of passwords across account categories. Because lower-level accounts may be easier to compromise [14, 1], such re-use is a risky security practice.

2.4 Research Questions

Password Reuse:

There are contradictory results in the literature regarding which passwords people re-use more often. Most password data that speaks to re-use is self report from user studies, in which people say that they tend to reuse weaker passwords more often than stronger passwords (e.g., Stobert and Biddle [33]). However, Egelman et al. [10] found that there was no difference in password strength between passwords created by subjects in their experiment who reported reusing existing passwords, and those who said they had not re-used passwords. And Ur et al. [36] found that subjects believed re-use would not be a problem for them, because they felt that the passwords they re-use are strong. In addition, when asked about why they re-use passwords, subjects in many studies self-report that it makes passwords easier to remember [16, 36]; this implies that due to memory constraints passwords that users have to enter more frequently should also be re-used on more different accounts.

In order to measure re-use directly, it is necessary to have access to repeated instances of password use over time by the same person, and a mechanism that makes it possible to compare passwords to find out whether a person has entered the same password on more than one account. Florêncio and Herley [12] had access to this kind of data, and found that strong passwords are re-used at fewer sites ($M = 4.48$); weak passwords are used at more sites ($M = 6.06$). However, Bailey, Dürmuth and Paar [1] found in a different dataset that password re-use is more common for the high-value accounts (e.g., financial accounts) which have stronger passwords, than for all accounts. In our study, we collected data from specific individuals over a period of weeks. This means that we can examine which passwords are reused more by specific individuals, and on how many different accounts the frequently-entered passwords are re-used. Therefore, we ask:

Do people reuse their strong(er) passwords more, or their weak(er) passwords more? Do people reuse frequently entered passwords more than infrequently entered passwords?

Password Intentions:

In some studies, people self-report that they do have some idea what strong versus weak passwords look like, and what they say mirrors common password advice. Generally speaking, people report that they know unique and random passwords are more secure [16]. Ur et al.'s [35] subjects knew

that, for example, it was better to put upper-case letters, digits, and symbols in the middle of passwords rather than at the beginning or end, and that randomly chosen digits are better than years or "obvious sequences". But, when people create passwords, analyses of leaked datasets and experiment passwords show that they do not behave consistently with this knowledge [7]. They choose passwords that are simpler and easier to remember [38]. There is evidence from previous research about software updates that users do not always enact their security intentions correctly [40], however, this has not been examined before with respect to passwords. In our study, we collected log data about individuals' behaviors and survey data asking about their intentions, so we can connect how users think about passwords with password strength and re-use more directly than was possible in previous work. Therefore, we ask:

Do peoples' intentions for the passwords they create correlate with the characteristics of their actual passwords, and with which passwords they reuse more?

3. METHOD

3.1 Methods for Studying Passwords

Researchers have used a number of different methods to study passwords. Each method has strengths and weaknesses. Interview studies like Stobert and Biddle [33] allow for in-depth questioning about a small number of users, but are hindered by the tendency to remember what one normally or typically does and not what one actually does. This is a problem for password research, because for most users passwords are a secondary task [30]. This can mean their memory for their past behavior is biased. Diary studies like Hayashi and Hong [19] help to get around that by asking users to record instances of password behaviors, but are only as accurate as subjects are able to adhere to the data recording protocol and routine, and can only be conducted with a small number of users. Surveys (e.g., Shay et al. [31], Ur et al. [35]) allow the researcher to gather data from many more people, on the order of hundreds to thousands, but are limited in that they are self-report which may be inaccurate, especially when it comes to security intentions which might not match actual behavior [40].

User studies conducted in the lab or online often ask subjects to create passwords under specific conditions, and typically take steps to create scenarios that closely approximate situations users are likely to encounter in the real world to increase external validity of the research (e.g., Egelman et al. [10]). Online user studies such as Komanduri et al. [23] using Amazon Mechanical Turk can potentially reach a large number of people. However, many people behave differently when creating passwords for a user study than they do normally [11].

Password datasets collected through partnerships with companies or organizations and leaked password datasets include users' actual passwords, and some of these datasets are quite large. The security community has used these datasets to learn more about the passwords users choose, and analyzed them for patterns of common password composition characteristics. However, these datasets typically include little information about the users who created the passwords. An exception is Mazurek et al. [24] which through a partnership with Carnegie Mellon University was able to analyze password data from every account holder. This study correlated demographic data about faculty, staff and students of the university with password characteristics, in addition to analyzing the guess resistance of the passwords. Two papers use data collected over days (in the case of Bonneau [3], 69.3 million users) or months (in the case of Florêncio and Herley [12], 544,960 users) to present findings at the user level as well as at the password level.

Demographic	#	%
Man	61	46%
Woman	71	53%
18–29 years old	127	95%
30–49 years old	7	5%
High School Diploma / Undergraduate student	98	73%
Bachelors degree / Graduate student	36	27%
Have children	4	3%
No children	130	97%
White	103	77%
Asian	13	10%
African American	4	3%
Hispanic	6	5%

Table 1: Demographics of our sample

The study by Florêncio and Herley [12] is the most similar study to ours. However, they only were able to collect "reuse events": instances when a password was reused across more than one website. We have more accurate data about how frequently a password is entered into each website, data about passwords that were only entered into a single website (69% of passwords in our study), and self-report data about user perceptions.

3.2 Data Collection and Participants

Our study combines survey methods asking subjects about beliefs, behaviors and behavioral intentions, with log data about actual behaviors over time. Subjects installed custom-written log data collection software on their personal computers and web browsers for a median duration of six full weeks, and also took a survey at the beginning and at the end of the data collection period. This allowed us to collect both self-reported beliefs, behaviors and behavioral intentions and log-based behavioral measures for the same subjects, which enabled us to correlate subjects' security beliefs and intentions with their actual password characteristics and re-use. In this way we can examine how knowledge, attitudes, and intentions match up with behaviors within a person.

Our data collection software consisted of a web browser plugin for both Google Chrome and Mozilla Firefox. This plugin collected web use data, and uploaded it to our server. The plugin recorded all URLs visited by the web browser, as well as any form submission on a web page. Additionally, the plugin recorded all security-related settings and recorded information about all add-ons (plugins, extensions) installed and/or running. The plugin did not record anything while the user was in Private Browsing mode (Firefox) or Incognito mode (Chrome); subjects were instructed to use these modes for activities they did not want recorded. All connections to our server were encrypted to protect user privacy.

When the plugin detected a password HTML element in a form submission, it recorded the password entry: when the user entered the password, what webpage the user entered a password into, which password was entered, and how strong each password was (entropy, following Florêncio and Herley [12]). We did not collect plain text passwords; instead our browser plugin measured password entropy on the client and then hashed the passwords with a per-user salt before the information was sent to our server. This enabled us to examine which passwords were re-used by each subject across different websites without knowing his or her actual passwords. Additionally, since we collected data for a number of weeks, we were able to identify which passwords were re-used by each subject, and on what websites. We were not able to compare plain text passwords across subjects.

We recruited subjects from a large midwestern university by asking the registrar to email a random sample of students (both undergraduate and graduate). Students in computer science and engineering were excluded from participating. We sent out a total of 15,000 emails in three waves, and had approximately 247 students respond to our recruiting mail (1.6% response rate). Of those 247, about 180 were eligible to participate in the study: they had a personal computer running Windows 7 or Windows 8 which they said they used regularly, used either Google Chrome or Mozilla Firefox as their main web browser, and responded to our instruction emails. They were also required to have the ability to install software on the computer, and be the only user of the computer. The first two constraints (Windows, web browsers) are a limitation of our data collection software—supporting other operating systems and web browsers was prohibitively complex, so we designed the software to support the most popular operating systems and web browsers.

Of those subjects that were eligible to participate in the study, we received usable data from 134 subjects (0.8% usable response rate). The remaining subjects mostly were excluded due to unforeseen bugs in the data collection software that prevented sending accurate data, or because subjects did not use their computer enough (e.g. had more than 7 consecutive days without using the computer, not counting spring break). Two subjects had hardware problems with their computer that caused them to withdraw, and two other subjects withdrew without explanation. Our sample is fairly representative of the population of the university. Almost all subjects were in the 18-29 age range. Close to the demographics of the student population, our sample was 52% female and 76% white. Approximately 76% of the subjects were undergraduates, while the remaining are graduate students. Only 3 of the 122 subjects had children. Table 1 has more details.

All subjects provided informed consent to the data collection. Subjects were compensated a total of $70 for their participation; those who withdrew early received partial compensation. Subjects had the ability to turn off the data collection software at any time using a control panel that we provided, and we also provided instructions as part of the sign-up procedure for how to use private browsing mode. Our study was approved by our institutions's IRB.

	Min	25%	50%	75%	Max
Password Entries per day	0.4	1.6	2.5	3.9	33
Unique passwords entered	2	8	12	17	58
Unique correct passwords	1	4	6	8	18
Average password entropy	35	46	49	57	83
Length of Passwords	6.0	8.0	8.7	9.8	15
Websites with password	5	12	16.5	22	67
Website-to-Password Ratio	1.0	2.3	3.0	4.1	18
Frequency of Password Entry	1.2	2.1	2.7	3.6	37
Uses Password Manager			Yes: 26 — No: 108		

Table 2: Summary statistics about per-subject password usage. The 50% column contains the value for the median user; the 25% and 75% columns contain the first quartile and the third quartile users, respectively. "Frequency of Password Entry" is the average number of times a password is entered into each website.

4. RESULTS

4.1 Description of Password Use

Our dataset allows us to have a fairly comprehensive view of how each subject uses passwords on the web on a daily basis, over a number of consecutive weeks. We were able to capture every time a subject entered a password into a web page, and associate that with a specific browser and the user of that web browser. Our subjects visited an average of 5,613 web pages during the study (SD = 5,002), which translates to an average of 118 web pages per day (SD = 104). The median user entered a password into a web page 128 times over their participation in our study, though often they entered passwords into the same web pages on different days, or multiple times in a single day. Subjects ranged from a minimum of 22 password entries to a maximum of 1,474 entries, though most fell between 78 and 158 entries. The median user entered a password on 70% of the days they participated in the study, for an average of 3.2 passwords entered per day (SD = 3.5).

Our subjects used a median of 12 distinct passwords, though the number of passwords per subject varied quite a bit. On the low end, one subject entered only 2 distinct passwords (into 11 different websites). On the high end, another subject entered 58 different passwords over the study period, though most subjects ranged between 8 and 17 distinct passwords. This is not very many different passwords, given how frequently subjects needed to enter a password into a web page.

We grouped web pages into websites by domain name. Subjects entered passwords into a median of 17.5 different websites. They entered passwords into as few as 5 different websites, and into as many as 69, though most ranged between 12 and 19 different websites. As these numbers are higher than the number of distinct passwords, it is clear that our subjects tend to re-use the same passwords across multiple websites. One hundred fourteen of our subjects (85%) had fewer unique passwords than they did websites that they entered passwords into.

4.1.1 Likely Correct Passwords

At times, some users will enter more than one password into a website. This may be because they entered a typo or they forgot their correct password and are guessing passwords

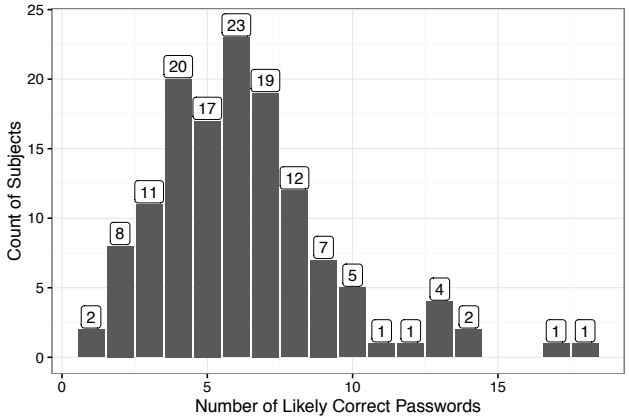


Figure 1: Histogram showing the number of passwords used by subjects in our sample.

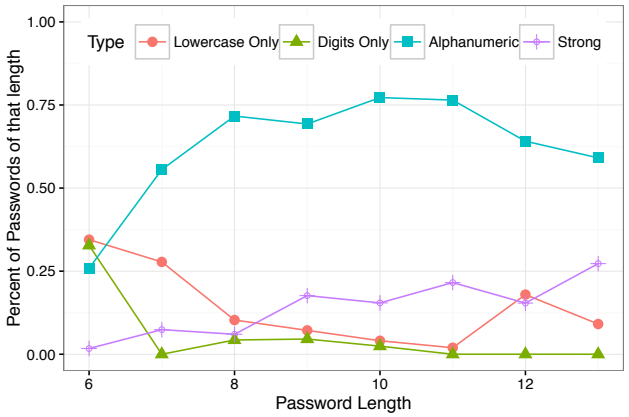


Figure 2: The type of password used, by length of password. Unlike in Florêncio and Herley [12], a clear majority of passwords we observed are alphanumeric passwords. We also observed significantly more strong passwords.

until they are able to successfully authenticate. It could be because they confused the password for one site with another site. It could be that the user has multiple accounts on the same website, or that they changed their password during the period of the study. Or it could be because they are guessing a friend's password for that website.

From our log data, we cannot tell which password was the correct password for a given website account. However, we can make an attempt to identify which password was *likely correct* based on usage patterns over time. We used a three step process for identifying which password is likely correct for a user on a given website:

1. The password that was entered most frequently into a given website is likely to be the correct one.
2. For websites where more than one password was frequently used, choose the password that was used on the larger number of days.
3. If there is still a tie (8% of websites), then choose the password that was used on the largest number of other websites by that user (the *Re-Use Assumption*).

This process successfully identified a likely correct password for 98% of websites. Most websites were fairly easy to choose a likely correct password; for example, one subject used 4 different passwords to log into his most frequently used website—3 were used once each, while the fourth was entered 96 times. Our subjects had a median of 6 likely correct passwords. Two subjects used only 1 likely correct password (which were correct on 10 and 18 different websites); one of our subjects correctly entered 18 different passwords over the study period, though most subjects ranged between 4 and 8 likely correct passwords.

4.1.2 Password Strength

For privacy reasons, we did not directly collect subjects' passwords. Instead, our data collection software calculated a standard *entropy* measure for each password before hashing the password and recording the hash. In our analysis, we use entropy not as a precise measure of how resistant a given password is to compromise, but to compare passwords created and used by the same person, across people, and across websites. Entropy allows us to roughly describe the complexity of a password, and as a result, how hard it might be for the user to remember. This is similar to how entropy has been used in several other studies of passwords. For example, Fahl et al. [11] used it to characterize relative differences between multiple passwords created by the same user. Florêncio, Herley, and van Oorshot [15] refer to entropy as a way to "represent user effort to remember a password". Egelman et al. [10] use entropy to quantify differences between groups of passwords created by participants in different conditions of their experiment.[2]

Averaging across all passwords that any of our subjects ever entered into a website, the average entropy is 49.2 bits (*SD* = 22.1). Passwords ranged in entropy from 4.322 bits (for a password consisting of a single symbol) up to 165.438 bits (for a 32 character alphanumeric password). These numbers, however, include all passwords that were entered into a website, including incorrect passwords and password guesses. If we only consider passwords that we identified as likely correct, then the average entropy across our sample is 49.5 (*SD* = 18.1). (The range is the same, as both the strongest and weakest passwords in our sample were likely correct on at least one website.) Subjects' strongest likely correct password had a median entropy of 65.5 bits, and the interquartile range was 53.6 bits to 82.7 bits.

4.1.3 Password Characteristics

Following Florêncio and Herley [12], we reverse engineered characters of passwords from the recorded entropy value. Given only a password's entropy and the knowledge of how it was calculated it is possible to reconstruct information about the password without ever knowing exactly what the

[2]Since we collected our data, Melicher et al. demonstrated a new technique based on deep learning to approximate the guessability of a single password in real-time [25]. We plan to use this in addition to entropy in future work. https://github.com/cupslab/neural_network_cracking

password was. A password's entropy was calculated by computing $entropy = \log_2 set\_size^{length}$. Rearranging, we see that $length = \frac{entropy}{\log_2 set\_size}$. For each possible size of character set, we can calculate an estimated password length. The correct set size and length is the one where the estimated password length is a whole number. This method does leave us with some possible limitations; it does not provide a way to differentiate between using lower case or upper case letters because they share the same character set size.

The average subject used passwords of length 8.98 (SD = 1.43) that used 2.29 (SD = 0.376) different character sets (from the set {Lowercase letters, Uppercase letter, Numbers, Basic Symbols, Extended Symbols}). Approximately 87% of passwords included a letter, 80% of passwords included a digit (number), and 14% included a symbol. Florêncio and Herley [12] found that the vast majority of the passwords that they observed were solely lowercase letters, with PINs (passwords consisting solely of numbers) the second most common; this is summarized in their Figure 9. Reproducing their Figure 9 using our dataset (Figure 2), we find that our subjects frequently used more complex passwords; the majority of our subjects use alphanumeric passwords, with "strong" passwords the second most common.

4.2 What passwords do people re-use?

Our median subject entered their passwords into 16.5 websites, and entered 12 distinct passwords into those websites. Overall, 31% of all passwords were entered into more than one website, and 20% of passwords were likely correct on more than one website.

Since the number of websites is larger than the number of passwords, this indicates that subjects re-used their passwords. We use our data to quantify password re-use: for each subject we calculate a website-to-password ratio — how many different websites on average each password is entered into by each user. A website-to-password ratio of 1.0 means that each website gets a different password. A website-to-password ratio greater than 1.0 suggests password re-use: passwords are entered into more than one website. A website-to-password ratio less than 1.0 happens when people change their passwords or enter incorrect passwords, thus entering multiple different passwords into a single website.

If we calculate a website-to-password ratio for each subject and then average them across subjects, we find that each password is entered into a median of 1.6 different websites. This number is likely a lower bound estimate for the true median website-to-password ratio, because it includes a number of incorrect passwords. If instead we only count passwords that we deemed to be *likely correct* for at least one website, then we find that the median subject in our sample entered a likely correct password into 3.0 different websites. This is because once incorrect passwords are removed, there remain a median of 6 correct passwords per subject.

In identifying which password was likely to be correct, we made a *re-use* assumption: among passwords used equally often, the one that was used on the most other websites was most likely correct. This assumption affected about 6% of user/website pairs, and it biases our re-use estimates toward higher levels of re-use. Thus, this second re-use estimate (each password is used on 3 websites) is likely an upper bound on the estimate for the true website-to-password ratio for the sample.

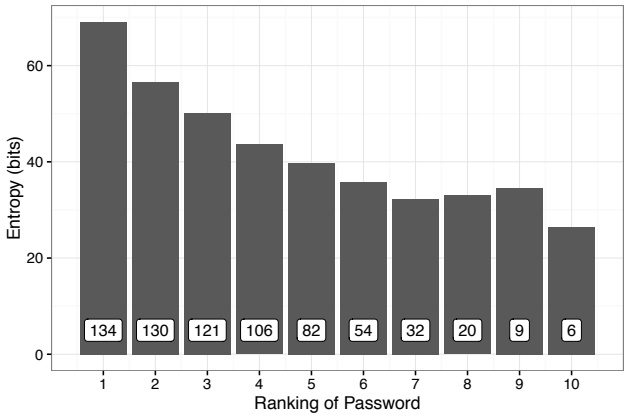


Figure 3: Average entropy for passwords at a given rank. The number near the bottom of each bar is the number of subjects with passwords at that rank.

4.2.1 People re-use strong passwords

Weaker passwords are easier to remember and to type, and most websites that require a password are not high-importance websites with complex password composition policies. People might therefore re-use their weaker passwords. On the other hand, they might re-use stronger passwords; memorizing a strong password takes more work so users might want to get the most out of that effort by re-using it wherever possible.

Among our subjects, people re-used their stronger passwords. There was a 0.063 correlation between the entropy of a password and the number of websites that password was entered into ($p = 0.007$). This positive, statistically significant correlation suggests that a subject's stronger passwords are the ones being re-used, though the small size of the correlation means that password strength isn't the whole story.

To better understand password re-use, we ranked all of the passwords that each subject used during our study by the strength (entropy) of that password. The password ranked #1 is the strongest password that subject ever entered, #2 is the second strongest, and so on down to password #N, the sybject's weakest password. This ranking is an individual ranking per subject, and allows us to examine whether it is the absolute strength of a password that influences re-use, or whether it is the strength relative to the subject's other passwords that matters. Figure 3 shows the average entropy of a password by individual ranking.

Putting both password entropy and password ranking into the same regression allows us to separately estimate the effects: for different passwords with the same entropy, does having a better ranking (lower number) lead to more re-use? And likewise, for password ranked the same, is the stronger one more likely to be re-used? However, password entropy and password ranking are highly correlated, (r=-0.628). Including both predictors in the same regression model can lead to collinearity issues. To address this, we ran three separate multi-level linear regression models (Table 3): one

	Entropy		Ranking		Both	
(Intercept)	1.82	***	2.70	***	2.61	***
Entropy	0.01	**			0.00	
Ranking			-0.06	***	-0.06	***
$R^2_{LMM_c}$	0.0040		0.0086		0.0086	

Table 3: Three multi-level linear models that analyze the effect of password strength on how many websites a password is re-used on. Each regression includes a random effect control for subject.

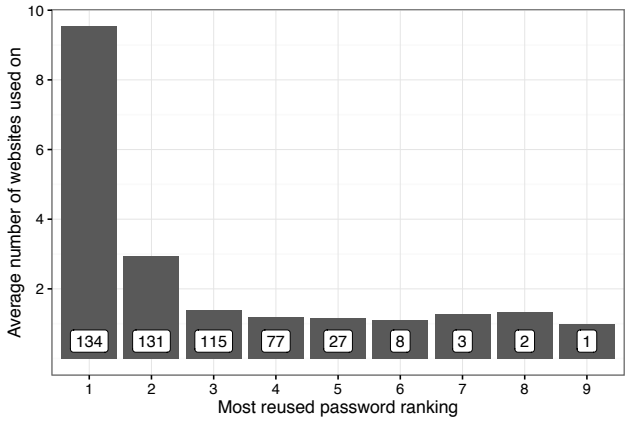


Figure 4: How often passwords are re-used. The leftmost bar shows the average for a subject's most-reused password; the second bar the second-most-reused password; and so on. The number near the bottom of each bar shows the number of subjects with passwords at that rank.

model with entropy, one model with ranking, and a model with both. By comparing the R^2 value for each model[3], we can see that the personal ranking is a better predictor of password re-use than absolute entropy, and indeed personal ranking explains almost all of the variance that both variables together explain.[4] In addition, Figure 4 shows that a subject's most re-used password is used far more often than any of that subject's other passwords. Thus, we conclude that it is not the absolute strength of a password that leads to re-use. People are re-using *their* strongest passwords, but not necessarily passwords that are objectively strong.

4.2.2 People re-use frequently entered passwords

Another possibility is that people are re-using passwords that they have to enter frequently. It is easier to remember a password if you have to enter it on a regular basis [5], and thus, passwords that need to be entered frequently might be easier to remember, and therefore easier to use. However, password re-use is correlated with the number of websites a password is entered into; passwords that are re-used on more websites will also naturally need to be entered more frequently. Instead of overall frequency, we looked at the number of times a password was entered, and divided it by the number of websites the password was used on to get a measure of the average number of times a password is entered into any single website. An average password from a median user was entered into each website it was used on 2.7 times.

[3]The R^2 measure for linear mixed models is the conditional R^2 for the whole model, $R^2_{LMM_c}$ from Nakagawa and Schielzeth [26].

[4]These R^2 numbers are are fairly low, which suggests that neither of these variables has much explanatory power. We just use these regressions to draw relative comparisons between the entropy and ranking variables to identify which predictor to include in future regressions. Our other regressions that we use to draw more substantive conclusions have more appropriate R^2 values.

	# Websites Correct	Password Re-used?	Non-univ Websites
(Intercept)	1.39 ***	-1.07 ***	0.87 ***
Ranking	-0.05 ***	-0.04 **	-0.01
Entry Frequency	0.11 ***	0.18 ***	-0.00
Uses Password Mgr.	0.03	0.00	0.01
University Password			3.20 ***
$R^2_{GLMM_c}$	0.069	0.320	0.124

Table 4: Multi-level regressions predicting password re-use. The left column is a linear regression where the DV is the number of different websites that a password will be re-used on. The center column is a logistic regression estimating the probability of a *likely correct* password being re-used. The right column is a linear regression where the DV is the number of *non-university* websites that a password will be re-used on. Each regression includes a random effect control for subject.

Using this measure of password *entry frequency*, we ran additional regressions to examine whether subjects re-used frequently-entered passwords (Table 4). We found that more frequently entered passwords are more likely to be re-used—much more likely. For every 9 times a password is entered into a website, that password is used on one additional website. Figure 5 shows a graphical representation of the probability that a password is reused that is generated by the logistic regression in the middle column of Table 4. The frequency that a password is entered (the x-axis) has a much larger effect on reuse than the password strength (the difference between lines). Also, the coefficient on Ranking is very similar in both in Table 4 and Table 3. This suggests that relative password strength within an individual and entry frequency are separate effects: people re-use their stronger passwords, and they also re-use passwords that they enter frequently into websites.

4.2.3 People re-use university passwords

One feature that all of our subjects have in common is that they all have accounts at the same university that they use on a regular basis for accessing email and other university services. This university has a password composition policy (passwords must be at least eight alphanumeric characters long) but does not require users to regularly change their password. It is possible that this commonality across subjects in our sample explains our results: our subjects use a strong password for their university accounts, and have to log into multiple university services frequently, so this is a natural password for them to re-use. However, we can

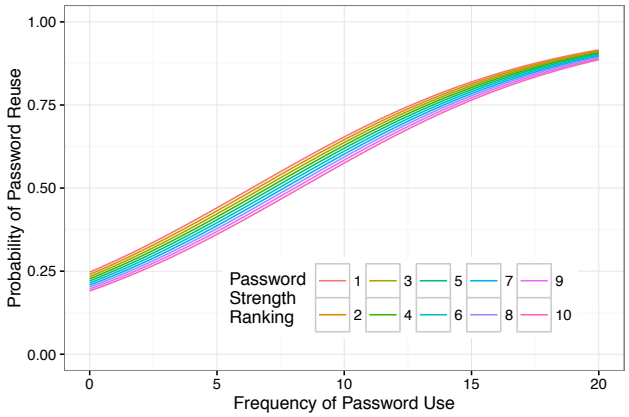


Figure 5: Predicted probability of re-using a password. There is a separate line for each strength of password: a password ranked #1 strongest for that subject, ranked #2 for that subject, etc.. How frequently a password is entered is a more important influence on password reuse than password strength.

test this. We identified all university websites and which password was the subject's likely correct password for those websites. Thirty-five subjects (23.1%) had their university password as their strongest password. Sixty-seven subjects (46.3%) had their university password as their strongest likely correct password. Also, 106 (79%) entered their university password more frequently than any other password.

To understand re-use across non-university accounts, we calculated a dependent variable consisting of the number of non-university websites where each likely correct password was entered. The rightmost column in Table 4 summarizes these results. The university password was heavily re-used across non-university websites; on average across all of our subjects, it was used on 3.2 additional non-university websites. Since for many of our subjects this was one of their strongest passwords, this means that they were using a relatively strong password (which is good), but were re-using a very high-value password (which is bad). The university strongly recommends against doing this; the first piece of advice on their password webpage says "Don't use your [university] ID and password for non-[university] accounts." [5]

4.2.4 *Password managers don't affect re-use*

In understanding re-use, one important consideration is password managers. Using a password manager makes it easier for people to use different passwords on every site because the passwords don't need to be remembered; they are stored by the computer instead. All of our subjects had the potential to use a password manager because both Google Chrome and Mozilla Firefox will save passwords for websites as they are used. Due to API restrictions, we were not able to identify when a password was filled in by the browser's built-in password saving feature.

However, our browser plugin recorded all add-ons such as browser plugins and browser extensions that were installed and/or enabled on each subject's web browser during the study. Manually looking through this list, we found that 26 of our subjects (19%) had a browser-based password manager enabled during the study. We saw six different password managers in use; the most popular password manager was Norton Identity Safe (9 users), followed by SimplePass (7 users).

[5] https://secureit.msu.edu/passwords/index.html, retrieved May 28, 2016.

	# Websites Incorrect
(Intercept)	0.59 ***
Ranking	-0.02 ***
Frequency	-0.01 **
# Websites where correct	0.03 ***
University Password	0.03
$R^2_{LMM_c}$	0.148

Table 5: Mutli-level linear regression predicting the number of websites a password will be *incorrectly* entered into. Each regression includes a random effect control for user.

The regressions in Table 4 include a subject-level variable indicating whether that subject used a third party password manager. A password used by a subject running a password manager was used on about 0.02 more websites than an equivalent password used by a subject without a password-manager, and this difference is not statistically significant. Third-party password managers do not significantly reduce password re-use across websites. However, we cannot tell if this is because many of our subjects are using the password saving features built-in to web browsers (everyone is storing passwords using a different mechanism), or if the subjects with password managers simply aren't using them or aren't using them effectively.

4.2.5 *People guess passwords from their other accounts*

When people forget their password for a website, they often guess passwords that they know they have used. We can learn a lot about what passwords people *think* are appropriate for a website by looking at the password that they incorrectly guessed. When we identified *likely correct* passwords, we separately identified password entries that we are fairly certain are incorrect guesses. We labeled as incorrect any password that was only entered once on a website where other passwords were used more often. We also labeled as incorrect any password entered into a website less than half as often or on less than half as many days as the password we identified as correct. Subjects entered incorrect passwords on 20% of websites.

Table 5 shows the results from a multi-level linear regression predicting how many times a password would be incorrectly guessed. A password that is correctly used on many other websites is more likely to be guessed incorrectly also. Subjects entered their commonly used passwords even into accounts where they were incorrect. Also, higher ranked, and thus stronger, passwords are slightly more likely to be guessed incorrectly. This is further evidence that re-use of stronger passwords on multiple accounts is an intentional strategy. Interestingly, the university password is not more

	Strength		Re-Use	
(Intercept)	40.54	***	2.95	***
"Use good passwords"	1.78	*		
"I use different passwords"			-0.30	*
Uses Password Manager	-1.13		0.29	
University Password	8.54	***	4.68	***
$R^2_{LMM_c}$	0.119		0.226	

Table 6: Muli-level linear regressions looking at the connection between intentions and behavior. The regression on the left examines whether self-reported password strength intentions predict password entropy. The regression on the right examines whether self-reported password re-use predicts the number of websites each of that subject's passwords is used on. Each regression includes a random effect control for subject.

likely to be guessed than any other password after controlling for how often it is used.

4.3 Do people self-report password use accurately?

Self-report questions are typically framed in one of two ways: self-reported intentions (future-oriented), or self-reported actions (past-oriented). However, it isn't clear whether people's self-reports regarding passwords accurately reflect their actual behavior. In a meta-meta-analysis by Sheeran [32], self-reported intentions in general have a 0.6 correlation with behavior across a number of domains. This is a high correlation, which is good; it suggests self-report can be fairly accurate. But Sheeran also found that the strength of the correlation can also vary widely by circumstance, which is why it is important to examine self-report accuracy in different areas to see what people can self-report accurately and what people do not self-report accurately.

In our survey, we included two intention (future-oriented) questions that are directly related to passwords and comparable with our log data:

- Password Strength: "Use good passwords (good passwords include uppercase and lowercase letters, numbers, and symbols)." [Scored Never (1) to Always (5), $M = 4.09$, $SD = 0.92$].
- Password Re-Use: "I use different passwords for different accounts that I have." [Scored Never (1) to Always (5), $M = 2.97$, $SD = 0.96$].

The first question is part of Wash and Rader's protection behaviors scale [39], and directly asks about subject intentions for password strength. The second question is part of the SeBIS behavioral intentions scale [9], and directly asks about subject intentions for password re-use. Twelve subjects did not provide an answer to the SeBIS question; those subjects have been removed from these analyses.

4.3.1 People understand password strength

Our subjects appear to be able to approximately self-report their intentions for using strong passwords. There is a 0.19 correlation between a subject's intentions to use strong passwords and the average entropy of the passwords that person entered during our study ($p = 0.027$). This statistically significant correlation is relatively small for an intention/behavior correlation, but it suggests that people do have some understanding of whether they are choosing stronger passwords.

	Entered Passwords	Correct Passwords
Strongest	0.12	0.11
Weakest	0.07	0.14
Avg by Password	0.19 *	0.19 *
Avg by Website	0.23 **	0.25 **
Avg by Use	0.16 .	0.15 .

Table 7: How well each measure of password strength correlates with a subject's self-reported intention to "Use good passwords". Each number is a Pearson correlation between the self-report measure "Use good passwords" and the indicated password or average of a set of passwords.

Table 6 contains more detailed regression results for the intention/behavior link. The left column uses a multi-level regression to predict a password's entropy using the subject's answer to the self-report survey question about password strength, while controlling for other differences across subjects. On average, a subject that chooses one higher answer on the scale from 'Never' (1) to 'Always' (5) will have passwords that are approximately 1.8 bits stronger. This is approximately equivalent to taking an all-letter password and replacing one letter with a number. This is not a large effect; there is a lot of variation in password strength that is not explained by self-reported intentions. But it is statistically significantly greater than no effect. When people self-report that they intend to use good passwords, their passwords are stronger—but only slightly.

4.3.2 What do people self-report?

When people self-report whether they "Use good passwords", which passwords are they thinking about? They could be thinking about their strongest password when answering this question; alternatively, they could think about their weakest password and evaluate whether they think it is strong. They could imagine all the different passwords they've created and mentally average their strength. They could look at the different websites they have entered passwords on recently and average the strength of the passwords on those websites. Or they could think about all the different times that they had to enter a password, and average the strength of the passwords that they entered. Each of these is a slightly different way of operationalizing which password(s) a person is thinking of when self-reporting.

Our dataset allows us to explore these different interpretations. We can look at different ways of aggregating a subject's passwords and see which aggregation most strongly correlates with a subject's self-reported intention to use strong passwords. Table 7 reports these correlations. The first row examines whether self-reported intention to use strong passwords is correlated with the strongest password each subject

has. The second row looks at the correlation with the weakest password, which is a measure of how strong *all* passwords are. The third row is the correlation with the average entropy if the subject thinks about each distinct password separately. The fourth row represents the correlation with the average strength of the passwords used on each different account. The last row is the correlation if the subject thinks about each time they enter a password and how strong that password is.

Comparing these correlations shows that subjects are not thinking about specific passwords as the strongest or weakest password; instead, when subjects answered this question they likely were thinking across all of the websites that they have accounts on, and looking at the average strength of those passwords. We suspect that when answering this survey question, subjects thought of each website as having a separate password (and thus, a separate choice for that website's password strength), even if he or she re-uses a password on multiple accounts.

4.3.3 People also understand password re-use

Our subjects also seemed to be able to self-report password re-use somewhat accurately. The correlation between a subject's self-reported intention to re-use passwords and their actual re-use (as measured by the ratio of websites-per-correct-password) is -0.12 ($p = 0.18$). This correlation is negative, which is the expected direction; subjects who self-report stronger intentions to use different passwords use each password on *fewer* websites.

Controlling for differences across subjects, we find similar results (Table 6, second column). Using a multi-level linear regression we find that on average, a subject who chooses one level higher on the scale from 'Never' (1) to 'Always' (5) for their intention to use different passwords will use each of their passwords on 0.3 fewer websites. This indicates that a greater intention to use unique passwords is related to less actual password re-use. Though, as with password strength, there is still a lot of unexplained variance.

4.4 Limitations

Our study has several limitations. Our subjects are undergraduate and graduate students at a large midwestern university, and all were from the same university; therefore, our results may not generalize to a wider population. For example, older users tend to select stronger passwords [3]. In addition, the specifics of the university's password composition policy and enforcement of frequent authentication are undoubtedly factors contributing to our results. However, our findings regarding the number of unique passwords and the amount of password re-use are in the same general range as other password studies.

We potentially do not capture all password entry events, either when subjects used private browsing mode, or because a website didn't use a recognized HTML form element. During development, we tested many websites and included special-case code to detect a variety of password forms. We capture password behavior for the majority of websites; for example, we have good data from at least 97 of the 100 most frequently visited websites in our dataset. For ethical reasons, we allowed users to disable data collection which may mean our results do not apply to sensitive online activity.

In addition, our data collection method does not allow us to differentiate between successful and unsuccessful authentication attempts. In other words, we do not know what the true correct password is for any of our subjects' website accounts. This also means that we can't tell if or when our subjects may have changed any of their passwords during the study period. Finally, approximately six weeks of data collection is not enough longitudinal data to make causal claims about these phenomena based on timing or sequence of events, and may have missed passwords entered less often than every six weeks.

5. DISCUSSION AND CONCLUSION

From prior literature, we know that people say they re-use passwords to reduce the difficulty of remembering too many passwords [27]. A median subject in our study used 6 unique passwords that we identified as *likely correct* for the websites they were entered on. While the median password was used on 3 websites, each subject's most re-used password was used on an average of 9 different websites (Figure 4). Subjects tend to re-use passwords that they have to enter frequently, and those passwords tend to be among the user's strongest passwords. In addition, likely correct passwords were also more likely to be entered incorrectly on other accounts, indicating that when subjects attempted to authenticate they naturally tried their "go-to" passwords.

Many studies that have examined password re-use have found that users have a similar number of distinct passwords that they re-use across their websites. Florêncio and Herley [12] found that users averaged 6.5 distinct passwords. Fahl et al. [11] found that people used between 2 and 5 passwords for most of their online accounts. Gaw and Felton's [16] subjects used an average of 3.31 distinct passwords. Stobert and Biddle's [33] subjects reported having between 2 and 20 unique passwords, with a median of 5 passwords. Rinn et al. [29] reported low-literacy subjects used between 1 and 9 unique passwords, with a median of 4. And our subjects mostly used between 4 and 8 passwords with a median of 6. This suggests that there may be a practical constraint that is a hard limit on the number of passwords that most people can remember.

Memorizing strong passwords is difficult for most users to do. Bonneau and Schechter [5] were able to influence 94% of their subjects to memorize a randomly generated 56-bit password by asking them to repeatedly log in 90 times over a period of two weeks with some clever interface manipulations. Logging in with a password frequently is an effective means of memorizing strong passwords. Florêncio and Herley [13] suggest that organizations for which there are no alternatives, such as one's bank, employer, or university, tend to have stronger password composition policies and require users to authenticate more often than websites where use is voluntary (e.g. social media, news websites). These organizations may be helping users memorize stronger passwords by forcing them to choose a long, complex password and enter it frequently. Once memorized, that password can then be re-used elsewhere. This may be what happened in our dataset: the university our subjects are associated with requires fairly strong passwords, and also requires users to enter them frequently.

Among our sample of non-technical users, how frequently a user had to enter a password was one of the strongest

predictors of password re-use. We suspect that once they had a strong password memorized, it was easier to use that password on other websites. This points to an unexpected interdependence between accounts: if users must memorize a strong password on a website where they have to enter it frequently, they then re-use it elsewhere. This results in stronger passwords on more websites. While this practice puts users at greater risk of cross-site password guessing attacks, it helps prevent within-site password guessing by spreading stronger passwords rather than weaker passwords. Since most non-expert users believe that password strength is more important than password re-use [21], it makes sense for them to adopt this strategy. This is evidence that users are trying to adapt their password practices to the security advice they are being told—"Use strong passwords as much as you can."

While people seem to have a mental model of what "stronger" passwords look like [35], our subjects' intentions for using strong passwords and choosing different passwords for each account were only weakly correlated with behavioral measures of password strength and re-use. Responses about password strength intentions were most correlated with the average entropy of the passwords used on each different account, indicating that when people think about what using strong passwords means, each account is considered separately and re-use is not considered.

Bonneau et al. [4] show that entropy is a poor measure of password strength. However, the weak correlation of password entropy with self-reported behavior suggests that when thinking about strong passwords, people think about something similar to complexity (which is what entropy measures).

Our results suggest that asking users about how well they adhere to common password advice from experts asks about password behavior in a way that does not approximate how people actually behave. The ideal situation for security experts would be no re-use: unique, random passwords for every account [7]. Expert advice tends to treat passwords as a black-and-white issue; anything less than the ideal introduces unacceptable vulnerabilities. When the ideal is used as the benchmark, it fails to reflect the reality behind users' choices, and our results speak to the size of the gap between the ideal for security and the realities users face.

Our results show that some amount of re-use might actually be good from a cost/benefit perspective, because if users have a few fairly strong passwords that they use on appropriate categories of sites (e.g., don't use the strong, high-value password on a weaker category of site [14]), they may be more secure than if they used weak passwords everywhere [15]. If the (stronger) university password is used appropriately, then this re-use pattern could lead to a positive effect on overall security. This presents an opportunity for organizations with the ability to force system use (e.g. large employers or universities) to help users memorize stronger passwords by requiring strong passwords and frequent re-authentication. Password composition policies [23] and feedback from password meters [10] can cause people to create stronger passwords than they would otherwise. This might help people use stronger passwords, and is often considered a good security practice by many organizations. However, this practice also puts the organization at greater risk; if the password is re-used on a site with lower security (which is an optimal strategy for some types of websites [28]), an attacker can learn the user's password and use it to compromise the organization. While forcing re-authentication solves one security problem, it creates another: it encourages re-use of the organization's password. Practically speaking, sites that are likely to be compromised need the strongest passwords so they can withstand an offline guessing attack, but users shouldn't have to spend their limited memory capacity and effort creating very strong passwords for sites that are unlikely to be compromised [14].

Unfortunately, defining appropriate categories of websites for re-use of passwords of varying strengths is an open area of research; should it be defined by how much the user values the information [14] or how much an attacker stands to gain [1], or by how much the website invests in security [28]? There isn't a consensus about this, and it seems to be an area of disagreement among researchers. Our study provides insight into how the human and technical constraints imposed on users shape their password choices and behaviors over time, which highlights additional constraints to consider: relative password strength within an individual, and how often the password must be used.

6. ACKNOWLEDGMENTS

We thank Kami Vaniea, Tyler Olson, Nick Saxton, Nathan Klein, Raymond Heldt, Ruchira Ramani, Jallal Elhazzat, Tim Hasselbeck, Shiwani Bisth, Robert Plant Pinto Santos, Meghan Huynh, and Simone Merendi for assistance in developing the software and analyzing the data. This material is based upon work supported by the National Science Foundation under Grant Nos. CNS-1116544 and CNS-1115926.

7. REFERENCES

[1] D. V. Bailey, M. Dürmuth, and C. Paar. Statistics on Password Re-use and Adaptive Strength for Financial Accounts. In *Proceedings of the 9th Conference on Security and Cryptography for Networks (SCN)*, pages 218–235, 2014.

[2] A. Beautement, M. A. Sasse, and M. Wonham. The compliance budget: managing security behaviour in organisations. In *Proceedings of the 2008 workshop on New Security Paradigms Workshop (NSPW)*, 2008.

[3] J. Bonneau. The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, pages 538–552, 2012.

[4] J. Bonneau, C. Herley, P. C. van Oorschot, and F. Stajano. Passwords and the evolution of imperfect authentication. *Communications of the ACM*, 58(7):78–87, June 2015.

[5] J. Bonneau and S. Schechter. Towards reliable storage of 56-bit secrets in human memory. In *Proceedings of the 23rd USENIX Security Symposium*, 2014.

[6] S. Chiasson, A. Forget, E. Stobert, P. C. van Oorschot, and R. Biddle. Multiple password interference in text passwords and click-based graphical passwords. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, pages 500–511, 2009.

[7] A. Das, J. Bonneau, M. Caesar, N. Borisov, and X. Wang. The Tangled Web of Password Reuse. In

Proceedings of the 2014 Network and Distributed System Security Symposium (NDSS), 2014.

[8] G. B. Duggan, H. Johnson, and B. Grawemeyer. Rational security: Modelling everyday password use. *Journal of Human Computer Studies*, 70(6):415–431, 2012.

[9] S. Egelman and E. Peer. Scaling the security wall: Developing a security behavior intentions scale. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 2873–2882, 2015.

[10] S. Egelman, A. Sotirakopoulos, I. Muslukhov, K. Beznosov, and C. Herley. Does my password go up to eleven? The impact of password meters on password selection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 2379–2388, 2013.

[11] S. Fahl, M. Harbach, Y. Acar, and M. Smith. On the ecological validity of a password study. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, 2013.

[12] D. Florêncio and C. Herley. A large-scale study of web password habits. In *Proceedings of the 16th International Conference on World Wide Web (WWW)*, pages 657–666, 2007.

[13] D. Florêncio and C. Herley. Where Do Security Policies Come From? In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, 2010.

[14] D. Florêncio, C. Herley, and P. C. van Oorschot. An administrator's guide to internet password research. In *Proceedings of the 28th USENIX conference on Large Installation System Administration (LISA)*, pages 44–61, 2014.

[15] D. Florêncio, C. Herley, and P. C. van Oorschot. Password Portfolios and the Finite-Effort User: Sustainably Managing Large Numbers of Accounts. In *Proceedings of the 23rd USENIX Security Symposium*, pages 575–590, 2014.

[16] S. Gaw and E. W. Felten. Password management strategies for online accounts. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, pages 44–55, 2006.

[17] B. Grawemeyer and H. Johnson. Using and managing multiple passwords: A week to a view. *Interacting with Computers*, 23(3):256–267, 2011.

[18] S. M. T. Haque, M. Wright, and S. Scielzo. A study of user password strategy for multiple accounts. In *Proceedings of the third ACM conference on Data and Application Security and Privacy (CODASPY)*, pages 173–176, 2013.

[19] E. Hayashi and J. Hong. A diary study of password usage in daily life. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 2627–2630, 2011.

[20] P. G. Inglesant and M. A. Sasse. The true cost of unusable password policies: password use in the wild. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 383–392, 2010.

[21] I. Ion, R. Reeder, and S. Consolvo. "... no one can hack my mind": Comparing Expert and Non-Expert Security Practices. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, 2015.

[22] P. G. Kelley, S. Komanduri, M. L. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, and J. Lopez. Guess Again (and Again and Again): Measuring Password Strength by Simulating Password-Cracking Algorithms. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, pages 523–537, 2012.

[23] S. Komanduri, R. Shay, P. G. Kelley, M. L. Mazurek, L. Bauer, N. Christin, L. F. Cranor, and S. Egelman. Of passwords and people: measuring the effect of password-composition policies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 2595–2604, 2011.

[24] M. L. Mazurek, S. Komanduri, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, P. G. Kelley, R. Shay, and B. Ur. Measuring password guessability for an entire university. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, pages 173–186, 2013.

[25] W. Melicher, B. Ur, S. M. Segreti, S. Komanduri, L. Bauer, N. Christin, and L. F. Cranor. Fast, lean and accurate: Modeling password guessability using neural networks. In *Proceedings of USENIX Security*, 2016.

[26] Nakagawa and Schielzeth. A general and simple method for obtaining r^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142, 2012.

[27] G. Notoatmodjo and C. Thomborson. Passwords and Perceptions. In *Proceedings of the Seventh Australasian Conference on Information Security (AISC)*, pages 71–78, 2009.

[28] S. Preibusch and J. Bonneau. The Password Game: Negative Externalities from Weak Password Practices. In *Proceedings of the Conference on Decision and Game Theory for Security (GameSec)*, pages 192–207. 2010.

[29] C. Rinn, K. Summers, E. Rhodes, J. Virothaisakun, and D. Chisnell. Password Creation Strategies Across High- and Low- Literacy Web Users. In *Proceedings of the 78th ASIS&T Annual Meeting (ASIST)*, 2015.

[30] M. A. Sasse, M. Steves, K. Krol, and D. Chisnell. The Great Authentication Fatigue–And How to Overcome It. In *Proceedings of the Cross-Cultural Design 6th International Conference (CCD)*, pages 228–239, 2014.

[31] R. Shay, S. Komanduri, P. G. Kelley, P. G. Leon, M. L. Mazurek, L. Bauer, N. Christin, and L. F. Cranor. Encountering stronger password requirements: user attitudes and behaviors. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, 2010.

[32] P. Sheeran. Intention-behaviour relations: A conceptual and empirical review. *European Review of Social Psychology*, 12:1–36, 2002.

[33] E. Stobert and R. Biddle. The Password Life Cycle: User Behaviour in Managing Passwords. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, pages 243–255, 2014.

[34] L. Tam, M. Glassman, and M. Vandenwauver. The psychology of password management: a tradeoff between security and convenience. *Behaviour & Information Technology*, 29(3):233–244, 2010.

[35] B. Ur, J. Bees, S. M. Segreti, L. Bauer, N. Christin, and L. F. Cranor. Do Users' Perceptions of Password Security Match Reality? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2016.

[36] B. Ur, F. Noma, J. Bees, S. M. Segreti, R. Shay, L. Bauer, N. Christin, and L. F. Cranor. "I Added '!' at the End to Make It Secure": Observing Password Creation in the Lab. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, pages 123–140, 2015.

[37] E. von Zezschwitz, A. De Luca, and H. Hussman. Survival of the Shortest: A Retrospective Analysis of Influencing Factors on Password Composition. In *Proceedings of Human-Computer Interaction–INTERACT*, pages 460–467, 2013.

[38] K.-P. L. Vu, R. W. Proctor, A. Bhargav-Spantzel, B.-L. B. Tai, J. Cook, and E. Eugene Schultz. Improving password security and memorability to protect personal and organizational information. *International Journal of Human-Computer Studies*, 65(8):744–757, 2007.

[39] R. Wash and E. Rader. Too much knowledge? security beliefs and protective behaviors among us internet users. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, 2015.

[40] R. Wash, E. Rader, K. Vaniea, and M. Rizor. Out of the Loop: How Automated Software Updates Cause Unintended Security Consequences. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, pages 89–104, 2014.

A Study of Authentication in Daily Life

Shrirang Mare
Dartmouth College

Mary Baker
HP Labs, Palo Alto

Jeremy Gummeson
Disney Research, Pittsburgh

ABSTRACT

We report on a wearable digital diary study of 26 participants that explores people's daily authentication behavior across a wide range of targets (phones, PCs, websites, doors, cars, etc.) using a wide range of authenticators (passwords, PINs, physical keys, ID badges, fingerprints, etc.). Our goal is to gain an understanding of how much of a burden different kinds of authentication place on people, so that we can evaluate what kinds of improvements would most benefit them. We found that on average 25% of our participants' authentications employed physical tokens such as car keys, which suggests that token-based authentication, in addition to password authentication, is a worthy area for improvement. We also found that our participants' authentication behavior and opinions about authentication varied greatly, so any particular solution might not please everyone. We observed a surprisingly high (3–12%) false reject rate across many types of authentication. We present the design and implementation of the study itself, since wearable digital diary studies may prove useful for others exploring similar topics of human behavior. Finally, we provide an example use of participants' logs of authentication events as simulation workloads for investigating the possible energy consumption of a "universal authentication" device.

1. INTRODUCTION

Car key, house key, corporate badge, bike key, RSA token, bus pass, credit card, driver's license, ATM card, ... Many of us carry several of these with us every day to access the doors, computers, and services we need (see Figure 1). We also use passwords, PINs, and fingerprints for devices, websites, and applications. These are all *authenticators* – ways to provide evidence that we are the right people to unlock the restricted resources in our lives. We expect people, especially those working in corporate environments, to carry these authentication tokens and remember complex passwords. This burden leads to *frustration* (when we forget our badges, keys, and passwords), *security breaches* (when we tailgate other people through secure doorways, write down our passwords, or leave our devices unlocked), and *IT expense* (when we call help desks to reset passwords or issue new authentication tokens). Password resets make up 10% to 30% of IT helpdesk calls and can cost from $50 to $150 each to resolve [34]. Even physical keys present an increasing risk, as new smartphone apps enable scanning an unattended key in a few seconds and then printing copies of it by mail order or at kiosks [11].

Figure 1: A subset of the authentication material carried by one participant, who also has to manage over 250 passwords.

Evidence and rationale suggests that password authentication can indeed be burdensome for users [5, 15], and experts provide several approaches for addressing this problem, such as using password managers [17], but how about other forms of authentication? If we aim to reduce the authentication burden for users, is it only worth considering passwords, or are physical authenticators like keys also worthy? Answers to additional questions will further help us tackle this area: How much authentication of different kinds do users actually do, and does it correspond to their own concept of the burden they face? How failure-prone are the different kinds of authentication? Do people generally agree about what kinds of authentication they like and dislike, or will it be hard to help the bulk of people in the same way?

To address these questions and gather a better understanding of the user authentication burden, we conducted a wearable digital diary user study of twenty-six people, including teenagers and adults, students, corporate employees, and others. We provided participants with a commercially avail-

This work was performed while all three authors were at HP Labs.

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22–24, 2016, Denver, Colorado.

able wrist wearable, the MOTOACTV [24] running our own logging application, and asked participants to log all their authentication events for a week. We used the wrist wearable because there are typically so many required authentication events during a day that we wanted participants to be able to log events in the moment, rather than try to remember what they did later. We designed a "slot machine" application interface to provide the wearable with an immediately available and streamlined logging process to help reduce the amount of under-reporting to which diary studies (including ours) are susceptible [18]. We also applied this logging approach because we are interested in authentication with physical infrastructure and not just online authentication, and we could not simply instrument all of the participants' targeted resources to log authentication events automatically. The product of the study includes 4,623 hours of logged events, interviews of each participant before and after the week of logging, and comments participants entered through daily surveys on their smart phones. Our results thus include quantitative information in the form of "traces" of user authentication behavior as well as qualitative information in the form of participants' opinions. Twenty-six participants is not a large enough population to make broad claims about the general population or any particular demographic, but it allows for close observation of a diversity of authentication behavior and opinions.

Our contributions are threefold. First, we hope the design of the wearable digital diary may be interesting to others performing similar studies, even though we believe this particular study suffers from several flaws such as the small sample size of participants. Second, our results incorporate information that might help others working to reduce people's authentication overheads. We find, for instance, that authentication using physical tokens is a sufficient burden to warrant addressing. On average, 25% of a participant's authentications employ physical authenticators – tokens such as car keys that users need to carry with them – and participants offered negative opinions about physical authenticators, not just passwords. We also find that people's authentication behavior and their opinions about authentication vary greatly, so it may be hard to please everyone in the same way. For instance, some people's favorite authenticators are others' least favorite, although several participants favor quick, effortless authentication methods even if they come with significant error (false reject) rates. We see surprisingly high failure rates across many types of authentication: 5% for passwords (with an even higher rate for PCs and websites), 3% for physical keys, and 3% for fingerprints. Our third contribution consists of the authentication event logs themselves, which we will make publicly available. We provide an example of how we use the logs as workloads for simulating energy consumption of a "universal authenticator" – a device that performs many varieties of authentication on behalf of its user.

2. RELATED WORK

There is a tremendous amount of existing and ongoing work related to ours, especially in the areas of authentication, user studies, and wearables. We confine our descriptions of related work to examples of user authentication behavioral studies that we perceive to be the most relevant to ours. We note that even definitions of authentication devices differ across these studies, with some including the presentation of "things you have" such as keyfobs, and others only including tokens that display or contain information specific to a single individual, such as a badge with the owner's photograph [29].

Several studies focus on smartphones and how people choose to secure them or not, and their results vary considerably. Based on 2,000 Android users' smartphone usage Hintze et al. report that on average people unlock their phones 25 times per day [16], whereas Harbach et al. find an average of 47 unlocks per day in their 52-participant study [13]. In our study we observed unlock usage of about 33 times per day. A 2013 study by Lookout [19] of 1,003 Americans (age 18 and older) found that 56% of users surveyed did not choose to enable a security lock for their phones, and that "people care [about privacy] but exhibit risky behavior." Other studies see fewer people choosing not to lock their phones [20]. Egelman et al. report that 8 of their 28 interviewed participants (29%) and 42% of their 2,418-person online questionnaire respondents did not lock their phones [9]. Bruggen et al. observed that 35% of phones out of the 149 running their software agent did not employ any locking mechanism [32]. In our study, 4 of 26 participants (15%) did not lock their phones, and we too observed risky authentication behavior in terms of password management and sharing.

Various other non-smartphone studies and essays explore passwords and how users manage, choose, and forget them [5, 10]. A study of the password habits of half a million users over a 3-month period used a component in Windows Live Toolbar on users' machines to record password strength, usage, and frequency metrics [10]. The study found users choose weak passwords and use them across multiple sites and that 4.28% of Yahoo users forgot them during the study. We see an even higher percentage of users who forget, struggle to remember, or reset a password at least once during our study (36%). Hayashi and Hong conducted a diary study with twenty-one participants, in which participants carried diaries and recorded information about password-based authentications, but the focus of the study was authentications only on laptops and desktops [14]. A *New York Times* study explores the meaningful personal information users embed in their choice of passwords [31]. All of these studies agree with ours in concluding that users find it hard and frustrating to manage passwords according to established rules of safety. Usable security that takes into account human limitations and strengths has become increasingly important [6].

A recent study of online safety covers opinions and practices of both experts and non-experts regarding how to stay safe online [17]. It is interesting to note that the reported expert security advice on password management differs somewhat from the requirements promoted by some of the participants' companies' IT departments. In particular, at least one IT department asks employees not to trust their passwords to third-party password managers, and yet it does not provide any in-house password manager. Experts promote the use of password managers, while non-experts surveyed by the study shared the IT department's distrust of password managers.

At least two studies include consideration of authentication other than with phones and passwords. A National Institute of Standards and Technology (NIST) study involved 23 NIST employees (ages 20 and above) carrying a written diary in which they recorded a wealth of information about their authentication events for a 24-hour period [29]. This study

covers not just smartphones, passwords, or online authentication behavior but also a few other types of devices such as badges. Their participants recorded an average of 23 events a day, which is significantly lower than the 45 average of our participants. This may be because of differences in the study sample (a majority of their participants were in their fifties, whereas the median age of our participants is 29 years) or differences in the event logging mechanism (a paper diary vs. a digital wearable diary). Some other results from their study correlate well with ours, such as finding no strong relationship between participants' amount of authentication and the frustration they express. Another study that considers physical authentication was performed by Bauer et al. in 2007, in which they instrumented doors at participants' workplace(s) for authentication using smartphones, and developed (and evaluated) access-control policies for unlocking those doors [4]. We are interested in all physical authentications in participants' daily lives, which ruled out instrumenting things for automatic authentication logging, leading us to use a self-reporting approach with a wearable digital diary.

We believe our study is unique in two ways. First, we enable the diary study with a wearable application to allow easier and more streamlined in-the-moment logging of authentication events. Our motivation is to reduce under-reporting and provide more accurate timing information for authentication events. Second, our study covers a wider range of authentication types, including authentication with locked cars, doors, bicycles, public transportation, and so forth. While there are studies that report on authentication with a few types of physical targets, we are unaware of a study that covers the breadth of physical authentication targets accessed by the participants in our study.

3. METHODOLOGY

In this section we define an authentication event and describe the wearable digital diary method for self logging and our study procedure.

3.1 Authentication event

We define an authentication event as one where an individual must demonstrate, actively, that he is the right person to gain access to a resource or service through something he *is* (or *does*), something he *knows*, or something he *has*. Examples include unlocking a phone, unlocking a house door, logging in to a password-protected website, or entering a PIN on an ATM machine. Accessing a website with cached credentials that does not even require a mouse click to choose among credentials involves no active user effort, so it does not count as an authentication event for our purposes, since we want to explore in-the-moment user authentication effort. Note that we also do not include lock or re-lock events. We define an *authentication target* as the device, resource, or service to which the individual requests access, and an *authenticator* as the evidence the individual provides to gain access. For example, when unlocking a phone with a PIN, the phone is the authentication target and the PIN is the authenticator; when opening a door with a badge, the door is the authentication target and the badge is the authenticator. Below is the list of targets and authenticators we use in the study. Note that some of the items represent a category. For instance, "Laptop" also covers desktop computers, while "Password" also covers passcodes, PINs, locker combinations or any knowledge-based authenticator.

Figure 2: Diary entry app on the smartwatch.

Authentication Targets: Laptop (also desktops), Phone, Tablet (also e-readers), Website (also online websites or any software), Door, Car, ATM, Public Transport, Bicycle (also motorcycle), Phone payment, Card payment, Bank check, Locker (also locked drawers), and Other.

Authenticators: Password (also PINs, locker combinations, etc.), Fingerprint, Face biometrics, Voice biometrics, Card (ID cards, credit cards, badges), Certificate (PKI), Mouse click (where the participant has to click to authenticate, e.g., to request autofill with a password manager), Lock key (physical key), Keyfob (remote key), Signature, 2-Factor, and Other.

We are also interested in whether an authentication succeeds and the location where it occurs. We asked our participants to log whether the authentication event was successful and the number of required attempts before it was successful. We wanted to collect semantic locations for authentication events, including Home, Work (includes School for student participants), Shop, Traveling, and Other. Thus, in our study an authentication event is represented as {*event-time, authentication-target, authenticator, success, location-label*}.

3.2 Wearable digital diary

To reduce the amount of under-reporting and poor recall that can affect diary studies [18], we wanted to enable immediate, easy logging of events. This is especially important for events such as authentication that can occur frequently and at times when it is inconvenient to pull out a paper diary and pen or even pull out a smartphone to bring up an app. We considered using a wearable voice recording device, but pilot study participants said they would not be happy talking to themselves when unlocking stuff. We chose a smartwatch (the commercially available Motorola MOTOACTV [24] Android smartwatch) as our primary logging device, as it is easily accessible and we could take over the display with our logging app for immediate entry; Figure 2 shows the logging interface available as a user raises his wrist. Indeed, most of our participants found logging events via the watch convenient compared to a smartphone; we further describe this in Section 3.3.1. Besides the logging app on the smartwatch, we also developed a companion smartphone app, where participants could view, edit, label, and comment upon their logged events using the bigger display.

3.2.1 Watch app

The MOTOACTV is not programmable out of the box. To use it as a digital diary we rooted the watch and installed our Android application, which always runs in the foreground so

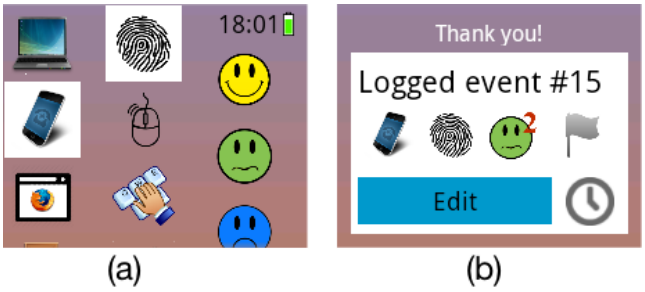


Figure 3: Watch App UI. a) Main watch app screen showing logging in progress for a phone event using a fingerprint unlock. b) Watch app screen confirming the logged event; the green icon with number two indicates that two retries were required to unlock the phone.

that it is immediately accessible to participants when they raise their wrists, which turns on the display (Figure 2). For an authentication event, we want to collect the authentication target, the authenticator, success or failure, number of attempts required (in case of a success), location, and time. The app automatically collects GPS location and time, but the participants have to log the other four details. Logging an event should be quick and easy so that it is less interruptive to the participants' current tasks, otherwise they are likely to delay logging and may later forget to do so. After several iterations and feedback from pilot study participants we came up with a novel "slot machine" like interface to log an authentication event with usually only two taps on the watch touchscreen. Figure 3 shows the logging interface. Participants generally liked the watch app interface: eight participants mentioned unprompted during their post-logging interviews that it was easy for them to log events through the watch. Participant P5 added that *"anything more than 2–3 taps is effort for me"*. Some participants did not like the watch's form factor: six participants wished the MOTOACTV watch had been smaller or more comfortable, and one participant chose not to wear or carry the watch and entered all his events from his phone.

Figure 3a shows the app logging screen. It presents three vertically scrollable columns of icons: the first column for authentication targets, the second for authenticators, and the third for success/failure. In Figure 3a the participant has selected phone (target) and fingerprint (authenticator). The success/failure column has three icons: (from top to bottom) a yellow happy face (for immediately successful authentication), a green unhappy face (for successful authentication that required more than one attempt), and a blue sad face (for a failed authentication or extremely problematic event). Examples of failed events include forgetting one's password or dropping one's car keys in the mud under the car. Tapping a face icon enters (logs) the event, except for the unhappy face (middle icon), which brings forth another column on the right side of the display. This fourth column contains a list of numbers (2, 3, 4, and 5+) indicating attempts performed for the successful authentication. The order of icons in each column is user-configurable for convenience, so participants can keep their most-often used targets and authenticators at the top for quick access. The app also caches the last chosen authenticator for each target and automatically selects it when the participant chooses a target, to reduce necessary taps in the common case. For example, choosing phone would automatically select fingerprint if the participant's last logged phone unlock event was with a fingerprint. Tapping the happy face would then log the event. With caching and configurable icon order, participants can log events with only two taps for their common cases.

The act of choosing a face icon enters the event and brings up a confirmation screen. Figure 3b shows a confirmation screen for a phone unlock event with a fingerprint in two attempts. The confirmation screen shows the authentication event logged and allows editing the event. It also allows flagging the event (flag icon) or adjusting the time of the event (clock icon) in case the event was performed in the past (e.g., 10 min ago). We asked participants to flag an event when there was something unusual about it or if they wanted to comment on it, which they could do on their smartphones when reviewing their event logs. We inquired about flagged events and any other odd events during the post-logging interviews. The confirmation screen persists long enough to allow users to edit the event if they wish and then returns to the logging interface.

3.2.2 *Smartphone app*

The watch allows participants to log an authentication event quickly without needing to reach for their phones, but its small screen size is not suitable for complex interactions such as viewing event logs or editing events in the log, so we provided participants with a companion smartphone app. The smartphone app periodically syncs with the watch and administers the daily survey at the participant's chosen time, usually in the evenings. The app also periodically syncs with the cloud, allowing us to monitor the study. The smartphone app provides a dashboard interface for participants where they can also manually sync the phone with their MOTOACTV watch, sync the app with the cloud, browse and edit their authentication logs, and take the daily survey.

Figure 4a shows an example of the event log UI, reachable from the dashboard or the daily survey. Each row represents an authentication event, with the time of the event displayed on the left, followed by the authentication target icon, the authenticator icon, an authentication success/fail icon, a comment icon (orange if the participant entered any comment for the event), a flag icon, and a location label. Tapping on any of these icons allows the participant to edit the field. An unassigned location label appears as "NA" and participants can tap on it to assign a label from a pop-down menu of five location labels (Home, Work, Shop, Travel, and Other). When a participant labels an event, the app automatically labels other events logged at the same location. Participant labeled events are orange; in the figure the top and bottom location labels were assigned by the participant and the other labels were assigned by the app. Although we chose the MOTOACTV watch in part due to its built-in GPS sensor, the GPS on the watch could not always provide a location, so the smartphone app collects GPS information every five minutes, and we also use this information to assign semantic location labels to events. When participants finished the study, we deleted the GPS information to keep only the semantic labels, as they are far less privacy-sensitive.

Figure 4b shows the survey we administer daily to the participants. In the survey we ask participants to go over the

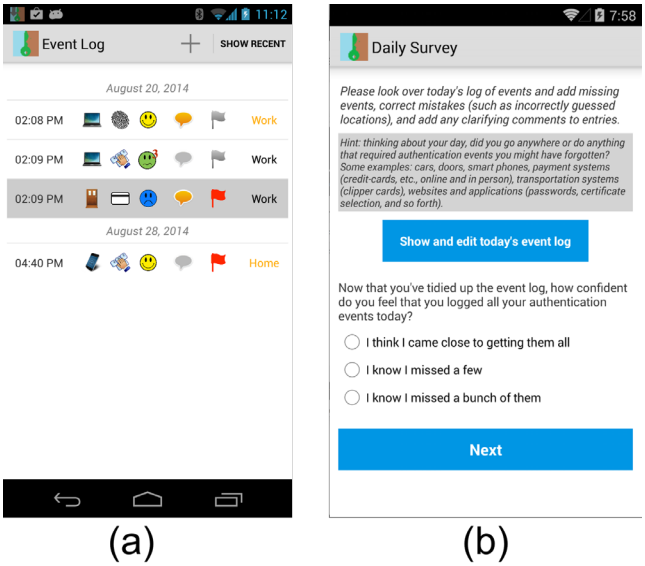


Figure 4: Phone App UI a) event log, b) end-of-day survey.

day's authentication event logs, add missing events, make any edits if necessary, add comments to events, and add location labels to the authentication events. To make it easier for participants to review their logs, the app only displays events that the participant logged since the last time they took the survey. Participants can see the complete log by choosing the 'show all events' option. In the survey we also ask them to rate how good they thought they were about logging all the events. On the Next screen in the survey we ask them to provide any comments they had, about the study or about their day, especially if they felt there was something unusual about the day.

3.3 Methodology successes and failures

Other researchers might be interested in deploying a wearable digital diary study for their own purposes, so we describe here the high-level successes and failures of our study methodology. We list other limitations of the implementation of our study in Section 5.

3.3.1 *Logging on the wrist versus the phone*

One of the reasons we created the phone app as well as the wearable app is that we worried many participants might prefer to log entries from their phones. After all, many people have their phones handy most of the time. We gave participants the choice of logging either from phone or watch. However, the immediate accessibility of the wrist wearable combined with our slot-machine style logging interface worked as intended. Except for three people, participants logged an average of 93% of their events on their watches. One participant (P25) did not like wearing a watch so he logged all events from his phone, and two other participants (P15 and P16) did not wear the watch because they thought it was not fashionable. Instead, they carried it clipped to their bags and logged about 40% of their events on the watch. Overall, the approach made logging easy enough that 84% of events in our study were logged from the wrist wearable despite the availability of the phone application. We suspect that this approach could lead to future wearable digital diary studies. The smartwatch was generally the preferred platform for logging in-the-moment events compared to the smartphone among our participants, and despite the clunkiness of the particular watch we used, one participant became a convert to watch use in general: *"I didn't used to wear a watch. I didn't think I liked them. But after this study I got used to just looking at my wrist and knowing what time it is. Now I want a watch." (P10)*

3.3.2 *Validity of self-reported phone events*

We captured phone unlock events in two ways: our phone app automatically logged phone unlock events (except for the five iOS users and a user whose phone was unable to do so), and a set of participants logged phone unlocks manually (including all the iOS users and a subset of the other users). Using the eight-person intersection of these groups, we compared their number of automatically and manually logged phone unlock events to get an estimate of under-reporting for phone unlock events. Under-reporting phone unlocks ranged from 7.8% to as high as 60.1% for one participant. We believe that participant decided not to worry about logging phone events, but since she did not explicitly inform us of this decision we count her data. On average we see 20.9% under-reporting, although one user over-reported by 31.9%. When queried, the over-reporter said that he was worried that phone unlocking was so automatic that he might not have recorded it so he would record it again just in case.

Automatic logging would be much better for accurately recording activities that involve many events, but where that is not possible, such as our case in which we cannot instrument all possible authentication targets, it is clear we must streamline the logging process however possible. Attempting to recall and record all authentication events after the fact seems close to hopeless. Some participants expressed a difference of feeling about logging phone events as compared to other events, saying that they were harder to recognize in the moment compared to other types of authentication and that they therefore had more trouble remembering to log them. Some participants either declined to log them or gave up logging them. These included our biggest users of phone unlocking, according to the automatically logged events. If compliance is inversely proportional to number of events, our participants' self-logging of event types other than phone unlocks may suffer from less under-reporting, but we have no good way to determine this. In addition, this makes comparisons between phone unlock and other authentication events less reliable.

3.4 Study procedure

We performed a 3-person 2-day pilot study to test the digital diary for logging, for our categorization of the authentication targets and authenticators, and to find bugs and refine our UI and procedure. We then executed the main study which we describe below.

3.4.1 *Recruitment and enrollment*

We recruited participants by word-of-mouth because our company's legal department required us to verify that participants be either affiliated with our company or US citizens at least indirectly known to members of our organization. These conditions also soothed management concerns that we be able to retrieve the smartwatches from participants after the study. We additionally screened participants to verify that they were comfortable using a smartphone. We provided

informed consent and information sheets to screened participants. If participants agreed with the documents, we invited them to come in person to our lab (or meet via Skype for remote participants) where they signed the consent form and we interviewed them and explained the study procedure. We required and received parental consent for participants under the age of 18. Enrollments occurred throughout the week, and participants were asked to perform the study for seven days from the enrollment day. We explained both in writing and in person what information we would collect. We also warned participants both in writing and in person to practice safe logging: "Please only log events on the watch and phone when it is safe to do so. Please do not use the devices while driving, biking, crossing streets, operating heavy machinery, or anything else that would be risky!" We also informed them that they could withdraw from the study at any time for any or no reason. One person did so, leaving 26 participants.

We gave each participant a MOTOACTV smartwatch with our app pre-installed and asked them to wear it on their wrist at all times (except when charging or in the shower or pool; the watch is not waterproof), but if they were uncomfortable wearing it on their wrists, they were allowed to attach it elsewhere via a provided clip. We installed the companion smartphone app on participants' Android smartphones. If a participant did not have an Android smartphone, we lent one for survey taking and syncing and editing event logs. Our study was approved by the ethics committee equivalent in our company. Study participants received $100 gift cards upon completion of the study.

3.4.2 Pre-logging interview

We conducted an in-person semi-structured interview with each participant to learn about their own pre-study perspectives on their daily authentication lives, the devices and resources they use, the authenticators they carry with them, and how they feel about various aspects of authentication. We asked participants to tell us about the authentication events they perform in a typical day by thinking through their daily routines and recalling their authentications. We also asked them to guess how often they might authenticate with various resources so that we could compare this information later with their reported data. We used a set of questions to guide these semi-structured interviews, but we allowed the participants to digress and describe their opinions and behaviors regarding authentication. See Appendix B for the list of interview questions. We answered any questions they had about how to enter various kinds of events.

3.4.3 Post-logging interview

We conducted another semi-structured interview with each participant after one week of self-logging. We asked them about flagged events, any logged entries that we did not understand, and about authentication failures they logged. We also asked about their thoughts on authentication, the watch UI, future inventions they would like to see in the area of authentication, their choice of best and worst authenticators, and about how their authentication behavior might have changed during (or as a result of) the study. See Appendix C for the list of questions.

4. USER STUDY PARTICIPANTS

The study includes 26 participants who logged their authentication events for one week each over the course of three months. Due to the conditions placed on our recruiting of subjects, our participants essentially form a "convenience sample" that is not as balanced across gender and other characteristics as we would have liked. We were able to aim for inclusion rather than balance. Participants' ages range from 13 to 64, with 7 participants each in age ranges 10 to 19 years and 20 to 29 years range, 8 participants in age range 30 to 49 years, and the remaining 4 participants in age range 50 to 64 years. The participants include 8 females and 18 males. Sixteen are from computer-related fields, 2 are from non-technical fields, one is from a medical-related field, and 7 are in grade-school. There are 10 students (3 are graduate students), and the rest are full-time employees. Our participants represent 7 different schools, 4 different companies, and 3 different regions of the US. Participants self-reported their ethnicities as 14 Caucasian, 2 African American, 7 East Indian, and 3 Asian.

5. LIMITATIONS

The goal of our study is to gain an understanding of authentication in participants' daily lives through self-reported quantitative data and qualitative interview data. Due to the nature of the data we obtained, the results should be interpreted carefully. We should avoid generalizing numerical results to a broader population, due to both the small number of participants and under-reporting of self-logged events. Instead, we can use the results to learn about authentication habits and the reasons behind them. With that in mind, we list the limitations of our study.

L1 *Small sample size.* Regardless of participant diversity, our convenience sample of twenty-six people is not a large enough group to give good statistics about the overall population or any particular demographic. We caution readers against generalizing the results.

L2 *Under-reporting.* We minimize the effort for reporting an event in our study, but it is not a zero-effort task, and participants failed to report some events, except for one participant who over-reported phone unlock events. Thus the number of self-reported authentication events in our study is generally a lower-bound of the actual number of authentication events performed by the participants. Moreover, whether a participant self-reports an event might be affected by context (e.g., current activity, location).

L3 *Self-logged vs. auto-logged data.* We asked some participants to report all authentication events, including phone unlock events, but our smartphone app also automatically logged phone authentication events. In our analysis (Section 6) the phone authentication events are from the automatically logged data for Android users (except one) and self-logged data for iOS users. The other (non-phone) authentication events are from participants' self-logged data. This exaggerates any differences between phone and other authentication events, which we should keep in mind when analyzing the results.

L4 *A snapshot of a week.* The data we obtained is a snapshot of authentications that participants encountered and reported during one week, which is not necessarily representative of their typical weeks. For instance,

there are most likely cases where a participant did not perform a type of authentication (e.g., using an ATM) during our study week that he might have performed during another week.

L5 *Participants with only daytime jobs.* None of our participants are night workers – they are all students or employees with daytime jobs – so we see only a few events at night.

L6 *Mostly Android participants.* Only five of our participants are iOS phone users with the rest being Android users. With more iOS users, for instance, we might see more fingerprint authentication, as fingerprint readers are less common on Android phones.

L7 *Missing location information.* GPS readings are not always reliable depending on location (for instance, indoors) and some participants' phones had more trouble getting frequent GPS readings than others'. As a result, about 25% of locations in the study have no semantic labels attached.

L8 *Extra phone unlock events.* There may be extra phone unlock events caused by the study, because participants might unlock their phones to take the daily survey or to log an event on the phone app. When queried, participants said they did not access their phones just for the survey or to log an event, but we have no way to verify this claim.

L9 *Change in participants' authentication behavior.* Participants may have changed their authentication behavior as a result of their participation in the study. One of our post-logging interview questions targets this concern. Three participants said they typed passwords more slowly to avoid errors, and one said he used his phone less often on some days, because he was embarrassed by how frequently he used it. Otherwise, all participants said they did not notice any change in their authentication behavior, but we have no way to verify their claims.

6. FINDINGS

In this section we present our findings from both logged events and participant interviews. We look at authentication patterns, the nature of the authentication burden on participants, and the rates of authentication errors participants experience. We see evidence that physical authenticators are part of the authentication problem for many people, and not just passwords. We find that people's authentication behavior and opinions vary greatly, and that many types of authentication suffer from high false reject rates. We report supplemental material about participants' estimates of how much authentication they do, their feelings about privacy, and further results about their authentication patterns in the Appendix.

Together, participants logged 7,225 authentication events: they manually logged 3,488 authentication events, and our phone app automatically logged 3,737 phone unlock authentication events. The log for one of our participants did not cover quite the full week; for calculations where this could affect results we use only data from 25 participants. We conducted semi-structured interviews with all participants before and after the data logging phase of the study. We took detailed notes from the interviews. Our notes include direct quotes from participants, summaries and paraphrases of participants' explanations, and descriptions about their authentication behavior and opinions. We identified themes and categories in our notes (coded) and formed a data matrix, with columns as themes and rows as participants [23]. As we describe our findings we include occasional quotations from participant interviews (with more in Appendix F) chosen because we found them especially interesting, representative of a particular point, or simply entertaining.

Table 1: Authentication targets and authenticators used in the study and the number of participants (N) who used them.

Targets	N	Authenticators	N
Laptop	26	Password	26
Website	26	Card	25
Door	24	Lock key	25
Phone	22	Keyfob	18
Car	20	Mouse click	15
Card Payment	20	Signature	13
Other	11	Fingerprint	6
Tablet	10	Other	8
Locker	9	Certificate	4
Bicycle	7	2-Factor	3
ATM	5		
Check	5		
Public Transport	3		
ID Verification	3		
Phone Payment	2		

6.1 Authentication patterns

We captured the different types of authentications that our participants performed, how often and where they authenticate, and various other characteristics. Table 1 presents a list of authentication targets and authenticators logged in the study and the number of participants who used them. Some targets and authenticators were very popular, but authentication behavior varied even at this high level. For instance, two participants, both teenagers, did not need to unlock doors during the study.

6.1.1 Distribution of events by authenticator

Overall we find that 74.4% of authentication events involve "things you know" (secrets such as passwords, PINs, swipe gestures, and locker combinations), 18.4% involve "things you have" (physical token-based authenticators such as badges, keys, cards, keyfobs, and 2-factor tokens), and 7.2% use other means, including biometrics and signatures, or "things you are or do."

Figure 5 shows the distribution of authentication events logged by each participant by category of authentication, secrets, physical tokens, and "Other." On average, 25% of a participant's authentications used a physical token for an authenticator. If we exclude the four participants who did not lock their phones, this number falls to 21%. Authenticating with keys and other physical tokens constitutes a significant part of most participants' authentication workload. There is high variance, though, as some participants performed almost no authentication with physical tokens.

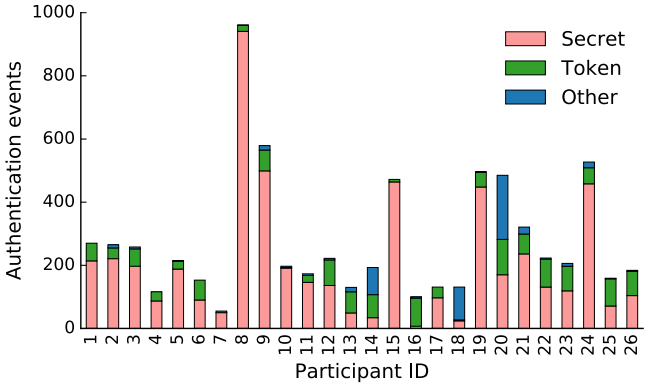


Figure 5: Authentication events for each participant, categorized by type of authenticator.

This is largely true for the teenage participants (P7, P8, P10, P15, and P19). The teenager P16 performed mostly token-based authentication, because he drives a car but does not lock his phone.

6.1.2 Distribution of events by target

We were also interested in learning how many authentication events involve digital versus physical targets. *Digital* events are those that require authentication to an electronic service or an electronic personal device, and *physical* events are those that require authentication to a physical resource or thing. Specifically, in our study, digital targets include Phone, Laptop, Website, or Tablet, and we consider all other targets physical. See Section 3.1 for a fuller definition of these targets. We debated using other possible categorizations, such as considering devices like phones and laptops to be physical infrastructure instead of digital targets. We use our current categorization so as not to overemphasize the importance of physical targets.

Among all the logged authentication events, 22.2% were physical authentication events. The average number of physical authentication events logged by each participant is 30.7% with a standard deviation of 20.2%. Again, we see substantial variations across participants, in part because of their widely varying ages. Middle-schoolers, for instance, do not need to unlock cars as often as adults, and most of them do not have credit cards. Alas, most of our adult participants drive cars more often than they bicycle.

6.1.3 Variation in authentication pattern

We also capture when, how often, and where participants authenticate themselves. Overall there is a high variance among participants. Authentication events per day across participants range from 0 to 208 with an average of 45 per day. Even in our day and age it is possible to have a day of zero authentications if you do not lock your phone and stay indoors the whole day. Authentication events per hour across participants in a 24 hr day range from 0 to 107 with an average of 2. Authentication events per hour across participants in a 9am–5pm day range from 0 to 83 with an average of 19.

We were interested in learning whether participants log more events on weekdays than on weekends, and when during the day they typically authenticate. We see no obvious distinction; only five participants (P3, P5, P6, P21, and P24) performed significantly more authentication events during a weekend than on a weekday ($p < 0.05$). We also analyzed the number of authentication events performed by participants at different hours of the day. All our participants have day jobs or generally follow a day-oriented schedule, and so we see more authentications between 9am–6pm, but there were authentications spread throughout the 24 hr day. There seems to be no hour of the day where someone isn't authenticating with something.

Table 2 shows the number of authentication events performed at different locations. We expected to see most events occur at Work (where school counts as work for students), but we were wrong. Home receives the largest number of authentications when averaged across all participants, and if we consider just phone unlock events, we see that participants unlock their phones 59.8% of the time when they are home and about 29.7% when they are at work. However, if we exclude teenagers we see that participants perform more authentication at work than at home (45% vs. 40%). For the overall participant pool there are roughly 10% fewer authentications on average at Work, with Shopping (which includes restaurants), Traveling and Other receiving far fewer events. Traveling includes driving, and unfortunately, we do see participants unlocking their phones while driving, as have others [19].

Table 2: Distribution of authentication events by location, across all participants and across participants excluding teenagers.

	All	Excluding teenagers
Home	43.6 %	40.1 %
Work	38.5 %	45.1 %
Shop	6.6 %	5.8 %
Travel	5.5 %	4.1 %
Other	5.8 %	4.8 %

Variation across age and gender. We see a slightly higher number of authentication events in teenagers and older participants (> 39 years) than those in their twenties and thirties, but we believe there are no general conclusions to draw from this and that it is likely due to individual behavior. We can, however, conclude that no participant escapes authentication.

We also compared authentication behavior between the 8 female and 18 male participants. Per person, both groups logged roughly the same number of authentication events, phone unlock events, and physical events. The average authentication events in a day logged by the female group and the male group were 42 (± 16) and 37 (± 33), respectively. The average number of phone unlock events in a day logged by the female group and the male group were 25 (± 23) and 22 (± 33), respectively. The high standard deviation highlights the wide variation in the study participants' authentication behavior. Overall, at least in our small sample size, we do not observe wildly different authentication behavior across gender.

6.2 Authentication burden

In this section we look further at whether, and in what ways, participants consider authentication a burden. We find that

Table 3: The number of authenticators carried by participants, added across all participants.

Authenticator	N	Comment
Credit card	60	Includes work, 4 not used
Loyalty/gift card	55	One gave no exact number
House/apartment key	27	One person carried 6
Membership card	22	
Car key	19	Regular or electric
Driver's license	18	
Other door key	17	One gave no exact number
Debit card	17	
Other ID card	16	2 expired
ID badge	15	Mostly corporate, also gym
Car fob	13	
Health insurance card	12	One noticed missing card
Transportation	10	Zip card/buses/metro
Car proof of insurance	9	Others kept these in car
Mail box key	7	One P.O. Box
Key of unknown function	7	
Scraps of paper	5	With writing
Bike key	4	Rest used combinations
Phone	4	Phone app for passwords
Motorcycle/scooter keys	4	Includes 2 trunk keys
Digital key	3	
Locker key	2	Rest used combinations
Blank checks	2	
Cabinet/drawer key	1	One attached to ID badge
Motorola skip	1	
Jewelry box key	1	
House alarm fob	1	
Work building fob	1	In lieu of ID badge

participants' opinions vary considerably, and that managing both "things you have" and "things you know" contribute to the burden.

6.2.1 Things people carry

While many problems with passwords are well documented, physical authenticators also offer challenges for users. Some of us have many resources we need to access frequently using physical authenticators, and this means we need to carry many authenticators with us. To find out more about this potential problem, we asked participants if they were willing to dump out the contents of their wallets, pockets, purses, bags, or other places where they carry authentication material. We told them we did not need to see what they dumped out, but that they could just enumerate for us what they found. Table 3 shows the results, added across participants.

There are several indicators that managing these carried authenticators can be troublesome. *"I don't like to carry around physical keys. It's just another thing to manage, and if I were to ever forget it...The Pebble is one exception 'cause it's always on your wrist. If it had a computer unlock I'd be totally happy." (P7)* Several participants attempted to divide up or stage their authentication material so that they did not need to carry all of it. For instance, one participant has bags for different purposes, with the appropriate ID cards or badges in the different bags. Another attaches a work cabinet key to her ID badge, and that key opens drawers with other cabinet keys in them. Another participant uses a phone cover with slots for cards in it. He carries his driver's license, a debit card, and his badge in the cover. The rest of his cards he puts in his wallet, which he keeps in his car and only carries on his person if he needs it in a store. Another participant stages his keys so that he carries a minimum but the keys he does carry allow him access to the rest of the authenticators. *"I'm at the limit of physical keys I can carry – can't tolerate any more. It's a layer system – the rest are kept in a pie tin at home. It's part of the family semaphoring system. Know who is doing what where...I have it set up usually so most things are automated and I don't have to carry as much. Never be without a house key – I teach all my kids that too." (P12)* Another participant rigged up his own "smartwatch" in the form of a Motorola skip clipped to his regular analog watch. He unlocks his phone by tapping it against the skip on his watch. Attaching it to the watch means he does not need to worry about carrying it – it is always with him since he wears his watch every day.

Another indicator of management burden is that people can't track what they carry. They carry authenticators with them that they no longer need or cannot identify. People carried expired school IDs, unused credit cards, and keys whose functions they couldn't remember. For instance, one carried two unidentified keys and said *"But I'm scared to remove them. They seem like they might have been important." (P21)* They also can't find material they were sure they were carrying. *"There should be two health cards – one for kids – but I can't see where that went." (P26)*

Some participants also make arrangements to carry authenticators on behalf of others. One teenager (P7) carries his brother's gym ID card "'cause he doesn't carry a wallet. We go together and my parents are worried he'd lose it." Another carries his own locker key and his friend's. Another carries his friend's house key, and two others carry their parents' house keys too. One participant carries loyalty cards shared with her husband.

A couple of participants volunteered that it's not just the hassle of carrying so much material that is the problem, but it's also their mental anxiety over wondering if they might have forgotten something. These people wanted someone or some tool to help them manage their keys and cards. *"From a technological point of view – [I want] someone [to] tell me your key is this place or your credential info is here...[It would help] best at home – [I] put my keys somewhere – depends on situation – baby crying, sofa, piano, and then I forget [where I put them]. But when I try to use car first have to find key or I can't use my car. So [if I could] have it be 'go to the car and someone gives me this key' that would be great!" (P13) "Did I forget something? Constant confusion if I forget something." (P8)*

Changes to routine also increase the chance that people won't have the authenticators they need with them. One participant mentioned *"Traveling has a problem with acquiring more keys and cards..." (P12)* Emergencies are a further problem: during a recent fire drill at a participant's company where emergency communications required particular tablets, *"The emergency crew didn't remember to bring the tablets with them when exiting the building, or they had them outside in their locked cars but didn't bring out the car keys." (P12)*

Finally, people carry scraps of paper in their wallets and bags with authentication material, sometimes obfuscated and sometimes not. For instance, one participant carries a paper with last year's gym locker combo on it *"'cause I was*

constantly forgetting it and asking the coach to open it for me." (P8) Another participant carries a paper in his wallet that is *"a letter of love to my wife – but it happens to be passwords encoded." (P11)*

6.2.2 Password management

Secrets constitute the largest portion of authenticators for our participants. They were used to unlock laptops, phones, lockers, bicycles, and even house doors. In our dataset, among all phone unlock events, 92% occurred with a PIN, 7.7% with a fingerprint reader, and 0.3% with a swipe gesture. For laptops, most participants used passwords, but in some instances (0.8%) the participant only had to click to login because the laptop was set to auto-login. For Website, which includes online services and access to software on phones or laptops, participants used passwords for 75.7% events and they used Mouse click (auto-filling the password with a password manager or cached passwords) for only 21.3% events. This surprised us, because we expected participants to use password managers or cache their passwords in the browser more often.

Many people feel that the rules around choosing and managing passwords have become onerous, especially in corporate environments. Across all our study participants, including those employed by our company, we found *no one* who followed all our company's workplace rules for passwords: change them frequently, don't reuse them, choose passwords of significant complexity, do not use the same password across multiple sites or accounts, do not write them down, do not cache them in browsers, and do not use a third-party cloud-based password manager to store them. *"It's awful. I'm dying...Everybody's got different rules and people are requiring I change them and then I can't remember them. Then life is hell...I use the same one [password] – I'm not a fool...All the tools to do my job are impossible to get to...This requirement that I change the password – They're causing us not to be able to remember the password, not to pick a good one, to use the same one and just change the postfix, or to write it down. They're forcing me into this corner – I don't know what to do. Maybe I'll write it on a sticky note and paste it on my computer." (P17)* Everyone "cheated" in at least some regard – and they were aware of it. Immediately after they told us about a bad practice, they confessed that it was bad or justified their action. *"About the management aspect – remembering a password – I reuse passwords is how I get around it, which is bad." (P2)* This may indicate that password management has become difficult enough that even otherwise conscientious tech-savvy employees are not willing to abide by the requirements.

To manage their many passwords, participants turned to a variety of tricks and tools: password-managers (n=9); password reuse (n=5); password reuse with permutations (n=8); passwords saved in an encrypted file (n=5); passwords saved in a plaintext file (n=3); passwords cached in browsers (n=5); passwords written on physical paper kept hidden (n=1); passwords kept in draft email (n=1); and passwords memorized (n=10). Several participants used more than one strategy. In a user study by Ion et al. 19% of non-expert users reused passwords, which matches our results [17]. We expected more participants in our study to use password managers, but only 34% of participants did, which is higher than the 24% of non-expert users but much lower than the 74% of expert users in Ion et al. [17] or the 81% of users in a study by Stobert et al. [30]. We suspect the low percentage of password manager use in our study is because many participants' organizations did not feel benign toward third-party cloud-based password managers. One participant mentioned that being able to share passwords was important for him, and that was one of the reasons he did not use password managers.

Table 4: Participant opinions regarding authentication and number of participants who gave a specific rating. N: normal ratings; N*: with volunteered ratings for when something goes wrong.

Opinion about Authentication	N	N*
(1) I don't even notice them.	1	1
(2) I notice them, but they rarely bug me.	9	6
(3) They bug me, but not too much.	10	8
(4) They bug me and I'd like to avoid them.	5	7
(5) They are extremely frustrating.	1	4

6.2.3 Opinions about authentication

Our participants' opinions on authentication vary widely. In the post-logging interview, we asked participants to rate their overall feelings about authentication on a scale from 1 to 5, with 5 being extremely frustrating. Table 4 shows that 16 participants found authentication at least somewhat burdensome. Seven participants, unprompted, gave two opinions when asked about how they feel about authentication: first for how they feel in the normal course of things (column N in the table), and second for how they feel when something goes wrong (column N* in the table), such as forgetting a password, losing a key, or having to change a password. Several participants gave fractional answers, which we rounded down. *"They bug me a little [rating 3] but they give me a sense of security. Shoots to a 5 when I have to set up an account or service or use the phone to enter 15 character password." (P20) "Most of the time it's just the cost of doing business [rating 2] – until it breaks. Then it's a 5 because it stops me doing what I need to do right now." (P12)*

We were interested in whether there was any correlation between participants' authentication opinions and the number of authentication events they performed or the failures they encountered. We expected participants who logged more events or encountered more failures would be more frustrated, but saw no correlation between number of authentications and opinion. This agrees with the NIST study findings [29]. Further, there is no strong correlation across number of failures and participant opinion. We also saw no significant difference between average opinion rating of female and male participants (2.9±1.0 vs. 2.8±0.9).

Best and worst authenticators. The kinds of authenticators participants most liked or disliked varied greatly, as seen in Table 5. Some participants' favorite authenticators were other participants' least favorite. Note that participants' answers were their own and not chosen from a predetermined list. (If they had been from a predetermined list, we might have seen more people choose authenticators such as "cached passwords" as most-liked.) While we supposed many people would complain about passwords, we were surprised by the number who disliked physical authenticators such as keys and badges. Participants also sometimes distinguished

Table 5: Authenticators participants most liked and disliked, and the number of participants (N) who did so. Flash-to-pass is one participant's authentication method that allows her to open her garage door by flashing her headlights, which then also unlocks her house from the garage entry.

Liked	N	Disliked	N
Fingerprints	7	Passwords	16
Badges and passes	6	Physical Keys	6
Pin codes	5	Pin codes	3
Key fobs	5	Badges and passes	2
Physical Keys	3	Fingerprints	1
Passwords	2	Credit cards	1
Cached passwords	2	Barcodes	1
Flash-to-pass	1	Key fob	1
		Everything	1

between the number of times they had to use a particular authenticator versus the amount of effort required each time.

Although the most disliked authenticator varied among participants, for most *having to remember* something (including carrying a physical token) was the explanation. *"I don't like my badge. I never remember to have it on me when I should." (P21) "Passwords, because I have to remember them." (P26)* Another reason to dislike an authenticator (especially keys) was the need to carry it: *"I don't like to carry around physical keys. It's just another thing to manage." (P7)*

Most participants liked an authenticator because it was either *automatic* (keyfobs or badges) or *quick* (4-digit PIN, fingerprint). Interestingly, participants who liked fingerprints and also used them during the study said they encounter failures with fingerprints often – indeed, we observed this in their logs – but they did not seem to mind the failures, because it was quick to try again. *"The fingerprint swipe for my phone [is my favorite]. It failed a lot but you don't have to do much." (P18)* Several participants who did not actually use a fingerprint reader during the study also said they like fingerprint authentication because of its speed and low effort. The need for quick, effortless authentication matches with the findings of De Luca et al. that participants did not favor Face Unlock because they found it slow [8]. Our results suggest that for the majority of participants, an authenticator being quick and effortless is more important than its being accurate in terms of false rejects. There were two exceptions, however. One participant whose wife has a fingerprint reader on her phone dislikes that mode of authentication due to its error rate. Another participant says *"I like the usability and quickness if I hold the phone correctly. But sometimes it really annoys me if there's water or something sticky – after washing my hands – it wouldn't work." (P20)* Perhaps we should require an authentication method to promote rather than punish good hygiene.

6.3 Authentication failures

Authentication failures add to users' frustrations. We observed a higher percentage of authentication failures than we expected. We compute failure rate for an authenticator as the ratio of failed attempts with that authenticator and total attempts with that authenticator. Failed attempts is the number of times a participant tries to authenticate to a resource and fails; for instance, if a participant had to try three times to unlock her phone, and succeeded in the third attempt, she had two failed attempts and a total of three authentication attempts. We did not see any significant difference in failure rates across gender or age. Note that these are all false reject failures, not false accept failures, as self-reporting of events is unlikely to tell us if any of our participants attempted to break into something they should not have.

The six participants who used a fingerprint reader logged a high failure rate of 25%, because one of the participants (P18) injured the finger he uses for fingerprint authentication and thus suffered many failures (44%). The participant reported that he could not authenticate via fingerprint with the injured finger; he would retry until his phone required him to type his password. Two other participants used fingerprint authentication less than five times with a 50% failure rate, so the average failure rate across participants is high. If we exclude the injured participant and two light users of fingerprint authentication, we see a fingerprint biometric failure rate of about 3.1% (± 3.0), which is still higher than we expected.

We saw a 5.6% (± 10.8) failure rate for Mouse clicks, which refer to an authentication event in which participants used a mouse click to authenticate (e.g., choose a certificate, auto-fill a password entry). The failures in mouse click authentication are instances when the participant accidentally chose the wrong certificate or accidentally auto-filled the wrong username and password (e.g., for websites where the participant has multiple accounts). The failure rate with physical keys was about 3.3% (± 7.9). Failures with physical keys were due to events such as a participant selecting the wrong key from her key bunch and trying it on the lock.

Password failure rates

Overall, we found a 5.1% (± 5.8) error rate for passwords among participants. If we look closer, the password failure rates differ based on the target (Websites, Laptops and desktops, Phones, Tablets, and Lockers/Combination locks). Websites have the highest failure rates (11.4% ± 16.8) even though website logins account for only 4.7% of authentication events. Laptop has the second highest failure rate, 7.9% ± 9.1, which surprised us, because laptop or desktop passwords are frequently used, often typed several times a day, so we supposed muscle memory would help reduce this error rate. We observed a 2.3% ± 3.5 failure rate for Phone passwords, a 0.4% ± 1.1 failure rate for Tablet passwords, and a 1.7% ± 2.4 failure rate for locker combination passwords.

In our post-logging interview we asked participants about their high failure rate with passwords. Several participants commented on making mistakes typing passwords (because of the length and/or complexity of the passwords) and forgetting a password, especially for websites that are not often used. This observation matches with past findings by Adams, Sasse, and Lunt that users have trouble recalling infrequently used passwords [1]. Several participants quoted "typing too fast" for getting passwords wrong, either out of habit or because they do not want anyone to see their password. *"I always type it super fast and get it wrong a couple of times." (P16) "It can be stolen easily, that's why I'm always in a hurry in typing a password – it's a mental thing, even if no one is around. It makes me type it quickly – it's instinct." (P14)*

Typing an old password (due to muscle memory) when the laptop password was recently changed or getting confused between devices they use, and typing the password of one device on other, are two more reasons why participants incorrectly entered passwords. *"It's muscle memory and I usually mess up when I update a password. I'll type old ones first and of course it fails." (P2) "I'm on autopilot typing my password, which is different on my PC versus my Mac, so I have an ordered search of passwords I go through until one works." (P1) "I get them [desktop and laptop] mixed up, and I type the wrong password – it's muscle memory." (P3)* One participant commented on not being able to see a password when it is being typed, especially for long passwords. Another participant (P1) complained about his phone being less responsive if he ran too many apps in the background, and "typing in the numbers [PIN] registered too slowly," so he had to retry.

Participants attributed Laptop authentication failures to incorrectly typed passwords, even though they knew the password. They further attributed mistyping passwords during Laptop authentication to their desire to login in quickly. Perhaps certain passwords are easier to type for a user than others of the same length. If the user frequently makes the same mistake when typing a password, perhaps the authentication target can suggest changing the password to the frequently mistyped password. On the other hand, this requires keeping more authentication material in potentially vulnerable places.

7. USE OF AUTHENTICATION LOGS

We engaged in this study to gather information in aid of projects involving authentication. As an example, we used our authentication event logs as workload "traces" for energy consumption simulations of an authenticator device called Mobius.

The Mobius Ring is a prototype of a "universal authenticator." The idea behind Mobius is that the ring will perform authentication tasks on behalf of the user, and will thus take the place of a user's many authenticators: passwords and other secrets, and physical tokens such as keys and badges. Ideally, if Mobius works well, a user would only need to remember one authentication secret (to activate the ring) and carry one authentication token (the ring itself). The ring must sense its presence on a user's finger when activated and deactivate itself when it senses its removal from a user's finger. Existing examples of universal authenticators, some with only a subset of these features, include Pico [28], the Nymi band [27], the NFC Ring [33], and the Java Ring [7].

There are many issues to explore for Mobius, including how to authenticate with the device and how vulnerable it is as a potential central point of failure. One usability concern we explored is whether users must remove the ring for recharging, or whether we can keep it perpetually charged via energy harvesting. If users remove the ring for recharging, they must reauthenticate with it when they put it back on, and they are without their authenticator while it recharges.

The ring prototype is a 3D printed enclosure and its required electronics, including both Bluetooth Low Energy (BLE) and Near Field Communication (NFC). The ring's components and functions consume energy except for the harvesting we can perform during authentication events that use NFC or while holding an NFC-equipped phone with the hand wearing the ring. Using experimentally determined measurements of component and functional energy consumption and harvesting, we use our logs of authentication events as workloads to estimate the energy neutrality of ring operation. We assume that interactions with mobile phones, transportation transponders, and card-based doorlocks use NFC, while other events use BLE. We treat phone unlocks as phone usage events (NFC harvesting opportunities) and vary the length of the associated time a user might hold the phone. We find that for the average session of 2 minutes across users as reported in the LiveLab traces from Rice University [26], it is feasible to keep Mobius perpetually powered using only harvested power given our observed workloads (see Appendix G).

8. SUMMARY

We present the design and implementation of a wearable digital diary study, our findings about participants' authentication habits and opinions, and an example use of our study's event logs as a workload for evaluating a potential authentication device. Overall we find that authentication is a noticeable annoyance in participants' lives, but they are creative in devising ways to cope with it. On average our participants performed 25% of their authentications with a physical token, and several participants expressed frustration over the authentication material they have to carry. Participants encountered authentication failure rates of about 3-5% during the study, with higher failure rates (7–12%) for PCs and websites. Participants' opinions about how burdensome authentication is to them vary greatly, as do their likes and dislikes about authenticators, with no one authentication method favored by everyone. In the study we used a smartwatch app with a novel slot-machine type interface for quick logging of events, and most of our participants favored the smartwatch over the smartphone for in-the-moment logging. We believe such wearable digital diary studies may be good platforms to conduct future studies that benefit from in-the-moment logging.

We are making our study data publicly available. Please contact the authors for more details.

9. ACKNOWLEDGMENTS

We thank the many people who greatly helped us with user study advice or feedback from pilot studies: Sunny Consolvo, April Mitchell, Iris Beneli, Alvin AuYoung, Ben Eric Andow, Animesh Srivastava, Kassem Fawaz, Jim Mann, Aarathi Prasad, and Denise Anthony. We also thank the SOUPS reviewers for their very helpful comments, and our paper shepherd, Cormac Herley, for his excellent help and encouragement. Any remaining errors, misapprehensions, or stupidities are entirely the fault of the authors.

10. REFERENCES

[1] A. Adams, M. A. Sasse, and P. Lunt. Making passwords secure and usable. In *People and Computers XII*, pages 1–19. Springer London, Jan. 1997. DOI 10.1007/978-1-4471-3601-9_1.

[2] AnalogDevices. ADXL362. Online at http://www.analog.com/en/products/mems/mems-accelerometers/adxl362.html. Last accessed June 2016.

[3] AustrianMicrosystems. AS3953. Online at http://ams.com/eng/Products/Wireless-Connectivity/Sensor-Tags-Interfaces/AS3953. Last accessed June 2016.

[4] L. Bauer, L. F. Cranor, M. K. Reiter, and K. Vaniea. Lessons learned from the deployment of a smartphone-based access-control system. In *Proceedings of the Symposium On Usable Privacy and Security (SOUPS)*, pages 64–75, 2007. DOI 10.1145/1280680.1280689.

[5] L. F. Cranor. What's wrong with your pa$$w0rd? TED, Mar. 2014. Online at http://www.ted.com/talks/lorrie_faith_cranor_what_s_wrong_with_your_pa_w0rd?language=en. Last accessed June 2016.

[6] L. F. Cranor and S. Garfinkel. *Security and usability: Designing secure systems that people can use*. O'Reilly Media, Inc., 2005.

[7] S. M. Curry. An introduction to the java ring. Java World, April 1998. Online at http://www.javaworld.com/article/2076641/learn-java/an-introduction-to-the-java-ring.html. Last accessed June 2016.

[8] A. De Luca, A. Hang, E. von Zezschwitz, and H. Hussmann. I feel like I'm taking selfies all day!: Towards understanding biometric authentication on smartphones. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 1411–1414, 2015. DOI 10.1145/2702123.2702141.

[9] S. Egelman, S. Jain, R. Portnoff, K. Liao, S. Consolvo, and D. Wagner. Are you ready to lock? In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, Nov. 2014. DOI 10.1145/2660267.2660273.

[10] D. Florêncio and C. Herley. A large-scale study of web password habits. In *Proceedings of the International World Wide Web Conference (WWW)*. ACM, 2007.

[11] A. Greenberg. The app I used to break into my neighbor's home. Wired, July 2014. Online at http://www.wired.com/2014/07/keyme-let-me-break-in/. Last accessed June 2016.

[12] J. Gummeson, B. Priyantha, and J. Liu. An energy harvesting wearable ring platform for gesture input on surfaces. In *Proceedings of the International Conference on Mobile Systems, Applications, and Services (MobiSys)*, pages 162–175. ACM, 2014. DOI 10.1145/2594368.2594389.

[13] M. Harbach, E. von Zezschwitz, A. Fichtner, A. D. Luca, and M. Smith. It's a hard lock life: A field study of smartphone (un)locking behavior and risk perception. In *Proceedings of the Symposium On Usable Privacy and Security (SOUPS)*, pages 213–230, July 2014. Online at https://www.usenix.org/system/files/conference/soups2014/soups14-paper-harbach.pdf. Last accessed June 2016.

[14] E. Hayashi and J. Hong. A diary study of password usage in daily life. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 2627–2630, 2011. DOI 10.1145/1978942.1979326.

[15] C. Herley. So long, and no thanks for the externalities: The rational rejection of security advice by users. In *Proceedings of the Workshop on New Security Paradigms Workshop (NSPW)*, pages 133–144, 2009. DOI 10.1145/1719030.1719050.

[16] D. Hintze, R. D. Findling, M. Muaaz, S. Scholz, and R. Mayrhofer. Diversity in locked and unlocked mobile device usage. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp Adjunct)*, pages 379–384, 2014. DOI 10.1145/2638728.2641697.

[17] I. Ion, R. Reeder, and S. Consolvo. "...no one can hack my mind": Comparing expert and non-expert security practices. In *Proceedings of the Symposium On Usable Privacy and Security (SOUPS)*, pages 323–346, July 2015. Online at https://www.usenix.org/system/files/conference/soups2015/soups15-paper-ion.pdf. Last accessed June 2016.

[18] I. Lillegaard, E. Løken, and L. Andersen. Relative validation of a pre-coded food diary among children, under-reporting varies with reporting day and time of the day. *European journal of clinical nutrition*, 61(1):61–68, 2007. DOI 10.1038/sj.ejcn.1602487.

[19] I. Lookout, Harris Interactive. Mobile mindset study, June 2012. Online at https://www.lookout.com/static/ee_images/lookout-mobile-mindset-2012.pdf. Last accessed June 2016.

[20] Lookout, Inc., Harris Interactive. Survey reveals consumers exhibit risky behaviors despite valuing their privacy on mobile devices, Oct. 2013. Online at https://www.lookout.com/news-mobile-security/sprint-lookout-mobile-privacy-survey. Last accessed June 2016.

[21] MaximSemiconductor. MAX17058. Online at http://datasheets.maximintegrated.com/en/ds/MAX17058-MAX17059.pdf. Last accessed June 2016.

[22] MaximSemiconductor. MAX17710. Online at http://datasheets.maximintegrated.com/en/ds/MAX17710.pdf. Last accessed June 2016.

[23] M. B. Miles and A. M. Huberman. *Qualitative data analysis: An expanded sourcebook*. Sage, second edition, 1994.

[24] Motorola. Motorola MOTOACTV. Online at https://motoactv.com/home/page/features.html. Last accessed June 2016.

[25] NordicSemiconductor. nRF51822. Online at https://www.nordicsemi.com/eng/Products/Bluetooth-Smart-Bluetooth-low-energy/nRF51822. Last accessed June 2016.

[26] C. Shepard, A. Rahmati, C. Tossell, L. Zhong, and P. Kortum. Livelab: measuring wireless networks and smartphone users in the field. *ACM SIGMETRICS Performance Evaluation Review*, 38(3):15–20, 2011. DOI 10.1145/1925019.1925023.

[27] H. Slade. Bionym inks $14m to get password-replacing wearable, Nymi out the door. Forbes, Sept. 2014.

[28] F. Stajano. Pico: No more passwords! In *Security Protocols XIX*, volume 7114 of *Lecture Notes in Computer Science*, pages 49–81. Springer-Verlag Berlin, Mar. 2011. DOI 10.1007/978-3-642-25867-1_6.

[29] M. Steves, D. Chisnell, A. Sasse, K. Krol, M. Theofanos, and H. Wald. Report: Authentication diary study. Technical Report NISTIR 7983, National Institute of Standards and Technology (NIST), 2014. DOI 10.6028/NIST.IR.7983.

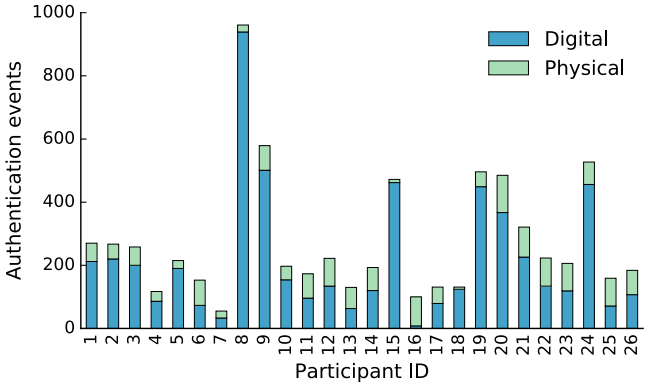


Figure 6: Authentication across participants, by category of target. (See Section 6.1.2 for target categorization.)

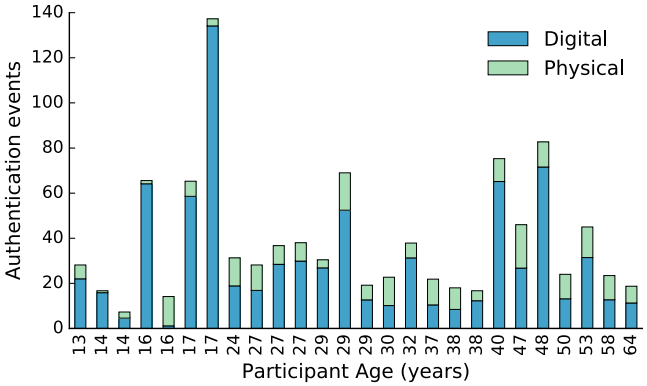


Figure 7: Average authentication events to digital and physical targets per day, by age of participants.

[30] E. Stobert and R. Biddle. The password life cycle: User behaviour in managing passwords. In *Proceedings of the Symposium On Usable Privacy and Security (SOUPS)*, pages 243–255, July 2014. Online at https://www.usenix.org/system/files/conference/soups2014/soups14-paper-stobert.pdf. Last accessed June 2016.

[31] I. Urbina. The secret life of passwords. New York Times, Nov. 2014. Online at http://www.nytimes.com/2014/11/19/magazine/the-secret-life-of-passwords.html. Last accessed June 2016.

[32] D. Van Bruggen, S. Liu, M. Kajzer, A. Striegel, C. R. Crowell, and J. D'Arcy. Modifying smartphone user locking behavior. In *Proceedings of the Symposium On Usable Privacy and Security (SOUPS)*, pages 10:1–10:14, July 2013. DOI 10.1145/2501604.2501614.

[33] R. Whitwam. NFC ring hands-on: Practice makes... pretty good. Android Police, Mar. 2014. Online at http://www.androidpolice.com/2014/03/09/nfc-ring-hands-on-practice-makes-pretty-good-video/. Last accessed June 2016.

[34] R. Witty and K. Brittain. Password Reset: Self-Service That You Will Love. Gartner Research, April 2002. Online at http://www.gartner.com/DisplayDocument?ref=g_search&id=354760. Last accessed June 2016.

APPENDIX

A. MORE AUTHENTICATION PATTERNS

Here we provide supplemental results for the authentication patterns covered in Section 6.1.

Distribution of events by target. Figure 6 shows the number of authentication events, for digital and physical targets, for each participant (see Section 6.1.2 for the categorization). All participants performed authentication with physical targets during the study, but there is high variance among them, with P8 logging only 2% of his authentications as physical and P16 logging 92% as physical. The average percentage of physical authentication events logged by each participant is 30.7% with a standard deviation of 20.2%.

Distribution of events by age. Figure 7 shows authentication events logged by participant age for digital and physical targets.

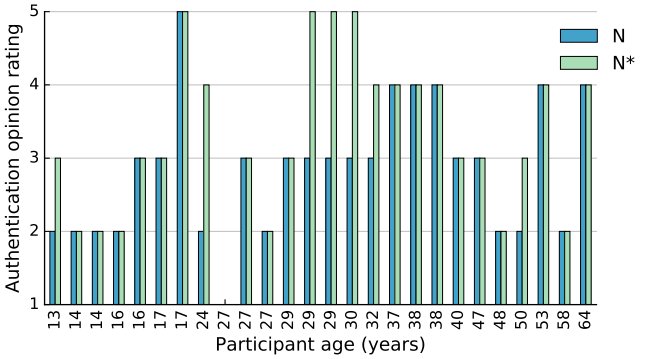


Figure 10: Participant opinions regarding authentication, by age of participants, both normally (N) and with ratings some participants volunteered for when something goes wrong (N*).

Weekday vs. Weekend. Figure 8 shows the average number of authentication events each participant logged during an average day Monday through Friday versus an average day on a weekend.

Distribution of events by hour of the day. Figure 9 shows the number of authentication events performed at different hours of a day, *averaged across both participants and days*. All our participants have day jobs or generally follow a day-oriented schedule, and this is evident from the figure, as there are few events after midnight. However, there is no hour where on average someone isn't authenticating with something.

Authentication opinion by age and logged events. In our post-logging interviews we asked participants to rate their authentication experience on a scale of 1 to 5, with 5 being extremely frustrating. Figure 10, Figure 11 and Figure 12 show the distribution of their opinion ratings by their age, the number of authentication events they performed, and the number of failed events they encountered in the study, both during the normal course of things (N) and when something goes wrong (N*). As we summarized in Section 6.2.3, we saw no correlation between participant opinions and their age, number of performed authentications, or number of encountered authentication failures.

B. PRE-LOGGING INTERVIEW

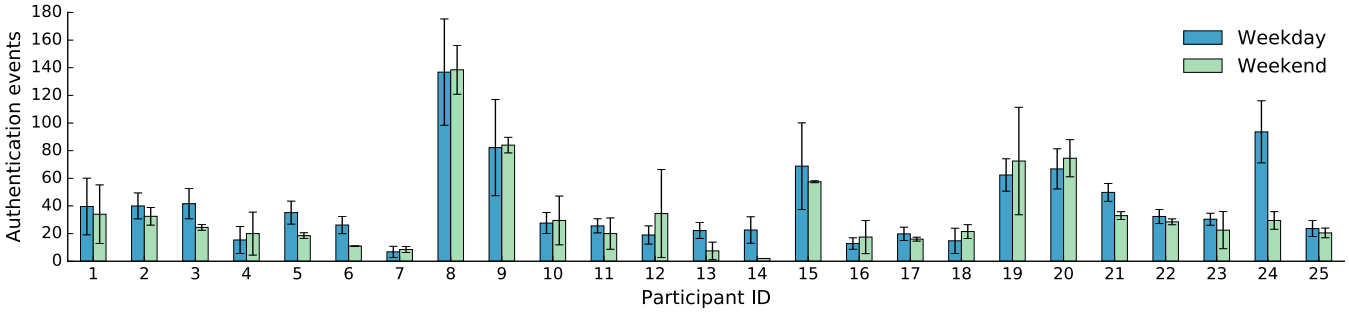


Figure 8: Number of authentication events participants performed on a weekday (averaged across Monday through Friday) versus a weekend day (averaged across Saturday and Sunday). Error bars show standard deviations.

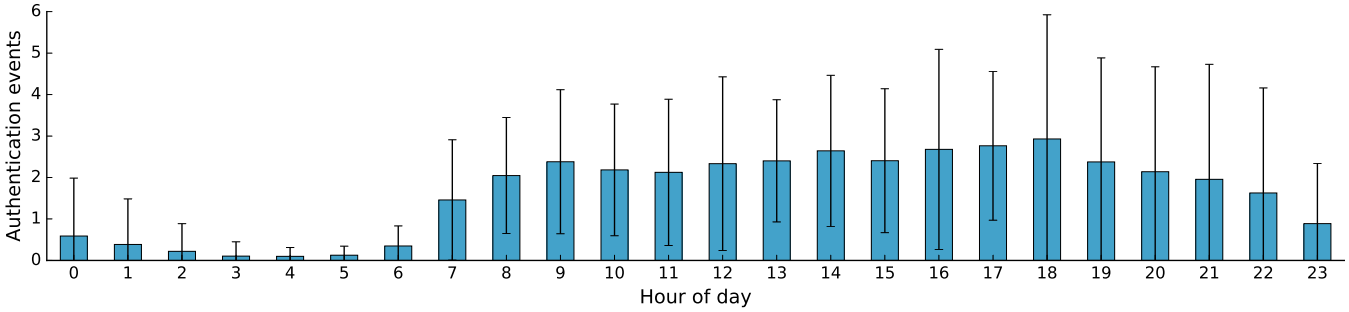


Figure 9: Number of authentication events performed at different hours of the day, averaged (± standard deviation) across participants and days.

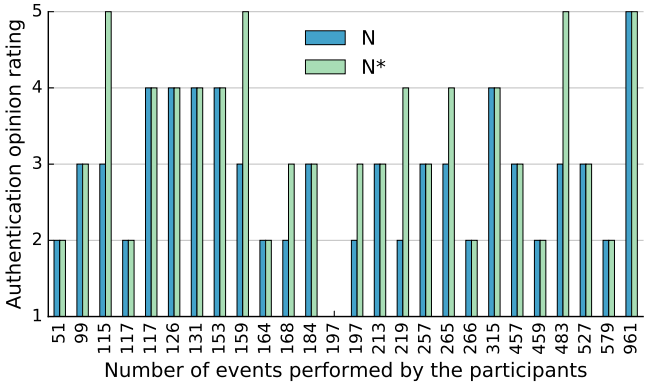


Figure 11: Participant opinions regarding authentication, by the number of authentication events performed by participants, both normally (N) and with ratings some participants volunteered for when something goes wrong (N*).

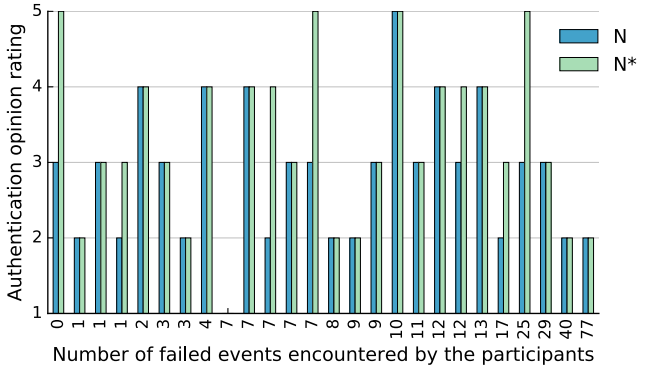


Figure 12: Participant opinions regarding authentication, by the number of failed events encountered by participants, both normally (N) and with ratings some participants volunteered for when something goes wrong (N*).

We used the following questions to guide our semi-structured interviews, before the participants began self-logging their authentication events. In addition to these questions, we welcomed topics and discussions about authentication initiated by participants.

- What is your typical day, in terms of authentication events?
- What targets and authenticators do you use? [We explained the meaning of targets and authenticators.]
- What do you carry with you? [We guided them to look in their bags, wallets, pockets, and purses.])
- How many times a day do you think you authenticate yourself with something?
- How do you manage your passwords?
- How do you choose/create passwords?

C. POST-LOGGING INTERVIEW

We used the following questions to guide our semi-structured interviews after participants logged their authentication events for one week. In addition to these questions, we probed participants about their authentication behavior, based on the logged data.

- What are your favorite authenticators?
- What are your least favorite authenticators?

- Did you log all events?
- Did your participation in the study lead you to be more aware of authentication events?
- How did your participation in the study change your authentication behavior?
- Did you notice any patterns in your authentication behavior?
- How do you feel about authentication events? (Multiple choice question)
  1. I don't even notice them.
  2. I notice them, but they rarely bug me.
  3. They bug me, but not too much.
  4. They bug me and I'd like to avoid them.
  5. They are extremely frustrating.
- How do you feel about passwords?
- In the future, what kinds of changes or inventions would you like to see related to authentication?
- Do you have any comments/suggestions/concerns about the study?

D. GUESSING ABOUT AUTHENTICATION

How accurate are people's feelings about how much authentication they perform in a day? In the pre-logging interview we asked participants how many times they believe they authenticate themselves every day on average. Fifteen participants overestimated their daily authentications and all but one of them did so significantly (by more than 25%). Eleven underestimated. Combined, the average of guessed daily authentication events was 47 ($\pm$ 31) vs. 39 ($\pm$ 29) logged authentications. Most of our participants overestimated, and only two participants came within 10% of their self-reported numbers.

E. PRIVACY ATTITUDES

One question many related studies consider is how much people care about privacy. We observed a higher level of care than we expected from our participants, with only two of the seven teenagers leaving their phones unlocked and two of the adults doing so. While our study does not include enough participants to make broad generalizations, we see evidence that teenagers and not just adults are interested in privacy and security, although teens may have less useful understandings of how to achieve it. We asked all participants why they chose to lock or leave unlocked their personal devices and resources. Both of the teenagers who did not lock their phones said it was because their phones always remained under their physical control, or in a safe environment (a desk at home). One of them also said he was careful not to keep anything private on his phone, and that he backed it up so nothing would be lost if his phone were lost. The other five teenagers all locked their personal devices with the intent of keeping them safe from the prying eyes of friends and sometimes parents and siblings. *"[I lock my phone] so people don't just go inside my phone – it's not pleasant for anyone this kind of snooping." (P8)* Three of the teenagers and seven of the adults also mentioned that besides having activity timeouts on their personal devices that automatically lock them, they deliberately lock their devices whenever they put them down or walk away from them, regardless of timeouts. On the other hand, both teenage girls (but none of the teenage boys) mentioned that they share their phone passwords with selected friends. This sharing seems to have social significance, and one of the teenagers suggested at the end of her post-logging interview that any kind of new authentication technology needs to support sharing of access. *"I want to use thumbprints on everything but I can't pass thumbprints to others – some friends can have access to my phone but not everyone." (P19)*

F. QUOTES FROM PARTICIPANTS

Here we include a few more participant comments, in addition to those already in the paper, because we found them especially interesting, representative of a particular point, or entertaining.

F.1 Feelings about authentication

"It's important – necessary, so you just do it." (P3)

"Most of the time it's just the cost of doing business – until it breaks. Then it's a 5 because it stops me doing what I need to do right now." (P12)

"It's kind of evil. It's a constant reminder that there are bad people. It makes me feel kind of bad, kind of angry." (P15)

"Sometimes it's annoying, but not all the time. I'm also very thankful for it." (P19)

"They bug me a little but they give me a sense of security. Shoots to a 5 when I have to set up an account or service or use the phone to enter 15 character password..." (P20)

"But when things go wrong, that's the worst. My worst was that I locked my keys in the car as I was getting out of it with two cats in two carriers to take them to their vet appointment. I also had my infant son with me in his car seat and I put down the carriers to go around to the other side of the car and get my son out, but I'd somehow locked the door when I closed it and my keys were inside the door and so my son was locked in the car. I couldn't leave him there and I couldn't leave the cats and it was horrible. But a guy in the parking lot was able to break into my car for me. I was never happier in my life to meet a competent criminal." (P21)

F.2 Likes and dislikes for authenticators

Likes:

"[phone PIN] my fingers know where to go on the keypad." (P6)

"Fingerprint, cause it's very quick. The rest all take significantly more time. Even for a key fob – you have to take something out – it would be great if I could use a fingerprint at the [company] entrance." (P14)

Dislikes:

"[Most effort are physical keys] first you have to find it in your purse, then pick out the right key from the ring, then get it in your hand correctly to unlock the door. There's a difference between fast but many times and lots of effort but only a few times. So keys were a lot of effort, and the phone unlock wasn't, but I had to do it most often so it adds up." (P3)

"Typing passwords on the phone and laptop took the most effort for frequency and chance for failure. I hate passwords!

We cannot do patterns or face recognition to get our work email...They have improved the initial pin interface on the Galaxy but still there's a greater than 20% or 25% failure. Initially the keyboard was large and mis-hitting was high. But still there is a problem when I am sleepy or in the dark or when I don't wear my glasses...I'm also not happy with door key unlocking. The door key at my home takes a lot of pressure and when carrying your son's books and toys and your bag on the other arm or carrying your son on one arm sleeping, it is really hard...Also my car if I carry the [fob] is supposed to unlock from a button on the handle. I don't have to press the fob. But to unlock it for everyone you have to press it twice from the driver side and only once from the passenger side. This is confusing and I lock it sometimes instead. So sometimes I bring out the fob and deliberately use it to unlock even though I'm not supposed to have to do that." (P4)

"Credit cards because you have to pull them out of your wallet." (P6)

"Passwords – they are complicated and annoying." (P8)

"You're booting some device and have to type in a long password and you can't be sure you got it right 'cause you can't see it. Is it typed wrong or is the keyboard in the wrong mode? You gotta preserve this password over all other values to keep the devices running." (P11)

"I like [the] key fob as opposed to physical stuff. Remote authentication without physical contact is a much better experience than physical contact or swiping. But the fob is too big – it's difficult to carry." (P14)

"Long-assed passwords for sites I rarely go to are obnoxious. But keys could also be bad...Which one is which and they all get tangled up and you have to find it and if it were my phone I could just do it and then I realize I'm just a bratty girl from Silicon Valley and I should be okay with taking the 15 seconds to do it." (P15)

"I don't like my badge. I never remember to have it on me when I should...Also, I feel embarrassed wearing it – kind of like I'm a kid in kindergarten with a name tag. And I hate my photo that's on it. If you forget it then you're kind of humiliated at the front desk in the lobby. It doesn't fit on my keychain, so where else should I put it? In my purse – 'cause I always bring my purse to work. But I have to put it in a special pocket or I can't find it in my purse and think I've left it somewhere even if I haven't." (P21)

G. MOBIUS RING ENERGY SIMULATIONS

The Mobius ring, depicted in Figure 13, includes the following components:

- 3D printed enclosure.
- Near Field Communications (NFC) using the AS3953 [3] NFC interface chip.
- Bluetooth Low Energy System on Chip (SoC), the Nordic Semiconductor nRF51822 [25]. We intend to use the flash memory of this SoC to store encrypted passwords and pins in our prototype.
- A low power 3-axis accelerometer, the ADXL362 [2], for tap detection for entering the activation pin for the ring (the one secret the user must remember).

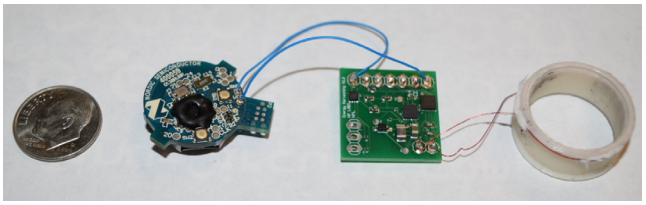


Figure 13: The components used in our current Mobius prototype are no larger than a typical Signet ring. A 10 mAh battery is behind the harvesting board (green).

- Pressure sensor (not yet implemented) mounted on the inside periphery of the ring to sense whether the ring is on the user's finger.
- The NFC interface stores the excess energy beyond what is required for authentication purposes in a small 10 mAh battery.
- We embed the NFC tag coil by winding a few turns of magnet wire around the circumference of the ring, similar to the approach used by Gummeson et al. [12].
- Prior to storage, the energy is conditioned by a MAX17710 energy harvesting chip [22], with charge state monitored using a MAX17058 fuel gauge IC [21].

Our first measurement result looks at how much energy we can harvest from NFC sources and effectively store in the ring's battery. To understand the end-to-end efficiency of energy storage, we monitor battery state using the onboard fuel gauge IC. Placing the ring within 5 mm above the NFC antenna embedded inside a Motorola Moto X, we observe an average harvesting rate of 1.67 mW.

Next, we look at the power consumption of different ring components to help understand the ring's steady state energy balance. The CPU portion of the BLE SoC consumes 1.08 μW of power in sleep state, and 4.32 mW while active. The BLE radio consumes 12.6 mW of power while transmitting at a power of -8 dBm and 23.4 mW of power while in receive mode. The accelerometer consumes 5 μW of power while actively detecting PIN entries, and consumes 270 nW while in a low power wakeup mode that is used to initiate authentication with a remote target. Using the 133.2 Joule buffer in Mobius, the ring can sustain itself in a low power wakeup mode for 132.6 days without any charging while polling its removal sensor at a rate of one hertz.

We model the energy consumed by a BLE authentication event by considering several steps of operation: 1) after a user taps the ring to wake it up, Mobius sends advertisement beacons to make authentication targets aware of its presence, 2) the authentication target initiates an unencrypted connection with Mobius, 3) Mobius and the authentication target encrypt the connection using a shared long term key that was previously established during bonding, 4) the ring sends an encrypted "unlock" command to the authentication target, and 5) the connection terminates.

When considering the power costs of processing and communication, an advertising interval of one second, a latency of four seconds to establish a connection, a BLE connection interval of 10 ms, and a BLE connection length of one second, Mobius consumes 419 μJ of energy per BLE authentication event. With our current energy buffer, we can handle

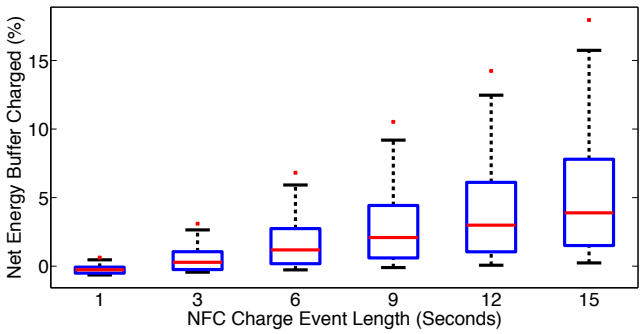


Figure 14: We model NFC charging events as interactions with mobile phones, transportation transponders, and card-based door-locks. We only need phone interactions to be an average of 15 seconds in length to keep the ring's buffer energy neutral across an entire week of authentication.

286,109 BLE authentication events – this assumes the last 10% of the battery is unusable due to low voltage.

Equipped with information about various hardware costs and the results from our user study, we seek to understand the feasibility of using Mobius as a perpetually powered universal authenticator. Since our hardware design is preliminary, our evaluation criterion is the overall energy neutrality of operation during the week we conducted the user study. During each simulation, Mobius' energy buffer is initialized to be at 50% capacity to avoid any coldstart effects.

The user study event logs allow us to estimate the impact a hypothetical Mobius workload has on the energy neutrality of operation. For our power simulations, we exclude data from participants for whom we have no automatically logged phone unlocks.

Our first results look at how changes in the length of mobile phone usage impact the energy neutrality of Mobius. In this experiment, we assume that doors unlocked with a card and transportation authentication targets each provide Mobius two seconds of charge time, but we vary the charge time provided through use of mobile phones. We assume that all other authentication targets use BLE for authentication and that when not authenticating, Mobius is in its low power mode where it seeks to detect removal events. We show the results of this study in Figure 14. When considering a very limited charge opportunity of one second during mobile phone use, no user experienced more than a ∼0.7% decrease in buffered energy, meaning that Mobius could run for more than 100 weeks before depleting its battery. After increasing the phone use length to 15 seconds, all but one user sees an overall increase in buffered energy after a week of operation; this user experiences a decrease of 0.01%. When we consider more realistic measures of the length of mobile phone use, such as an average of two minutes across users as reported in the LiveLab traces from Rice University [26], it seems feasible to keep Mobius perpetually powered using only harvested power.

Our final evaluation considers how decreasing the size of the energy buffer impacts the availability of Mobius for authentication. Our current design does not use the current battery for any fundamental reason; it was available off the shelf and amenable to the ring form factor. Since the battery used in our implementation is bigger than it needs to be, we currently do not experience any failures in authentication due to energy starvation. If we scale the energy buffer size down, we start to see failures in BLE-enabled authentication events based on their temporal distribution among charging opportunities and energy lost to sleep. For example, a significant amount of energy will be lost at night when users are sleeping rather than accessing their mobile phones. Figure 15 shows the number of failures across all users for five orders of magnitude of energy buffer size; we find that there are no authentication failures as a result of energy starvation for an energy buffer greater than or equal to one Joule in energy capacity. This result shows that the battery we are currently using is more than two orders of magnitude larger than it needs to be, indicating that there are opportunities for further platform miniaturization.

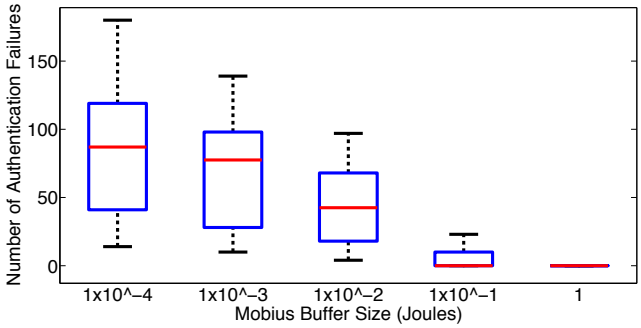


Figure 15: The battery currently used in our Mobius prototype is more than two orders of magnitude larger than it needs to be. A one Joule buffer has sufficient energy capacity to completely avoid failures due to energy starvation.

Our simulation study has several possible sources of inaccuracy that affect our ability to calculate how well charged we are able to keep the ring. First, the number of phone unlock events does not tell us how long the user keeps his phone in his hand after unlocking it. This means we do not know the length of time the ring can recharge due to its proximity to the NFC reader in the phone. However, we make a conservative assumption that is smaller than the unlock durations observed in the LiveLab traces. Second, the user does not necessarily hold his phone in the hand wearing the ring and the specific hand placement will result in variation of harvesting power – we leave a more detailed harvesting study to future work.

Use the Force: Evaluating Force-Sensitive Authentication for Mobile Devices

Katharina Krombholz
Ruhr-University Bochum, Germany and
SBA Research, Austria
kkrombholz@sba-research.org

Thomas Hupperich
Ruhr-University Bochum, Germany
thomas.hupperich@ruhr-uni-bochum.de

Thorsten Holz
Ruhr-University Bochum, Germany
thorsten.holz@rub.de

ABSTRACT

Modern, off-the-shelf smartphones provide a rich set of possible touchscreen interactions, but knowledge-based authentication schemes still rely on simple digit or character input. Previous studies examined the shortcomings of such schemes based on unlock patterns, PINs, and passcodes.

In this paper, we propose to integrate pressure-sensitive touchscreen interactions into knowledge-based authentication schemes. By adding a (practically) invisible, pressure-sensitive component, users can select stronger PINs that are harder to observe for a shoulder surfer. We conducted a within-subjects design lab study (n = 50) to compare our approach termed *force-PINs* with standard four-digit and six-digit PINs regarding their usability performance and a comprehensive security evaluation. In addition, we conducted a field study that demonstrated lower authentication overhead. Finally, we found that force-PINs let users select higher entropy PINs that are more resilient to shoulder surfing attacks with minimal impact on the usability performance.

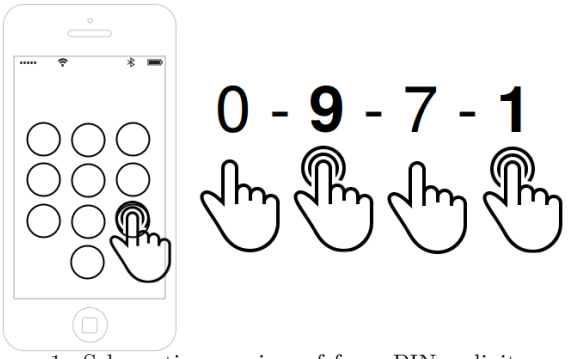


Figure 1: Schematic overview of force-PINs: digits can either be entered with shallow or deep pressure on a pressure-sensitive touchscreen, enhancing the space of four-digit PINs to $20^4 = 160,000$ by an invisible component. The user receives vibration feedback as soon as deep pressure is recognized.

1. INTRODUCTION

With the introduction of pressure-sensitive touchscreens (e. g., Apple recently introduced *3D Touch*[1]), many new kinds of user interaction for smartphones become possible that could also be used to enhance existing authentication schemes. The scientific community has already examined the shortcomings of unlock patterns, PINs and passcodes [2, 16, 19, 25] and presented alternative authentication schemes.

However, none of the proposed systems has shown to be capable of replacing passcodes and unlock patterns as means of authentication. On the one hand, many approaches, e.g., [15, 17] rely on customized hardware that is not available off the shelf and thus makes large-scale deployment infeasible. On the other hand, many alternative approaches, e.g., [13, 23] are time-consuming and therefore increase the authentication overhead. As shown by Harbach et al. [19] in a field study on smartphone unlocking behavior, (un-)locking smartphones produces a significant task overhead. This highlights the need for novel authentication methods that perform equally fast as or even faster than currently deployed systems in terms of authentication speed.

Recently, biometric approaches such as fingerprint sensors and face recognition have found their way into the mobile ecosystem. As with previous authentication methods, however, they have shown to be easy to break by attackers and difficult to use for certain groups of users. For example, Apple's fingerprint sensor as found in some recent iPhone models was soon hacked after being introduced [11] and excludes users with weak fingerprints (e. g., due to manual labor). Furthermore, classic biometric methods and implicit authentication based on user behavior still require users to use a PIN for fallback authentication in case the primary authentication methods fail. Bonneau et al. [8] presented a benchmark to evaluate authentication schemes. Their evaluation shows that many schemes only offer minor improvements over passwords (if any) and that many systems offer a number of benefits in theory but show severe limitations in practice. These observations highlight that it is still worth focusing on improving knowledge-based authentication on smartphones as no other authentication method has proven to be as secure and usable as passwords.

[1] `https://developer.apple.com/ios/3d-touch/`

Symposium on Usable Privacy and Security (SOUPS) 2016, July 22–24, 2015, Denver, Colorado, USA.

In this paper, we propose that device manufacturers integrate pressure-sensitive touchscreen interactions available on mobile and wearable devices into knowledge-based authentication schemes. Our goal is to improve PIN security by enhancing the password space without compromising usability factors such as authentication time, error rate and memorability. This approach enhances traditional four-digit or six-digit PINs with tactile features using pressure-sensitive touchscreens as found in modern consumer hardware. We refer to these enhanced PINs as *force-PINs* and Figure 1 provides an overview of the proposed scheme.

In theory, force-PINs offer the benefit of a larger PIN space by design. Hence they are more difficult for an attacker to guess and are more resilient to shoulder-surfing attacks due to the invisible pressure component. To estimate the task overhead introduced by this security feature, we present a comparative evaluation of force-PINs and standard four-digit and six-digit PINs as currently deployed in modern smartphones. We conducted a lab study with $n = 50$ participants to compare four-digit force-PINs against four- and six-digit standard PINs and performed a small shoulder-surfing experiment.

We found that entering force-PINs is more time-consuming than entering digit-only PINs. However, we also found that the difference in authentication time between six-digit and force-PINs was not statistically significant. The number of both critical and standard errors were rather low for force-PINs even though the participants from our lab study were using force-PINs for the first time. According to our survey results, the participants liked the invisible pressure component as an additional security feature.

In a small shoulder-surfing experiment, we found that the force component is more difficult for an attacker to observe: none of the force-PINs entered while being observed by an attacker was guessed correctly. However, the attackers were able to guess some of the digit sequences correctly. We also analyzed the user-chosen force patterns alongside with the entered digits and found that users create higher entropy PINs. In an additional field study, we collected evidence on learning effects and showed that authentication time decreases with training.

In summary, our contributions in this paper are:

- We propose an enhancement to digit-only PINs with an invisible force component via pressure-sensitive touchscreens.
- We implemented a prototype of the proposed scheme called *force-PINs*.
- We performed an evaluation of force-PINs, including a lab study with 50 participants, a security evaluation, and a field study with 10 participants.

The remainder of this paper is structured as follows: In Section 2, we discuss related work and in Section 3, we introduce the attacker model, the concept of force-PINs, and describe the objectives of this work. Section 4 presents the design and results of our lab study. In Section 5, we provide a security evalution and in Section 6.3, we present the results of a field study to show learning effects of force-PINs deployed in a real-world environment. Sections 7 and 9 discuss our work and its limitations and we conclude this paper in Section 10.

2. RELATED WORK

Given the importance and the practical impact, it is not surprising that there has been a significant amount of work on authentication schemes. In the following, we briefly review work closely related to our approach. We also refer to the work by Bonneau et al. [8], who presented a benchmark for evaluating authentication schemes.

Malek et al. [24] proposed a haptic-based graphical password scheme. They complement graphical passwords with personal entropies based on pressure and argue that the password space is increased. However, they did not conduct a user study to evaluate usability factors and do not provide empirical evidence that supports the theoretical calculations of a larger password space. Furthermore, they did not evaluate their approach against a shoulder-surfing threat model.

Bianchi et al. [3–6] proposed several authentication approaches based on tactile feedback with an emphasis on accessibility and multi-modal feedback. In comparison to our approach, they rely on a tactile wheel to interact with the system, a component which is not available in off-the-shelf devices.

To make smartphone authentication resilient to shoulder surfers, De Luca et al. [15, 17] presented an authentication mechanism that allows users to enter passwords at the front and the back of their device. While their approach offers benefits with respect to shoulder-surfing resilience, a major limitation of this approach is that there is no such device available at this time that provides users a touch-sensitive back.

Harbach et al. [19] performed a real-world study on smartphone unlocking and found that users spend a significant amount of phone usage time on unlocking their device with PINs and unlock patterns. On average, their study participants unlocked their phones about 47 times throughout the day. This finding shows that mobile device unlocking introduces a severe task overhead and highlights that authentication time is an important factor regarding the usability of the method. It also implies that any time-consuming method is potentially disadvantageous for usability and will therefore have difficulties in getting accepted by users. De Luca et al. [14] found that increased authentication time was a reason for Android users to stop using *Face Unlock* (called *Trusted Face* in later Android versions). Their study also revealed that usability factors are the primary reason keeping users from adopting biometric authentication on mobile devices and that privacy and trust issues only play a secondary role.

A new trending topic in authentication research is implicit authentication. E.g., Buschek et al. [10] studied the feasibility of mobile keystroke biometrics and found that they can be used for user authentication with relatively low error rates. As shown by Khan et al. [22], current methods for implicit authentication are not capable of replacing knowledge-based authentication because their real-world accuracy is significantly lower than in lab settings. Furthermore, they require a certain number of interactions to classify a user correctly. Therefore, these systems are often perceived as disruptive in cases where authentication fails and fallback authentication methods come into play.

3. CONCEPT AND OBJECTIVES

Our approach is based on PIN-based authentication and pressure-sensitive touchscreens as found in modern smartphones (e. g., *3D Touch* available in the iPhone 6s). In the following, we first describe the attacker model and then discuss the design and implementation of *force-PIN*.

3.1 Attacker Model

Throughout the rest of this paper, we assume that the attacker is able to perform a shoulder-surfing attack: she is in close vicinity to the user while authentication takes place and can observe the typing behavior (e.g., in a crowded public or semi-public environment). The key element of a successful shoulder-surfing attack is the ability to clearly observe all sensitive information being entered on the touchscreen.

We also assume that an attacker can gain possession of the user's device. In case the device gets lost or stolen, the design of force-PIN makes a PIN harder to guess due to the theoretically larger PIN space and the pressure component.

3.2 Force-PIN Design

Force-PINs are designed to be more resistant to observation due to the unobtrusive pressure component that helps to obfuscate PIN components and thereby complements regular PIN entry: a user enters a digit either via a shallow or deep pressure on a pressure-sensitive touchscreen. The user receives tactile feedback when entering a digit with deep force. The tactile component and vibration feedback may implicitly help users to memorize force-PINs [9].

An example force-PIN could be 0-**9**-7-**1** where bold and underlined numbers should be pressed more deeply than others on a pressure-sensitive touchscreen (see also Figure 1). The design is not only simple, it is also cheap and easy to deploy as it relies on off-the-shelf hardware. We expect that users who are already using pressure-sensitive touchscreens will find force-PINs as easy to learn as digit-only PINs as they are based on interactions they are already familiar with.

3.3 Implementation

For our study, we implemented a prototype app for iPhones with touch-sensitive screens. The app lets users set a force-PIN and presents a lock screen that looks just like a common lock screen from off-the-shelf iPhones. A force-PIN consists of four digits and a force pattern with two different pressure levels, namely shallow and deep press.

The design decision was based on a small pre-study with 9 participants where we evaluated subjective perceptions on different types of pressure encodings. We evaluated both relative and absolute differences in pressure with different thresholds, respectively. As two-stage pressure with a constant threshold for shallow and deep press performed best; we implemented the prototype app accordingly. We also tested different thresholds and to our surprise it was often not easy to distinguish which threshold was higher and which one was lower. Therefore, we then set the threshold for deep pressure to 50% or more of the maximum possible pressure supported by the hardware.

For our user study, we also implemented apps for four-digit-only and six-digit-only PINs for a comparative lab study and a slightly modified force-PIN app for our field study. The app for the field study had a different main screen and allowed users to submit additional comments to gather in-situ data. Furthermore, the app issued a daily notification to remind the participants of the study task. Each app stored the entered PINs and measured authentication time and failed attempts. The apps with force-PINs also stored the selected four-digit force pattern and arrays of force gradients that were measured for every touch interaction with a pressure-sensitive digit button.

4. LAB STUDY

In the course of a usability lab study, we evaluated force-PINs against digit-only four-digit and six-digit PINs. We chose to evaluate four-digit standard and force-PINs against six-digit standard PINs as they were introduced as the new default in iOS 9. We did not evaluate six-digit force-PINs as we wanted to minimize the additional task overhead. In this section, we describe the methodology and results of this lab study.

4.1 Design and Procedure

Our study is based on a within-subjects design, i.e., every participant is exposed to all conditions. This allows us to perform a comparative evaluation of all subjects exposed to our conditions. We assigned every participant a unique ID and a random order of conditions to reduce learning effects. The three conditions were as follows:

- (C1) four-digit PINs
- (C2) six-digit PINs
- (C3) four-digit force-PINs with shallow and deep pressure

We recruited participants around the university campus over bulletin boards and personal communication mentioning that the study was about their preference of different types of PINs. All of our participants were either employed or currently enrolled as students at the university. We recruited 50 participants for our lab study. They were compensated with a voucher for the university's cafeteria. Table 1 shows the demographics of our participants. All participants were frequent smartphone users and had used digit-only PINs before. To reduce the risk of biased interpretation, we presented the three PIN entry methods equally and did not provide any hints on which method was potentially more secure or not. The participants were not told that the study placed an emphasis on evaluating force-PINs.

The lab sessions proceeded as follows: First, the participants were briefed about the purpose of the study. A subsequent training session allowed them to get familiar with the different types of PINs. This was necessary to minimize the bias introduced by the comparison between a well-known and well-trained authentication method and a newly introduced scheme that users have not yet been exposed to.

Then the participants chose a PIN of the first assigned PIN type and afterwards authenticated with the respective PIN until they had completed three successful authentication sessions. After completing this task, the participant proceeded to the next condition, selected a new PIN and authenticated three times. We instructed the participants to select PINs that they thought were as secure as possible and asked them to remember the PINs just like their own ones in real life. We refrained from assigning PINs as it is a common scenario in the smartphone ecosystem that users

can choose their own PINs. For the same reason, we did not explicitly disallow PIN-reuse.

The metrics we used for our usability evaluation were authentication speed and error rate as defined by De Luca et al. [15]. They defined *authentication speed* as the time between the first touch and the last touch of the authentication session and only counted successful authentication attempts. Regarding the *error rate*, we differentiate between basic and critical errors (as also proposed by De Luca et al. [15]) where basic errors refer to errors within an overall successful authentication session (failed attempts) and critical errors refer to completely failed authentication sessions. Hence, successful authentication sessions may contain failed attempts that influence authentication speed.

In addition to the data collected through our smartphone apps, we gathered quantitative and qualitative data via a questionnaire consisting of 15 closed and open-ended questions to study the perceived security and usability of the three different types of passcodes. The reason why we chose to use open-ended questions was that we wanted to collect meaningful participant statements using their own knowledge, perceptions and interpretations. The questions can be found in Appendix A. After completing the experiments, all participants filled out the questionnaire on a laptop provided by the experimenters.

The participants had to provide their previously assigned experiment ID on the first page of the questionnaire to link the data sets. Except for age, gender and whether the participant had an IT background, no personal data was collected in order to preserve the participants' anonymity. We also collected data on smartphone usage and asked the participants which authentication method they were using at that time on their own smartphones.

The qualitative responses were coded using an iterative coding approach. Two researchers independently went trough the participant responses and produced an initial set of codes. Then, the researchers discussed reoccurring codes, topics and themes, and agreed on a final set of codes. Based on this set, one researcher coded the answer segments for further analysis. As most answers where short and to the point, we did not perform a reliability test of the final coding.

4.2 Results

Given our sample consisting of 50 participants, the quantitative results of our study are based on $3 * 3 * 50 = 450$ authentication sessions (three conditions, every pin type was entered three times by 50 participants). Our study has a repeated-measures design, i.e., every participant was exposed to every condition. Therefore, we analyzed our data with repeated measures ANOVAs. We removed 2 authentication sessions that lasted longer than 30 seconds from the dataset as those occurred when participants where distracted from the study task.

4.2.1 Authentication Overhead

Authentication Speed.

As proposed by De Luca et al. [15], we measured authentication speed from the first to the last touch of a successful authentication session. Hence, an authentication session can also contain a maximum of two failed attempts. After the third failed attempt, the user was locked out of the app.

Table 1: Participant characteristics from the lab study. n=50

Demographic	Number	Percent
Gender		
Male	31	62%
Female	19	38%
Decline to answer	0	0%
Age		
Min.	19	
Max.	56	
Median	25	
IT Background		
Yes	4	8%
No	46	92%
Smartphone		
Android	32	64%
iPhone	14	28%
Windows Phone	2	4%
Other	2	4%
Used Authentication Method		
4-digit PIN	26	52%
6-digit PIN	2	4%
Password (digits/characters)	3	6%
Unlock Pattern	14	28%
Fingerprint Sensor	7	14%
Face Recognition	0	0%
Android Smartlock	1	2%
None	5	10%

The participants had to start the sessions by clicking on a button.

We only considered successful authentication sessions to measure authentication speed. As every user entered every PIN type three times, we calculated the average authentication speed for every user and every authentication method and used this value for further analysis. Overall, 56 force-PINs were selected by our participants. Five of them decided to change their PIN during the experiments, one participant renewed the PIN twice. The participants did not mention any reasons for these decisions. The authentication time was measured based on the most recently selected PIN. Table 2 shows the mean authentication time in seconds and error rate. Figure 2 shows the collected authentication speed measures for all participants and PIN types.

To reveal significant effects regarding authentication speed, we performed a one-way repeated-measures omnibus ANOVA across the 3 PIN types. The results show significant differences in authentication time ($F_{2,147} = 10.19, p < 0.001$). A pairwise t-test with with $t_{0.95,98} = 1.9845$ revealed significant main effects comparing the authentication speed of four-digit with six-digit PINs ($p < 0042$). In addition, authentication speed of four-digit PINs was significantly faster than of force-PINs ($p < 0.001$). The difference in authentication speed between six-digit and force-PINs was not statistically significant ($p = 0.12$).

Errors.

An important factor when estimating the overhead of an authentication method is the number of errors. Similar to

Table 2: Mean authentication time in seconds and error rate with different levels of the independent variables.

Authentication Speed	Mean	SD
4-digit	2.34	1.21
6-digit	3.33	1.56
Force	3.66	1.96
Error Rate	**Basic**	**Critical**
4-digit	21	0
6-digit	22	0
Force	36	4

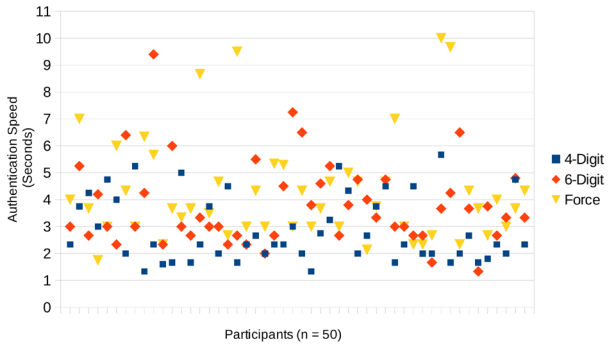


Figure 2: Mean authentication time per participant.

De Luca et al. [15], we distinguished between basic and critical errors. For our authentication scenario, we defined a *basic* error as an erroneous attempt to enter a PIN code. An authentication session can be successful overall, but may take a user two or three times to enter the PIN correctly. We considered an error as *critical* if the entire authentication failed, i.e., a user was locked out after three erroneous attempts as commonly deployed in off-the-shelf smartphone operating systems.

Out of 450 total authentication sessions, four authentication sessions failed (0.9%). All failed sessions involved force-PINs. 36 (8.0%) failed attempts (basic errors) were registered with force-PINs. 22 (4.8%) failed attempts were registered with six-digit PINs and 21 (4.6%) with four-digit PINs.

4.2.2 Perceived Usability and Security

As explained above, participants were asked to fill out a short questionnaire after completing the PIN selection and authentication tasks. In addition to the measurements collected via our iPhone apps, we were interested in participants' perceptions of the three suggested PIN types regarding usability and security. We presented users with closed-ended questions asking which PIN type they thought was the easiest/hardest to remember, fastest/slowest and most/least error-prone to enter and generally most/least secure. The results of these questions are shown in Figure 3.

91% of our participants reported that they thought four-digit PINs were the least secure of the three tested PIN types. 95% also thought that four-digit PINs were the fastest PIN type to enter and 80% thought that they were the easiest to remember. 62% thought that force-PINs were the most secure of the three methods but 55% also thought that this was the most time-consuming PIN type to enter. In

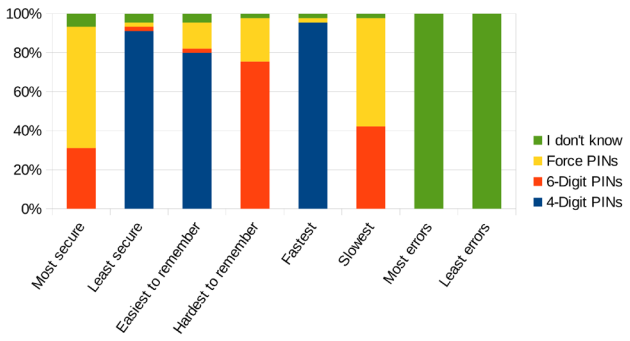


Figure 3: Self-reported usability and security estimation in percent.

comparison, only 31% thought that six-digit PINs were the most secure but 75% also thought that they were the hardest to remember.

To our surprise, all participants chose the "I don't know" option regarding most and least errors when entering any of the suggested PIN types.

On the last page of the online survey, we asked participants three open-ended questions related to their perception of force-PINs. This was the only part of our study where force-PINs received particular attention. These questions were asked at the very end of our lab sessions to minimize the risk of biased interpretation.

After coding the data segments collected through these questions, we found that 38 of the 50 participants thought that a major benefit of force-PINs was the resistance against observation due the haptic and invisible component. 10 participants also stated that they think force patterns are easier to remember than additional digits, as would be the case with longer PINs. Eighteen participants reported that they still think that it requires additional effort to enter digits with different levels of force as they are still not used to this new interaction method with touchscreens.

4.2.3 Informal Participant Statements

In this section, we present informal participant statements and also quote some of the qualitative statements gathered via the open-ended questions from our post-experiment survey. These direct quotes are presented as they were given by the participants prior to coding.

Overall, we were surprised by how easy it was to recruit participants irrespective of the promised reward. We had the impression that all of them found the topic of PIN security important. Based on their comments, we had the impression that most of them seemed to be aware of the richness of private data stored on their smartphones. Most participants also asked for further help in protecting their devices after participating in our study. After their participation, they were given the opportunity to have their questions answered by the experimenters. Even though a few authentication sessions with force-PINs failed, all participants understood the concept of force-PINs and were able to use them. To our surprise, the participants found the concept natural and intuitive even though most of them were using pressure-sensitive touchscreens for the first time.

- *"I like the additional dimension. It is invisible and therefore makes my PIN more secure." (P5)*

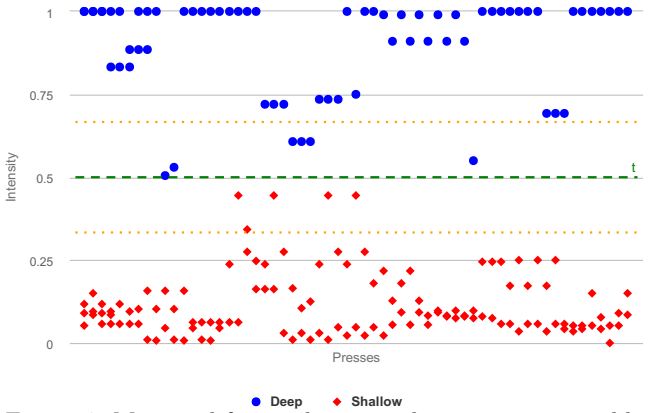


Figure 4: Measured force relative to the maximum possible force. The green line at y = 0.5 represents the threshold for distinguishing between deep and shallow presses. The grey lines at 0.25 and 0.75 indicate two potential thresholds for a three-step force scale (e.g., *shallow-medium-deep*.)

- *"If someone observes me entering my PIN, which is not that secure and probably easy to guess, at least the force component is harder to guess. (P28)"*
- *"I think it might take a while to fully get used to it, as this concept is new to me. (P23)"*
- *"Why not use a six-digit force-PIN? (P12)"*

4.2.4 Force Pressure

As stated in Section 3, we based our design for a two-step scale on our pre-testing with people who had never used *3D Touch* before. Due to the low experience with pressure-sensitive screens, they could not easily distinguish different thresholds to separate deep and shallow press. The app also provided vibration feedback as soon as the user entered a digit with force. Through our lab study, we collected the exact values of the force registered by the device and then used it to evaluate how close or far the registered force was from the threshold and the upper and lower boundaries. Figure 4 shows the force intensities of all logged force-PIN digits during the lab study in percent of the maximum possible force.

5. SECURITY EVALUATION

Based on the data collected during the lab study, we performed an additional security evaluation to evaluate shoulder-surfing resistance and PIN entropy.

5.1 Shoulder Surfing

To evaluate our approach to the attacker model, we performed a small shoulder-surfing experiment in the lab. Similar to the study design of De Luca et al. [15] and von Zeschwitz et al. [26], the attacker tried to shoulder surf the force-PIN entry from the victim. For our evaluation, we considered direct observation, i.e., the attacker was physically standing behind the victim and tried to guess the entered force-PIN and then performed an additional evaluation based on separately recorded video material. Our evaluation is based on the 50 force-PINs which were collected in the course of our lab study and then used for our evaluation of authentication speed and error-rate.

The direct observation attack was performed during the lab study. One experimenter acted as a shoulder surfer and was in close proximity to the victim. Our participants were aware of their entered PINs being tracked via the device used during the experiments but they were not told that one of the experimenters acted as a shoulder surfer. The shoulder-surfing experimenter was perceived as trustworthy. Therefore, the participants did not apply additional measures to prevent their PINs from being observed. We chose this experimental setting as we believe that situations where victims are not aware of being observed are the most dangerous. We furthermore believe that any authentication method should be resilient to direct observation regardless of a specific situation and the user's awareness. In addition, an experimenter entered the collected PINs with their corresponding force patterns while being filmed. Each PIN was entered only once. Another two volunteers, who were university students (one male, one female), then tried to guess the force-PINs based on the recorded material. Each of them tried to guess 25 PINs. They were allowed to re-watch the video sequence up to 5 times if they wanted to.

This first look at shoulder-surfing resistance suggests that force-PINs are capable of making digit PINs more resilient against shoulder-surfing attacks. Out of the 50 entered force-PINs, the shoulder surfer was not able to guess a single one completely. However, 21 out of 50 PINs were partially guessed (i.e., the attacker correctly guessed the digits but not the force pattern). Similar to the direct observation attacks, the attackers in the camera-based attacks were not able to completely guess the force-PINs from the recorded material, but managed to guess 39 of the shown digit sequences correctly. We did not evaluate whether individual digits (with or without force) were guessed correctly.

5.2 Entropy

In theory, the PIN space of four-digit force-PINs is larger than for standard four-digit and smaller than six-digit PINs. In our lab study, we used user-assigned PINs. We gave participants a password policy, namely to choose a PIN that, in their opinion, is as secure yet as memorable as possible and where at least one digit within the four digit pattern is entered with a deep press.

Obviously, the number of possible combinations is $10^4 = 10,000$ for four digit passwords and $10^6 = 1,000,000$ for six digit passwords. Force-PINs augment the four-digit password space to $20^4 = 160,000$ possible PIN codes including four-digit PINs with all digits entered with shallow pressure. As we defined a policy for the lab study which forced participants to choose at least one digit with deep pressure, the password space decreases to $150,000$.

As done by Cherapau et al. [12], we calculate the zero-order entropy, which is a theoretical measure of the entire search space of all possible secrets of a given length and the size of a given alphabet assuming that each character is selected randomly. Zero-order entropy is measured in bits and calculated as $L * \log_2 N$, where L is the length of the secret and N the size of the character set. Hence, for force-PINs, the length is 4 and the character set 20. Thus, the zero-order entropy for force-PINs is 17.28 bits, while four-digit PINs have a zero-order entropy of 13.28 [12] and six-digit PINs 19.93 bits. These theoretical measures are upper bounds for real-world entropy.

In theory, the augmented PIN space is a major improvement compared to standard four-digit PINs. In practice however, users often do not fully exploit this benefit but se-

lect PIN codes and passwords from a much smaller subset that are often easy to predict [21]. Therefore, the search space is smaller and the PIN is therefore easier for an attacker to guess. We therefore evaluate the distribution of force patterns and digit-pressure combinations.

Table 3 shows the occurrences of force patterns selected by our participants. Our results suggest that more than half of our participants selected a force pattern where only a single digit is entered with deep press. In our sample, the most popular positions in the digit sequence were the first and second one with a probability of 14.0%. Even though this trend indicates that our participants did not fully make use of the theoretically larger PIN space and therefore create lower entropy PINs in practice, this is already an improvement over standard four-digit PINs. Our dataset of 56 PINs is relatively small and therefore not sufficient to determine the practical entropy of force-PINs. To provide a rough indicator, we calculate the entropy of the binary force component based on the force-PINs chosen by our study participants. Furthermore, to estimate the entropy gain over digit-only PINs, we compare our results to those from a related study on iPhone passcodes with a larger sample size. In theory, if force patterns were evenly distributed, the theoretical entropy gain would be 4 bits. We calculate the practical entropy gain as $-\sum_{i=1}^{n} p_i * log_2(p_i)$ where p_i is the probability of a certain pattern occurring. Based on our observed probabilities from 56 user-chosen force patterns (as presented in Table 3), the practical entropy gain is 3.41 bits. Bonneau et al. [7] calculated the entropy of four-digit PINs from iPhone users as 11.42 based on a dataset of 204,508 PINs. Comparing our findings with Bonneau et al. [7], an additional binary force component provides an entropy gain of approximately 23% to digit-only PINs of length 4.

Table 3: Force patterns selected by the lab study participants where S = shallow press, D = deep press. n = 56 user-selected PINs. The table is sorted in descending order. The pattern SSSS was excluded as the PIN selection policy required participants to enter at least one digit with deep press.

Force Pattern	Number	Percent
DSSS	8	14.0%
S**D**SS	8	14.0%
SSS**D**	7	12.2%
SS**D**S	6	10.5%
DSS**D**	6	10.5%
S**DD**S	5	8.7%
DDDD	5	8.7%
SS**DD**	4	7.0%
S**DDD**	2	3.5%
S**D**S**D**	2	3.5%
DDSS	1	1.7%
DS**D**S	1	1.7%
DDS**D**	1	1.7%
DDDS	0	0.0%
DS**DD**	0	0.0%

6. FIELD STUDY

In addition to the lab study, we conducted a field study to show that authentication time for four-digit force-PINs decreases with training. The latter is an important metric when comparing the usability performance of digit-only PINs with force-PINs as we assume that users will initially perform better with digit-only PINs as they are already trained to use them.

Table 4: Digits and their occurrence entered with either shallow or deep press. Deep pressed digits are in bold; sorted in descending order.

Digit (shallow/deep press)	Number
1 (shallow)	27
0 (shallow)	22
5 (shallow)	16
4 (shallow)	15
3 (shallow)	14
2 (shallow)	12
0 (deep)	12
1 (deep)	12
6 (shallow)	11
2 (deep)	10
9 (deep)	10
3 (deep)	9
6 (deep)	9
9 (shallow)	8
4 (deep)	7
7 (shallow)	6
7 (deep)	6
8 (deep)	6
5 (deep)	6
8 (shallow)	5

6.1 Study Design and Procedure

We recruited 10 participants and deployed an iOS app on their personal devices and asked them to enter as many force-PINs as possible (we required a minimum of 300 successful authentication sessions) over a period of two weeks. At the end of this period, we conducted short debriefing interviews with the participants. In contrast to the lab study, the participants were aware that the focus of the study was to evaluate force-PINs.

Due to the low propagation of compatible iPhones in our region, we were able to recruit only 10 participants. In spite of the relatively low number of participants, we still believe that the gathered data provides useful insights and rough indicators on learning effects. Furthermore, deploying force-PINs in a real-world environment helped us to gather in-situ reactions on authentication problems with force-PINs.

We based our study design on findings from Harbach et al. [19], who found that users unlock their phone on average 47.8 times a day (about three unlocks per hour assuming a user is awake for 16 hours per day).

Due to the restrictions in iOS, we were not able to replace the actual PIN scheme on the participants' devices with force-PINs. We also had to reject our plan to issue notifications based on the participants' unlocking behavior as iOS does not offer to activate third-party apps after an unlock event. Therefore, we were not able to collect the respective data from the users' own devices. As everyday routines and smartphone usage habits are highly diverse, we refrained from requiring force PIN entries at fixed timepoints throughout the day and opted for a more realistic and

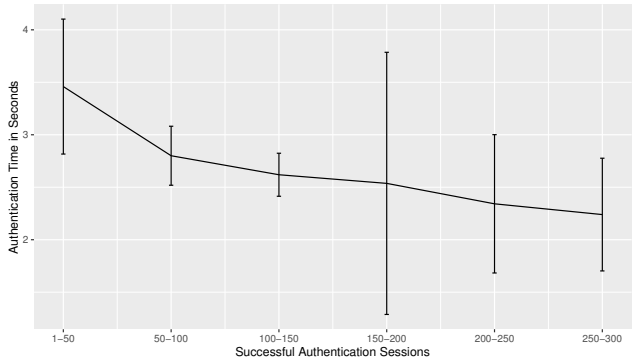


Figure 5: Authentication time development based on the first 300 successful authentication sessions across all participants.

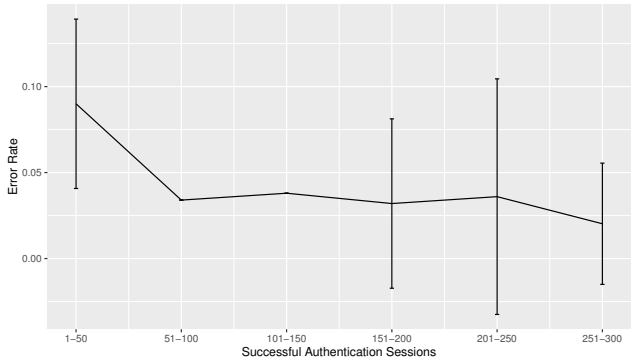


Figure 6: Error rate development (basic errors) based on the first 300 successful authentication sessions across all participants.

less disruptive setting. To evaluate different timing options for notifications, we conducted a small pilot study with different notification patterns. The participants from this pilot study perceived the notifications as disruptive and annoying regardless of whether they were issued at fixed or adaptive time points. Based on the participants' responses, we decided to reduce the number of daily notifications to a single daily reminder at an arbitrary point in time and left it up to the participants when and how often to entered their force-PINs. We are confident that this study design reflects realistic usage habits and reduced the risk of participants dropping out early from the study.

We instructed our participants to enter force-PINs whenever they took out their phone before or after their primary task. We suggested they distribute the PIN entries over the given period of time (i.e., about 20 PINs a day), but also told them that it was their own decision when exactly and how often to enter them. The participants were also instructed to choose as secure and memorable PINs as possible with at least one digit entered with force.

The main screen of our app had a button that redirected the participants to a lock screen to start an authentication session with a force-PIN. It was designed to look exactly like the standard iPhone lock screen. Our app also displayed a counter of successful authentication sessions and provided users with two extra buttons, one to send us an e-mail in case of questions and another one to leave a comment to a situation. We also provided users with an option on the main screen to set a new force-PIN. Upon clicking on this button, a password-forgotten event was logged and the participants were able to set a new force-PIN.

6.2 Results

Overall, our participants successfully completed 3,748 authentication sessions with force-PINs. The results are summarized in Table 5. Among the successful sessions, 254 failed attempts (basic errors) were registered and five participants had entirely failed authentication sessions (critical errors). The number of critical errors (i.e., failed authentication sessions) was low. The entirely failed authentication sessions were registered at the very beginning of the study. The error rates in Table 5 are given in percent of authentication sessions completed by the user. For the quantitative analysis, we removed authentication sessions that lasted longer than 30 seconds from our sample. As observed in our lab study, authentication sessions longer than 30 seconds usually occurred when the participant was interrupted or distracted from the study task.

The mean authentication speed over all authentication sessions was 2.69 seconds (median=2.26, SD=0.59), which is an improvement over the results from the lab study. The shortest authentication session was only 1.02 seconds long. In comparison, Harbach et al. [18] determined the average authentication speed for digit-only PINs as 1.9 seconds.

All participants attended the debriefing session and participated in the debriefing interviews. One participant did not complete the initially requested 300 successful authentication sessions and had only 210 completed authentication sessions. Although this did not meet our desired goal, we included the data and conducted the debriefing interview with the participant as the number of participants was low.

Just like in the lab study, we measured the authentication time of each session as time from the first touch until the user was successfully authenticated (including potentially unsuccessful attempts made during the session). As per the study design, we expected the PIN entries to be unevenly distributed over time across the participants. Our results show that the participants did not make use of the given time and completed the study task in a few days regardless of our daily notifications. Five participants completed their authentication sessions on a single day. They distributed their PIN entries over the morning, late afternoon and the evening of that day. Four participants completed the study task in two or three days and entered their PINs mostly in the morning and late afternoon/evening of these days. One participant spent four days on the study task and distributed the PIN entries over various times of the day. We therefore refrain from a time-based analysis and compare the results based on authentication sessions.

For our analysis of authentication time and error rate, we consider the first 300 successfully completed authentication sessions from all participants. In order to visualize a trend over multiple completed authentication sessions, we grouped the results in bins of 50 sessions across all participants. We selected a bin size of 50 to approximate the average number of phone unlocks per day as determined by Harbach et al. [19]. We believe that this is a good way to simulate a trend over a reasonable period of time. Figure 5 provides a comparison of the average authentication time grouped by 50 successful authentication sessions based on the median

Table 5: Summary of field study results. n=10

Subjects	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10
Completed Authentication Sessions	534	336	453	387	407	335	210	386	343	357
Basic Errors	13	41	69	20	4	26	16	17	21	27
Basic Error Rate	2.4 %	12.2 %	15.2 %	5.2 %	0.9 %	7.7 %	7.6 %	4.4 %	6.1 %	7.6 %
Critical Errors	0	3	0	0	0	1	1	1	1	0
Critical Error Rate	0 %	0.8 %	0 %	0 %	0 %	0.2 %	0.4 %	0.2 %	0.2 %	0 %
Forgot Force-Pin	0	2	1	0	0	0	0	0	0	0
Force-Pins	**5225**	**0229**	1234	**5795**	**5968**	**0000**	1**70**3	**0171**	22**0**4	9**9**99
	-	0**229**	741**2**	-	-	-	-		-	-
	-	**1397**	-	-	-	-	-	-	-	-

authentication time per participant. These results suggest that the authentication time decreases with training. Figure 6 shows that the error rate also decreases with training.

6.3 Debriefing Interviews

During the debriefing sessions, we asked the participants in which situations they used force-PINs and whether they found them feasible in these scenarios. According to the participants, most force-PINs where entered either while they were at home, in their office, or on public transport. Eight participants reported that they found force-PINs a good way to protect their digit PINs from shoulder surfers even though they estimated their susceptibility towards direct observers as relatively low. Three participants said that they would like to use force-PINs to make their existing PINs more secure against close intruders such as family and friends who could easily guess their PIN as it was an important date. According to them, the risk of a close acquaintance spying on their phones was higher than that of shoulder surfing attacks in public spaces.

Nine participants reported that their perceived authentication time decreased with training when they used it several times a day. However, five of them reported that they still think that simple digit PINs are faster for authentication. All participants reported that they did not find force-PINs harder to remember than simple digit PINs.

Participants were also asked if they would prefer to use force-PINs over simple digit PINs. All of them said that they generally liked the idea of an additional invisible component and six participants said that they would maybe use them if deployed on their device. Eight participants reported that they found the training phase in the beginning annoying. Three expressed interest in multiple-step pressure difference.

7. DISCUSSION

Previous research [19] has shown that the task overhead of smartphone authentication is relatively high. Therefore, we argue that the overhead of a technology to replace simple digit PINs should not be higher than the state of the art.

The results from our lab study suggest that the task overhead of force-PINs is initially higher than for digit-only four- and six-digit PINs. Our security analysis and the participants' responses indicate that force-PINs can increase PIN entropy and improve the resilience towards shoulder-surfing attacks. The results from our field study revealed learning effects after a certain number of interactions with the invisible component, and indicate that authentication time and error rate decrease with training and converge towards the metrics for four-digit PINs.

We collected evidence on frequently used force patterns and determined a practical entropy gain of 3.41 bits based on the force-PINs chosen by our study participants. Similar to other user-chosen secrets, the practical entropy does not meet the theoretical measures but still suggests a major improvement when compared to entropy estimations of digit-only PINs.

Apart from the metrics we used to evaluate the performance of the respective PIN types, the self-reported data from our participants suggests that force-PINs were perceived as more secure than six-digit PINs. The open-ended survey questions revealed that this was mainly due to the force component, which our participants perceived as a good countermeasure against observation.

Only two participants forgot and renewed their force-PINs from the field study. The number of critical errors was also low.

Hence, our results suggest that our scheme is able to improve security with a reasonably low impact on task overhead. In comparison to other solutions, our design improves security without requiring the user to memorize longer sequences of digits, which have been shown to be more difficult to remember [20].

To our surprise, none of the 50 participants provided an estimation of which of the PIN schemes was most/least error prone. While our collected data does not explain reasons, we believe that this is because of the manifold sources of errors: As authentication sessions in the wild usually take place in diverse situations, their successful completion is influenced by environmental and situational constraints beyond the design of the authentication method.

According to a study by Harbach et al. [19], users are generally aware of risky situations but this does not influence their general opinion about this threat, which is that this risk is only considered in a low number of everyday situations. However, just because users do not perceive situations as risky does not mean that they are not. Hence, physically shielding the PIN from an observer can only mitigate an attack if the user is aware of the threat and therefore actively taking precautions. Our results suggest that force-PINs can help to protect users from shoulder surfers regardless of their risk awareness, while minimizing the additional effort the user has to invest.

Modern smartphones offer biometric authentication as an alternative. While supporters of these methods often argue that they are harder to replicate and therefore not susceptible to shoulder surfing, it is commonly acknowledged by the scientific community that these methods are non-revocable and can easily be broken [1, 11]. Furthermore, they still rely on passwords for fallback authentication. These examples highlight that it is worth putting effort into making knowledge-based authentication resilient to shoulder surfing.

Our prototype app was implemented for iPhone 6s. Other smartphone models, such as the Huawei Mate S, also have pressure-sensitive screens and are therefore suitable for force-PINs. Furthermore, force patterns like in force-PINs can also be added to character/digit passwords with variable length and Android unlock patterns to make them resilient to shoulder surfing attacks. As future work and as soon as a compatible API and device are available in our region, we plan to evaluate force patterns in combination with unlock patterns and other alternative authentication schemes, respectively.

8. LIMITATIONS

We now discuss limitations of our methodology and the conducted studies.

As we recruited our participants at the university campus, the level of education and technology affinity among our sample were higher than expected from the general population. As the results might differ for other demographics, our results cannot be generalized to the entire population of smartphone users. Since only 28% of the participants in the lab study were iPhone users, we cannot determine whether the measurements based on their input were biased by the lack of practice. However, as this study had a repeated-measures design, we were able to perform a comparative evaluation of all subjects exposed to our conditions. All participants in our field study took part with their own devices and had therefore been exposed to a force-sensitive screen before and were already familiar with the iOS user interface and lock screen, respectively.

It is possible that users would improve even more over a longer period of time and usability metrics converge to those of four-digit standard PINs. Regardless of our suggestion to distribute the authentication sessions over the two weeks, the participants tried to complete the study task as fast as possible and therefore entered all force-PINs within the first three days. Also, the number of successful authentication sessions varies widely across the participants. As the participants did not spread out the PIN entries over the given time, we can neither perform a time-based evaluation nor seriously evaluate memorability. The fact that our participants from the lab study thought that force-PINs are more memorable speaks for the system but does not obviate the need for a future long-term evaluation. Regardless of these limitations, we are confident that our study design reflects real-world usage behavior and due to its flexibility ensured that participants would not drop out early.

A major limitation of this work is that the participants from both the field study and the shoulder surfing experiment participated voluntarily and did not receive a compensation for their participation. Therefore, the motivation for the shoulder surfers was rather low to actually break the system. Another limitation is that they were new to the concept of force-PINs and therefore perceived the task as particularly challenging. Also, force-PINs do not provide visual feedback and the vibration for digits entered with force is very subtle and therefore not audible on the video material. The participants reported finding it hard to focus on both the digits and the force patterns. The person who entered the PINs in front of the camera was a faculty member who was aware of the hypothesis being tested just like the experimenter who tried to shoulder surf the PINs from the lab study. These limitations imply that further investigation is needed to determine a lower bound for shoulder surfing resistance.

9. ETHICAL CONSIDERATIONS

Our university does not have an ethics board but has a set of guidelines that we followed in our research. A fundamental requirement of these guidelines is to preserve the participants' privacy and to limit the collection of person-related data as far as possible. For both our studies, we did not collect any personally identifiable information, except for age and gender. A major ethical challenge was the collection of PINs. The PINs were chosen by the participants and they were aware that the PINs they selected were being collected. However, we cannot preclude that those were real PINs. Keeping this data confidential and making it impossible to map a physical person with a certain PIN was therefore our primary concern. In similar shoulder surfing studies, participants were re-corded with video cameras to perform attacks based on the recorded material. Although this was our initially planned study setting, we decided not to film the participants directly while they entered their PINs. This decision was made based on the results and feedback from our pilot study, where our participants expressed discomfort about being filmed while entering information as sensitive as a PIN. We therefore chose to let a separate person enter all force-PINs in front of a camera and then used the resulting material for our camera attacks.

10. CONCLUSION

In this work, we proposed integrating pressure-sensitive touchscreen interactions into knowledge-based authentication. These force-PINs enhance digit-only PINs with a force pattern, i. e., an additional pressure-sensitive component that allows users to select higher entropy PINs that are harder for a shoulder surfer to observe.

We were able to collect evidence on the security benefits of force-PINs and their impact on usability. We conducted a lab study with 50 participants and showed that authentication speed of force-PINs is not significantly slower than that of six-digit standard PINs, but still significantly slower than that of 4-digit standard PINs. We also showed that the error rate is rather low in spite of the fact that most participants had not yet been exposed to pressure-sensitive touchscreen interaction. Furthermore, we conducted a small shoulder-surfing study where an attacker tried to observe and guess force-PINs. The attackers were not able to guess a full force-PIN consisting of a digit sequence and a force component. These results suggest that force-PINs can help to mitigate shoulder-surfing attacks in public spaces that are potentially noisy and crowded. In a security evaluation of the collected force-PINs, we showed that the practical entropy is still higher than for standard four-digit PINs although users do not make full use of the larger PIN space. In

an additional field study with 10 participants, we deployed force-PINs in the wild and showed that users improve after being exposed to the technology over a longer period of time.

Our results imply that small enhancements such as an additional pressure component allow users to select higher entropy PINs that are more resilient to shoulder-surfing attacks, while keeping the impact on usability metrics such as authentication speed and error rate low. This is important as users enter their PINs multiple times a day and therefore require methods that do not increase the task overhead.

11. ACKNOWLEDGMENTS

We would like to thank the reviewers for their constructive feedback. We would also like to thank our shepherd Marian Harbach for his suggestions that were very helpful in improving our paper. This research was partially funded by COMET K1, FFG - Austrian Research Promotion Agency.

12. REFERENCES

[1] Android Authority. Android Jelly Bean Face Unlock 'liveness' check easily hacked with photo editing. http://www.androidauthority.com/android-jelly-bean-face-unlock-blink-hacking-105556/, last accessed 2/10/2016.

[2] Aviv, Adam J and Gibson, Katherine and Mossop, Evan and Blaze, Matt and Smith, Jonathan M. Smudge Attacks on Smartphone Touch Screens. *WOOT*, 10:1–7, 2010.

[3] Bianchi, Andrea and Oakley, Ian and Kostakos, Vassilis and Kwon, Dong Soo. The phone lock: audio and haptic shoulder-surfing resistant PIN entry methods for mobile devices. In *Proceedings of the fifth international conference on Tangible, embedded, and embodied interaction*, pages 197–200. ACM, 2011.

[4] Bianchi, Andrea and Oakley, Ian and Kwon, Dong Soo. Spinlock: a single-cue haptic and audio PIN input technique for authentication. In *Haptic and Audio Interaction Design*, pages 81–90. Springer, 2011.

[5] Bianchi, Andrea and Oakley, Ian and Kwon, Dong Soo. Counting clicks and beeps: Exploring numerosity based haptic and audio PIN entry. *Interacting with computers*, 24(5):409–422, 2012.

[6] Bianchi, Andrea and Oakley, Ian and Lee, Jong Keun and Kwon, Dong Soo and Kostakos, Vassilis. Haptics for tangible interaction: a vibro-tactile prototype. In *Proceedings of the fifth international conference on Tangible, embedded, and embodied interaction*, pages 283–284. ACM, 2011.

[7] J. Bonneau, S. Preibusch, and R. Anderson. A birthday present every eleven wallets? the security of customer-chosen banking pins. In *Financial Cryptography and Data Security*, pages 25–40. Springer, 2012.

[8] Bonneau, Joseph and Herley, Cormac and Van Oorschot, Paul C and Stajano, Frank. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 553–567. IEEE, 2012.

[9] A. Bragdon, E. Nelson, Y. Li, and K. Hinckley. Experimental analysis of touch-screen gesture designs in mobile environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 403–412. ACM, 2011.

[10] Buschek, Daniel and De Luca, Alexander and Alt, Florian. Improving Accuracy, Applicability and Usability of Keystroke Biometrics on Mobile Touchscreen Devices. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1393–1402. ACM, 2015.

[11] Chaos Computer Club. Chaos Computer Club breaks Apple TouchID. http://www.ccc.de/en/updates/2013/ccc-breaks-apple-touchid, last accessed 11/11/2015.

[12] Cherapau, Ivan and Muslukhov, Ildar and Asanka, Nalin and Beznosov, Konstantin. On the impact of touch id on iphone passcodes. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 257–276, 2015.

[13] S. Chowdhury, R. Poet, and L. Mackenzie. Passhint: Memorable and secure authentication. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 2917–2926, New York, NY, USA, 2014. ACM.

[14] De Luca, Alexander and Hang, Alina and von Zezschwitz, Emanuel and Hussmann, Heinrich. I feel like i'm taking selfies all day!: Towards understanding biometric authentication on smartphones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI*, volume 15, pages 1411–1414, 2015.

[15] De Luca, Alexander and Harbach, Marian and von Zezschwitz, Emanuel and Maurer, Max-Emanuel and Slawik, Bernhard Ewald and Hussmann, Heinrich and Smith, Matthew. Now you see me, now you don't: protecting smartphone authentication from shoulder surfers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2937–2946. ACM, 2014.

[16] De Luca, Alexander and Lindqvist, Janne. Is Secure and Usable Smartphone Authentication Asking Too Much? *Computer*, 48(5):64–68, 2015.

[17] De Luca, Alexander and Von Zezschwitz, Emanuel and Nguyen, Ngo Dieu Huong and Maurer, Max-Emanuel and Rubegni, Elisa and Scipioni, Marcello Paolo and Langheinrich, Marc. Back-of-device authentication on smartphones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2389–2398. ACM, 2013.

[18] M. Harbach, A. De Luca, and S. Egelman. The Anatomy of Smartphone Unlocking. In *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems, CHI*. 2016.

[19] Harbach, Marian and von Zezschwitz, Emanuel and Fichtner, Andreas and De Luca, Alexander and Smith, Matthew. It's a hard lock life: A field study of smartphone (un) locking behavior and risk perception. In *Symposium on Usable Privacy and Security (SOUPS)*, 2014.

[20] Huh, Jun Ho and Kim, Hyoungshick and Bobba, Rakesh B and Bashir, Masooda N and Beznosov, Konstantin. On the Memorability of System-generated PINs: Can Chunking Help? In *Eleventh Symposium*

On Usable Privacy and Security (SOUPS 2015), pages 197–209, 2015.

[21] P. G. Kelley, S. Komanduri, M. L. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, and J. Lopez. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 523–537, 2012.

[22] Khan, Hassan and Atwater, Aaron and Hengartner, Urs. A comparative evaluation of implicit authentication schemes. In *Research in Attacks, Intrusions and Defenses*, pages 255–275. Springer, 2014.

[23] R. Kuber and W. Yu. Tactile vs graphical authentication. In *Haptics: Generating and Perceiving Tangible Sensations*, pages 314–319. Springer, 2010.

[24] Malek, Behzad and Orozco, Mauricio and El Saddik, Abdulmotaleb. Novel shoulder-surfing resistant haptic-based graphical password. In *Proc. EuroHaptics*, volume 6, 2006.

[25] Song, Youngbae and Cho, Geumhwan and Oh, Seongyeol and Kim, Hyoungshick and Huh, Jun Ho. On the Effectiveness of Pattern Lock Strength Meters: Measuring the Strength of Real World Pattern Locks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2343–2352. ACM, 2015.

[26] von Zezschwitz, Emanuel and De Luca, Alexander and Brunkow, Bruno and Hussmann, Heinrich. SwiPIN: Fast and secure pin-entry on smartphones. In *Proceedings of the Conference on Human Factors in Computing Systems, CHI*, volume 15, pages 1403–1406, 2015.

APPENDIX

A. LAB STUDY QUESTIONNAIRE

The following questions were answered by the participants of the lab study after they used the three different types of PINs in a randomized order (four-digit/six-digit/force-PIN).

Demographics.

1. What was your ID during the lab experiments?
2. Gender
3. Age
4. Are you studying IT security or are you working in an IT security-related field? *(yes/no)*
5. What kind of smartphone are you currently using? *(single-choice: iPhone, Android, Windows Phone, Other, I don't use a smartphone)*
6. What methods are you currently using to unlock your smartphone? *(multiple-choice: 4-digit PINs, 6-digit PINs, character and digit password, unlock pattern, fingerprint sensor, Android Smartlock, none)*

Estimated security and usability of the three PIN types.

1. Which of the three PIN methods do you think is the most secure? *(single-choice: 4-digit PINs, 6-digit PINs, force-PINs, I don't know)*
2. Which of the three PIN methods do you think is the easiest to remember? *(single-choice: 4-digit PINs, 6-digit PINs, force-PINs, I don't know)*
3. Which of the three PIN methods do you think is the least secure? *(single-choice: 4-digit PINs, 6-digit PINs, force-PINs, I don't know)*
4. Which of the three PIN methods do you think is the most time-consuming? *(single-choice: 4-digit PINs, 6-digit PINs, force-PINs, I don't know)*
5. Which of the three PIN methods do you think is the hardest to remember? *(single-choice: 4-digit PINs, 6-digit PINs, force-PINs, I don't know)*
6. Which of the three PIN methods do you think is the least time-consuming? *(single-choice: 4-digit PINs, 6-digit PINs, force-PINs, I don't know)*

Open-ended questions.

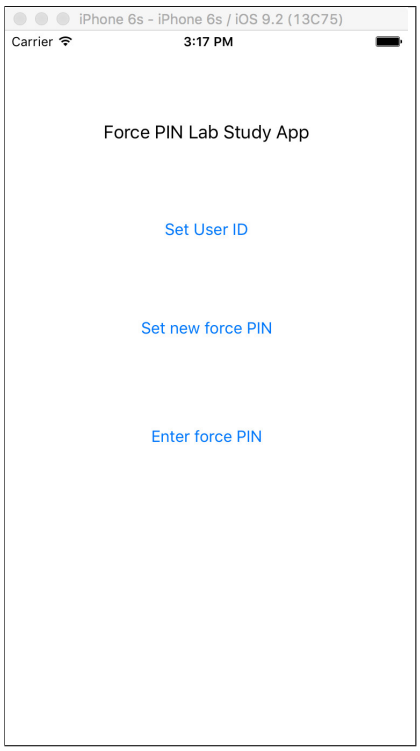
1. What did you like about force-PINs?
2. What did you NOT like about force-PINs?
3. Can you think of a situation where force-PINs would be particularly useful?

B. FIELD STUDY DEBRIEFING INTERVIEWS

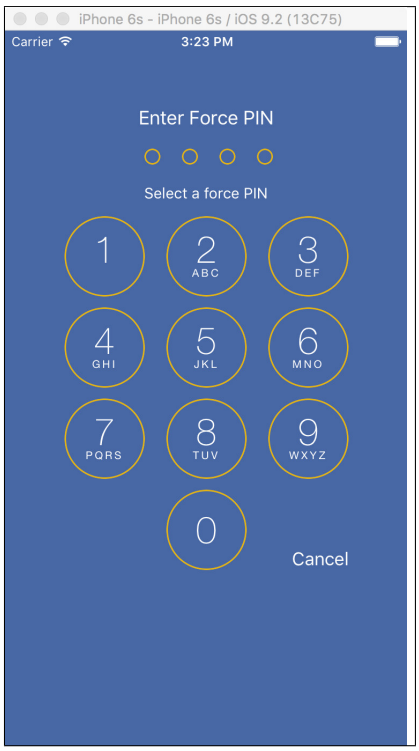
1. Where did you use force-PINs?
2. What did you like about force-PINs?
3. What did you NOT like about force-PINs?
4. Can you think of a situation where force-PINs were particularly useful?
5. Can you think of a situation where force-PINs were annoying?
6. Is there anything else you would like to let us know?

C. STUDY APPS

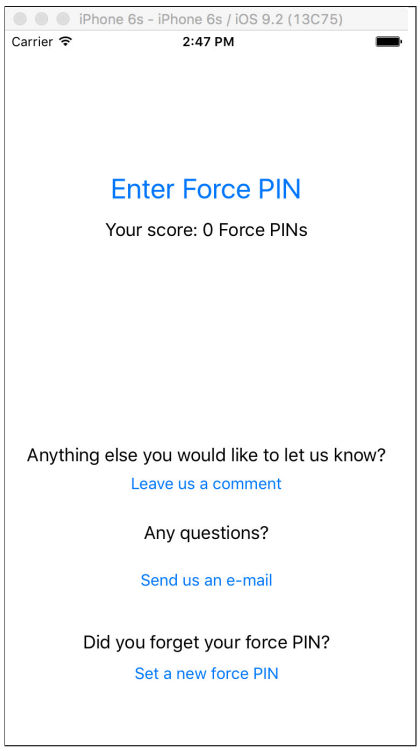
The following screenshots show the user interface of the apps used for the lab and field study. Figure 7a and Figure 7b were used to evaluate force-PINs in the lab study. The apps for the other two conditions had the same layout but evaluated four-digit and six-digit PINs, respectively. Figure 7c shows the main screen of the app used in the field study.



(a) Main screen of the lab study app.



(b) Lock screen of both the lab study and field study app.



(c) Main screen of the field study app.

Figure 7: Screenshots of the study force-PIN apps.

Ask me again but don't annoy me: Evaluating re-authentication strategies for smartphones

Lalit Agarwal, Hassan Khan and Urs Hengartner
Cheriton School of Computer Science
University of Waterloo
Waterloo, ON Canada
{lagarwal, h37khan, urs.hengartner}@uwaterloo.ca

ABSTRACT

Re-authenticating users may be necessary for smartphone authentication schemes that leverage user behaviour, device context, or task sensitivity. However, due to the unpredictable nature of re-authentication, users may get annoyed when they have to use the default, non-transparent authentication prompt for re-authentication. We address this concern by proposing several re-authentication configurations with varying levels of screen transparency and an optional time delay before displaying the authentication prompt. We conduct user studies with 30 participants to evaluate the usability and security perceptions of these configurations. We find that participants respond positively to our proposed changes and utilize the time delay while they are anticipating to get an authentication prompt to complete their current task. Though our findings indicate no differences in terms of task performance against these configurations, we find that the participants' preferences for the configurations are context-based. They generally prefer the re-authentication configuration with a non-transparent background for sensitive applications, such as banking and photo apps, while their preferences are inclined towards convenient, usable configurations for medium and low sensitive apps or while they are using their devices at home. We conclude with suggestions to improve the design of our proposed configurations as well as a discussion of guidelines for future implementations of re-authentication schemes.

1. INTRODUCTION

The increased usage of smartphones to access personal and corporate data requires authentication at multiple levels. A device-level authentication scheme, such as a PIN or fingerprint recognition, is required to protect access to the device while text-based passwords may be required to further establish identity for social networking, banking or enterprise apps. Existing studies have shown that the short and frequent nature of smartphone sessions creates usability issues for device-level authentication schemes [17] whereas constrained keyboards on smartphones are a bottleneck when users are authenticating using text-based passwords [29]. To mitigate these usability issues, researchers have proposed several techniques that reduce the authentication burden by leveraging user behaviour [21, 32, 37], device context [16, 24, 25] or the sensitivity of launched apps [17].

While these schemes reduce the authentication burden on the user, they may require mid-task re-authentication. Schemes that leverage user behaviour need re-authentication in case of a behaviour mismatch against the current phone user. Similarly, device context-based schemes may need to establish a user's identity in case a contextual source (e.g., ambient noise) changes. Taking the sensitivity of launched apps into account for authentication may also require mid-task re-authentication. For instance, some users have indicated that for a messenger app only opening old messages should trigger re-authentication [17].

Preliminary evaluations show that users like the convenience offered by these schemes [4, 16, 17, 19, 24]; however, a field study of behaviour-based authentication shows that re-authentications are a potential issue [19]. More specifically, the evaluated scheme used a (simulated) behaviour-based authentication scheme that focused on the user's touch input behaviour. Whenever re-authentication was required, the user's current task was interrupted and a re-authentication prompt with dark background, similar to the standard Android authentication prompt, appeared immediately. Non-surprisingly the unpredictability of a re-authentication and the context switch due to the task interruption were annoying to some users.

While re-authentication is unavoidable to preclude misuse of a device or an app, the unpredictability of re-authentication can be reduced by delaying the transition between the current task and the re-authentication prompt through a fade-in effect. During the fade-in, the user is allowed to continue interacting with their current task on the device. In addition to the fade-in effect, the re-authentication prompt can be configured to have varying levels of transparency to provide a visual of the user's current task in the background. The fade-in effect should reduce the unpredictability of the re-authentication and a visual of the current task of the user should reduce the context switch overhead due to re-authentications. Together these controls have the potential to provide increased usability at the cost of reduced security.

In this paper, we evaluate different configurations of explicit authentication schemes (such as PINs or pattern-locks) when used for re-authentication. Our focus is on the fade-in

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22–24, 2016, Denver, Colorado.

effect and the transparency of the re-authentication prompt. We choose behaviour-based authentication as a target use case to evaluate the different configurations; however, our findings can be generalized to other authentication proposals that require re-authentications. In addition to the re-authentication configuration used in the previous work [19], we select three configurations of explicit authentication schemes for re-authentication: (i) The authentication prompt appears immediately (no fade-in) and the background of the authentication prompt is transparent to provide a visual of the user's current task in the background; (ii) the authentication prompt appears immediately and the background of the authentication prompt gradually transitions from transparent to opaque for improved security; and (iii) the authentication prompt appears after a four second fade-in delay and the background of the authentication prompt gradually transitions from transparent to opaque.

We perform lab experiments using synthetic tasks to evaluate the security perception, ease of use, obstructiveness and annoyance of PIN and pattern-lock-based re-authentication based on the default configuration from the earlier study [19] (as a baseline) and the modified configurations. In addition to these qualitative usability metrics, we collect quantitative data on the task efficiency and the task error rate for a multifaceted evaluation of these configurations. Finally, we conduct interviews to gather participants' perceptions on the sensitivity of different kinds of apps and of participants' preferred configuration of the re-authentication prompt for different apps and different environments.

Our study was completed by 30 participants. Though our findings indicate no differences for the user performance (in terms of task efficiency, task error rate, and context switch overhead) against these configurations, participants found all three modified configurations to be less annoying and less obstructive as compared to the default configuration. The modified configurations were also at least as easy to use as the default configuration. As expected, the perceived security level of the modified configurations was quite low when compared to the default configuration. While the low perceived level of protection was a bottleneck in the adoption of the modified configurations in high-risk environments and for sensitive content, a significant number of participants preferred the proposed configurations over the default configuration for less sensitive content and for low-risk environments. We also communicate suggestions by the participants on how to improve the design of our proposed configurations and we discuss guidelines for future implementations of re-authentication schemes.

2. MOTIVATION

Implicit factors have been proposed to reduce authentication overhead on the web [2], personal computers [22] and smartphones [17, 25, 32]. Our focus is on smartphones. The implicit factors for authentication on smartphones leverage behavioural biometrics [32], device context [16, 24, 25] or the sensitivity of launched apps [17]. We next describe each of these three implicit factors and their potential need to re-authenticate a smartphone user.

2.1 Re-authentication Scenarios

Implicit authentication (IA): IA uses behavioural biometrics to conveniently authenticate users without requiring their explicit input. Various IA schemes have been proposed that authenticate users through their touch input behaviour [13, 21, 37], keystroke behaviour [8, 10, 14], gait behaviour [12, 27] or device usage behaviour [32, 33]. Several IA proposals have been shown to provide over 95% accuracy [13, 21, 37] and researchers have proposed to use them as a primary authentication mechanism for users who do not lock their device or as a secondary authentication mechanism to compliment the existing primary authentication schemes.

There are scenarios when an IA scheme is unsure about the identity of the user. This uncertainty may be caused by an adversary using the device or it could be the result of a false reject. False rejects occur when legitimate users are misclassified as adversaries. When an IA scheme is unsure about the identity of the user, it uses an explicit authentication mechanism to re-authenticate the user. Furthermore, if an IA scheme relies on the input behaviour of the user, the false rejects can occur mid-task and re-authentication requires interrupting the current task of the user [19].

Context-aware authentication: Several schemes have been proposed that leverage device context to reduce authentication overhead [16, 24, 25, 28]. These schemes rely on a variety of contextual sources, including location, proximity to WiFi and Bluetooth devices, and ambient light and noise. An evaluation of CASA [16] shows that it can reduce explicit authentications by 68% and a lab study of the scheme proposed by Riva et al. [28] indicates that it can reduce the number of explicit authentications by 42%.

Context-aware schemes can be deployed to sense and assist in authentication only when users begin their interaction with the device. However, to preclude attacks from informed attackers (such as friends and coworkers), a continuous authentication scenario is more suitable. For instance, a continuous proximity sensing scheme will not allow an informed malicious coworker to unlock the device at the workplace and then move to a secluded place to access personal data on the device. Since such scenarios may arise with the legitimate user of the device (e.g., the device owner moves out of the proximity range while using the device, or an ambient noise sensor may switch off), the device owner may be subjected to mid-task re-authentication.

App-specific authentication: Hayashi et al. [17] show that all-or-nothing access to smartphones does not align with user preferences. They find that while the majority of the users prefer to be authenticated for select apps only, for a subset of apps the users want some functionality to be available always and some functionality to be available after authentication. For instance, browsing existing entries (such as contacts) in an app should always be available while modifying or deleting entries should require authentication. Similarly, looking at recent messages should not require authentication while browsing old messages should require user authentication. These scenarios require mid-task re-authentication of the user.

2.2 Need for Better Re-authentication Schemes

User studies on IA show that users find IA to be more convenient and easier to use than traditional authentication schemes [4, 19]. Evaluations of the context-aware schemes show that the reduced authentication overhead is found to

be useful and the users indicated that they would use the evaluated scheme if it was available on their devices [16, 24]. A similar positive experience was reported for an app sensitivity based authentication scheme [17].

While users agree that these schemes are useful and are interested in adopting them, most of these evaluations have not investigated the effect of re-authentications with the exception of Khan et al. in their usability study of touch input-based IA [19]. Khan et al. find that for 35% of the participants, re-authentications due to false rejects were a source of annoyance. The participants found the re-authentications to be frustrating due to their unpredictable nature and the accompanying context-switch due to authentication interrupts. The context switch was also responsible for reducing the overall task completion time of the participants.

Since unavoidable re-authentications are a potential issue in the adoption of IA, we investigate whether the unpredictable nature and the context-switch due to authentication interrupts can be reduced by modifying how a user is re-authenticated. We assume that our concepts can mitigate these usability issues and thus reduce barriers to the adoption of novel authentication schemes that require re-authentication.

3. STUDY DESIGN & OBJECTIVES

In this section, we first outline different approaches that can be used for re-authentication. We then provide the rationale for our selection of a slightly modified version of the existing authentication prompts through two configuration parameters: *time delay* and *screen transparency*. Finally, we outline the security and usability trade-offs introduced by these parameters, our constructions of re-authentication prompts with different configurations of these parameters and the usability expectations from our constructions.

3.1 Re-authentication Approaches

Several re-authentication schemes are possible. During the design phase, we considered the following:

Split-screen configuration: In this configuration, the authentication prompt and the current user task equally share the screen space (screenshots are provided in Appendix B). This enables the user to authenticate within a timeout period with their task in sight. However, it is difficult to ensure that the authentication prompt is displayed at a location that the user is focusing on. In case the authentication prompt appears in the location where the user is focusing on, it results in the aforementioned usability issues. Nevertheless, this approach is worth exploring once gaze tracking solutions for smartphones have matured [23, 26].

Alternate authentication mechanisms: Alternate authentication mechanisms have been proposed to counter shoulder-surfing attacks, which reduce the size of the authentication prompt [20] or allow the user to enter the PIN using simple up and down gestures [35]. Similar to the split-screen configuration, a challenge for these approaches is the identification of the most suitable placement of the authentication prompt for re-authentication. Another option is to use mechanisms that provide security using obscurity. For instance, De Luca et al. [7] have proposed a mechanism that allows users to enter the secret discretely through the back of the device. In another proposal, the user is expected to enter an incorrect character to authenticate when the phone vibrates [6].

These approaches are promising; however, they may introduce confounding factors as they have not been adopted widely. The missing experience of the participants with these new configuration design may affect their usability perceptions. Since several usability issues can be traced to the unpredictability and context-switch effects of re-authentication [19], we perform experiments to investigate whether the unwanted effects stemming from unpredictability and context-switches can be minimized for widely deployed authentication mechanisms. Therefore, the main objective of this study is to investigate whether widely deployed authentication schemes can be modified to make them more usable for re-authentication scenarios without significantly compromising on security.

3.2 Configuration Parameters

We introduce two configuration parameters for existing authentication prompts: *time delay* and *screen transparency* and define the possible values for each of the parameter. The *time delay* represents the time it takes between the transition from the current task of the user to the appearance of the re-authentication prompt. This variable supports two possible values: immediate lock (Imm-Lock) and gradual lock (Grad-Lock). In the Imm-Lock case, the re-authentication prompt appears immediately (without any delay) whereas for the Grad-Lock case, the re-authentication prompt appears after a predefined interval with a fade-in effect. During this fade-in, the user can continue to interact with the current task. The two possible values provide different usability and security trade-offs: the secure Imm-Lock bars the user from interacting with the current task, while the less secure Grad-Lock is not abrupt and provides the user with an opportunity to interact with the current task during the fade-in effect thereby potentially allowing the user to reduce the effect of interruption. For example, the user can finish reading a sentence.

For our experiments, we chose a four second time delay. Our selection was based on the results from previous studies and our experiments with both shorter and longer delays. Ferreira et al.'s [11] study on understanding micro-usage patterns for various smartphone apps revealed that 40% of the application usage lasts less than 15 seconds and is sufficient for a user to read or reply to a message. In a study conducted by Yan et al. [38], they find that 50% of the smartphone interactions last fewer than 30 seconds. With such brief periods of interactions, it is therefore necessary to lock the device quickly to prevent any misuse. For the grace period, we considered and tested delays between two to seven seconds. During our empirical tests with four participants, we found that the four seconds delay period allowed the participants to prepare for re-authentication prompts. The shorter delay values did not provide the users with enough time to prepare for the re-authentication prompt, whereas the longer delay values made the users anxious in anticipation of the re-authentication prompt.

The *screen transparency* variable affects the visibility of the current task by configuring the background of the re-authentication prompt to be instantaneously dark (Imm-Dark, see Figure 1a), gradually fade from transparent to dark (Grad-

Dark, see Figure 1b) or remain transparent (Imm-Trans, see Figure 1c and 1d). Similar to the *time delay* variable, the three possible states of *screen transparency* provide varying degrees of security and usability. The Imm-Dark state is the most secure one because it hides sensitive data displayed in the current task; however, the context-switch overhead should be the most in this case since the user's task is not visible anymore. The Imm-Trans state covers the other extreme where sensitive data displayed in the current task remains visible behind the re-authentication prompt; however, the context-switch overhead should be the least since the user's task remains visible while the user is interacting with the re-authentication prompt. The Grad-Dark state provides a grace period during which the user can authenticate to resume the task at hand; however, if the user fails to do so in a configurable amount of time, the background of the re-authentication prompt becomes dark thereby hiding the user's current task.

3.3 Re-Authentication Prompt Configurations

The four configurations of re-authentication prompts that we construct using the different meaningful combinations of the two configuration parameters are as follows:

1. **Immediate Dark, Immediate Lock (Imm-Dark-Imm-Lock):** We evaluate the default lock scheme on most Android smartphones to establish a baseline for when it is used for re-authentication. In this configuration the re-authentication prompt appears immediately with a dark background, which completely hides the content of the current task, and the user can no longer interact with the current task. The re-authentication prompt asks the user to enter a PIN or pattern-lock and the user is able to access the current task again only after correctly answering the re-authentication prompt. This configuration was also used in the earlier work by Khan et al. [19], as discussed in § 2.2.
2. **Immediate Transparent, Immediate Lock (Imm-Trans-Imm-Lock):** The re-authentication prompt appears immediately in this configuration and the user can no longer interact with the current task. However, the background of the re-authentication prompt remains transparent, which allows users to observe the contents of their task.
3. **Gradual Dark, Immediate Lock (Grad-Dark-Imm-Lock):** In this configuration, the re-authentication prompt appears immediately and the user can no longer interact with the current task. Furthermore, the background of the re-authentication prompt is initially transparent and the contents of the current task are visible. Then, the background of the re-authentication prompt gradually fades into a dark screen and hides the contents of the current task from the user. If the user manages to authenticate before the screen has darkened completely, this configuration keeps the user's current task visible in the background.
4. **Gradual Dark, Gradual Lock (Grad-Dark-Grad-Lock):** In terms of task visibility, this configuration is similar to the Grad-Dark-Imm-Lock configuration described above. That is, the background of the re-authentication prompt is initially transparent and then turns into dark. However, this configuration also allows the user to continue interacting with the current task for a grace-period of four seconds before the re-authentication prompt appears. During the grace period, the brightness of the current task is reduced to indicate the forthcoming re-authentication prompt to the user. After the re-authentication prompt appears, the users can no longer interact with their task.

3.4 Study Aims

We expect the following properties from our re-authentication prompt configurations:

- Imm-Dark-Imm-Lock is the most obstructive therefore it should be the most annoying. Furthermore, since it provides no visual clues on the current task of the user, task efficiency should be reduced.
- Imm-Trans-Imm-Lock also immediately locks out the user but its presentation of the re-authentication prompt is less intrusive and it provides visual clues on the current task of the user. Therefore, it should be less annoying and more task efficient as compared to the Imm-Dark-Imm-Lock configuration.
- Grad-Dark-Imm-Lock has similar properties as Imm-Trans-Imm-Lock but it provides additional security by making the current task of the user invisible after a predefined time interval. Therefore, it should score similar to Imm-Trans-Imm-Lock in terms of usability with a relatively better security perception.
- Grad-Dark-Grad-Lock enables the user to interact with the current task for a grace period and this may increase the task efficiency of the users. However, the user may not take advantage of the grace period and instead wait for the re-authentication prompt to appear, which may increase the anxiety and annoyance of the user.

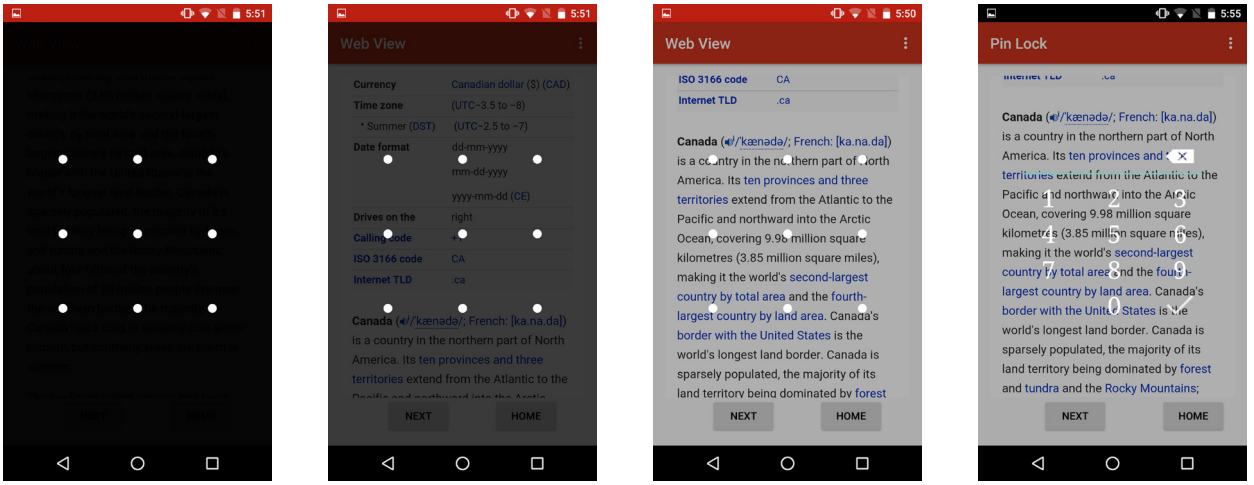
In the rest of this paper, we evaluate whether the four re-authentication prompt configurations provide the aforementioned usability properties.

4. STUDY DESIGN

In this section we outline our design of a user study to evaluate the four re-authentication prompt configurations. To measure the properties of each configuration, we perform a lab-based evaluation where participants are invited to experience each configuration by performing predefined synthetic tasks. After the users experience these configurations, they are asked to rate and provide qualitative feedback in terms of usability, security perception and their willingness to use these configurations. In addition to the user feedback, we measure the task efficiency, context switch overhead, and task error rate against each configuration. Our evaluation and feedback setup are designed to elicit the efficacy of these configurations for re-authentication in different scenarios. Our study was reviewed and received approval from the IRB of our university. We now provide details of our study design in terms of experimental setup and our methodology.

4.1 Apparatus

While several use cases exist for re-authentication (see § 2.1), we choose IA as the representative use case in this work because it was easier to explain and conduct than the other re-authentication cases outlined in the paper. Our choice of IA



(a) Imm-Dark (b) Grad-Dark (c) Imm-Trans (d) Imm-Trans

Figure 1: The proposed configurations with varying values for *screen transparency*. Figures (a), (b) and (c) show the three possible values when a pattern-lock based re-authentication prompt is used. Figure (d) shows a sample value for a PIN-based re-authentication prompt. For the Grad-Dark configuration, the background of the re-authentication prompt gradually turns from transparent into dark.

is also motivated by the prior work of Khan et al. [19] in the IA domain that highlights the issues with re-authentications in case of false rejects. To ensure that each participant experiences a certain number of false rejects, we use a simulated IA scheme, as was also done by Khan et al. In particular, our scheme simulates IA schemes based on a user's touch input or keystroke behaviour.

For our experiments, we select two widely used authentication mechanisms on Android: a 4-digit PIN and the Android pattern-lock (with the same constraints on possible patterns as in Android). The user interface of both schemes was similar to the Android lock screens (see Figure 1).

The four re-authentication prompt configurations introduced in § 3.3 are evaluated using two synthetic activities — a text entry activity and an email activity (screenshots are provided in Appendix A). We choose these activities since they represent common smartphone activities (i.e., reading and composing emails and text messages or interacting with social media apps).

- **Text entry activity:** This activity displays a 12-digit number to the participants. It also contains a text box and the users are asked to enter the displayed number in the text box using the numeric keyboard of the device.
- **Email activity:** In the email activity, users are asked to read an email in an email app. The user interface for the email app developed for this activity looks similar to the Android Gmail app. Once a participant has read the email, they are asked to answer a multiple choice question related to the email on a laptop. The emails composed for this activity contained sensitive data, which emphasized the need to protect the emails from adversaries (see Figure 10b for an example).

The design of the text entry activity ensures that the interaction of the users with the app can be measured, which enables us to compute several metrics in terms of context-switch overhead and errors made by the users. For the email activity, since the emails contain sensitive material, the users performing the email activity should consider the security implications of a re-authentication prompt configuration in addition to its usability aspects.

These activities were bundled in two separate Android apps, which allowed users to perform tasks. We define a task as completing the text entry or the email activity along with a mid-task re-authentication of the user using either the PIN or the pattern-lock in one of the four configurations. For the text entry task, the users were interrupted at predefined intervals, which were triggered based on the key presses by the users. The number of key presses required to trigger re-authentication changed across different text entry activities for each user but it stayed constant across users for those tasks for results to be comparable. Similar to the text entry task, the users were interrupted with a re-authentication prompt after a predefined number of swipes for the email task. The apps were instrumented to gather the timestamps of events, including input events by the user and the display and dismissal events of the re-authentication prompts. The apps also collected the errors made by the users for the text entry activity and during the re-authentication. We also logged the user interactions, including the keystrokes and screen touch events, during the grace period for the Grad-Dark-Grad-Lock configuration. The data collected by the apps was instrumental in computing the task completion rate, context switch overhead and the error rate against each re-authentication prompt configuration.

4.2 Evaluation Methodology

We evaluate the four re-authentication prompt configurations using the text entry and email tasks. Each scheme was evaluated in a round that consisted of four text entry tasks and two email tasks. Each user was subjected to

five rounds and in each round a different re-authentication prompt configuration was evaluated. For the first round, the participants performed the tasks without any authentication, which allowed us to establish a baseline. The participants were allowed to take a break between each task and each round. The order of the four re-authentication prompt configurations was randomly chosen for the participants.

The participants shortlisted for this study were invited for an hour long lab-based study. The participants were first asked to fill a demographic survey, which asked about their age, gender, and current occupation. They were then asked to fill a security preferences survey. In terms of security preferences, we asked the participants about their device locking habits, their preferred authentication scheme, and the adversaries that they wanted protection against. These pre-study surveys are provided in Appendix D. After the pre-study surveys, the participants were introduced to IA, the possibility of false rejects in IA, the tasks and apps used during the study, and the different re-authentication prompt configurations. The participants were also told that false rejects were simulated for the purpose of this study. We gave participants the option to select their preferred lock scheme (PIN or pattern-lock) and a corresponding secret for the study. We did not assign participants a specific scheme to avoid any bias due to their inexperience with it. This design decision prohibited us to counterbalance the authentication methods. The authentication times varied across participants. To cater for this, we report within-subject relative differences instead of absolute values. The participants experienced the different configurations in multiple rounds. After the completion of each round, they were asked to rate the usability and perceived security of the configuration that they experienced and to give an overall ranking in terms of their preferences by taking both the usability and the security of the evaluated configuration in account. Participants were also asked to indicate their preferences for the evaluated configurations under different device usage scenarios and were subjected to a semi-structured interview to gain further insight into their feedback. A researcher was present to respond to any questions the participants had.

4.3 User Feedback

The evaluated schemes trade off security for usability and since different users have different security preferences for different apps and different scenarios, we seek feedback from the users against four apps for three different scenarios. Previous studies have shown that users prefer a strict security setting for financial and email apps, which contain highly sensitive data, whereas they prefer a relatively relaxed security setting for contacts and other utility apps [17]. We sought feedback from the users for four apps: a banking app, an email app, a photos app, and a contacts app. These apps are commonly used and contain varying levels of sensitive data of the smartphone user. The participants were asked to consider the following device usage scenarios with the aforementioned apps available on the device.

- **Bus Scenario:** The participants had to consider a situation where they are traveling on a bus and they accidentally leave their smartphone behind. A stranger picks up their device and starts using it.
- **Office Scenario:** This scenario asks the participants to consider a work environment where one of their colleagues starts using their device when it is left unattended. For this scenario, the apps on the device may be used for a limited time by someone known by the smartphone owner.
- **Home Scenario:** In this scenario, we asked the participants to consider that their spouse accesses their device while it is left unattended or when they are asleep. The number of adversaries is limited in this scenario as compared to the others and the users may or may not want to protect their data from their spouse.

A researcher presented the scenarios to the participants and was available during the interview to answer any questions participants may have. Participants were given sufficient time to consider the presented scenarios. For each scenario, the participants were told that the re-authentication prompt would get activated in case the system notices any suspicious activity. We also reminded them of false rejects and the fact that they may be subjected to re-authentication while they are using the device. In order to inquire about the security perception of an evaluated re-authentication prompt configuration, the participants were told that for the purpose of these scenarios, they should consider that only IA is protecting their device. Since different users may have different security preferences for each configuration and each usage scenario, we initially asked the users to establish the sensitive nature of the apps and usage scenarios. Then the participants were asked to provide feedback in terms of security perception, usability and preferred re-authentication prompt configuration for each of the four apps under each of the three device usage scenarios. The feedback questionnaire is provided in Appendix E.

Finally, at the end of the study, we conducted a short semi-structured interview (provided in Appendix F) to gain insight into participants' overall impression of the configurations that they evaluated.

5. RESULTS

The data collected through the user studies and the interviews were recorded and analyzed. The audio responses of the participants were transcribed by one of the researchers. We report both the quantitative and the qualitative results from the study in this section. For statistical significance, we used paired t-tests when comparing continuous data for the within-subjects condition such as the inter-stroke rate for each user between grace and non-grace periods. We used one-way ANOVA when comparing continuous data for the within-subjects condition for the four authentication configurations (e.g., context-switch overhead). We used chi-squared tests when comparing participants' responses to categorical Likert-type questions.

5.1 Study Participants

We advertised the study through our university-wide mailing list and through the graduate student research portal of our university. The study was advertised with the title "Evaluating authentication schemes for smartphones" and we recruited only those users who had prior experience with using smartphones. Participants received $10 for their participation for an hour of study.

We recruited 30 participants for the study (see Table 1 for their demographics). All the participants were students from

N=30		
Gender	60%	Females
	40%	Males
Age	33%	Under 20 years
	57%	21-25 years
	7%	26-30 years
	3%	31-35 years
Lock device?	26 (87%)	Yes
	4 (13%)	No
Authentication scheme	13/26	Pattern-lock
	5/26	PIN (4 digits)
	6/26	Fingerprint
	2/26	Password
Protecting from?	25/26	Strangers
	16/26	Friends
	14/26	Room-mate
	14/26	Coworker
	3/26	Spouse, own children

Table 1: Demographic information and the device lock usage pattern of the participants.

our university. The majority of our participants (87%) reported that they locked their device. The security preferences of participants who locked their devices are provided in Table 1. We asked the four participants who did not lock their devices for their reason to do so: two indicated that they had nothing to protect, two wanted their emergency contacts to be available and one considered authentication to be inconvenient (multiple answers were possible).

5.2 Quantitative Results

Out of 30 participants, 18 participants chose to use a pattern-lock during the study, while the remaining participants chose to use a PIN. Participants were subjected to five rounds in total. During the first round, participants were not interrupted for re-authentication. This round was used to establish a baseline and we use the term BASE_ROUND to refer to it. For the remaining rounds, participants tested one of the four configurations in each round. The order of the configurations was random during the four rounds.

During each round, participants completed four text entry tasks and two email tasks. They re-authenticated once for every email and text entry task during all rounds except BASE_ROUND. The high rate of re-authentication is not representative of a real-world scenario; however, our motivation was to get participants acquainted with the configurations and to collect sufficient data to evaluate the metrics used in this section. During the study each participant re-authenticated themselves 16 times during the text entry activity (four times per configuration) and eight times during the email activity (twice per configuration). In total, 120 re-authentication events, 120 text entry tasks and 60 email tasks were logged per configuration by our apps.

5.2.1 *Effect on task completion overhead*

The task completion time is the time taken by the users to complete a text entry or an email task. It also includes the time taken by the users to re-authenticate themselves while evaluating one of the configurations. The task completion overhead is the additional time taken to complete a text entry task as compared to the BASE_ROUND in which a user is not interrupted to re-authenticate. For the task completion overhead, we only take into account the text entry activity since the emails used for the email activity were of a different nature and length during each round. Our goal is to find if there are any re-authentication prompt configurations that assist the users in completing their text entry tasks faster.

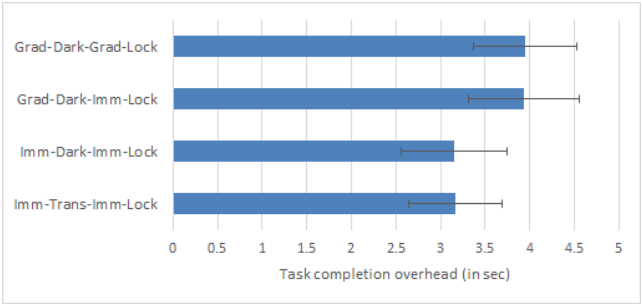


Figure 2: Task completion overhead time for the text entry activity relative to the BASE_ROUND (error bars represent 95% confidence interval).

We found that on average users took 3-4 seconds longer when they had to re-authenticate during a text entry task (see Figure 2). A one-way between subjects ANOVA was conducted to compare the effect of the four configurations on the task completion overhead, which indicated no significant differences across the four configurations *(F(3,116)=2.31, p=0.08)*.

Discussion: Our expectation that the Imm-Dark-Imm-Lock configuration is less efficient as compared to the modified re-authentication prompt configurations turns out to be incorrect. Though, we did not find any significant differences in the performance of the configurations, the participants mentioned during the study that they felt that their performance was affected during the Imm-Dark-Imm-Lock configuration:

> "It kind of freaks me out because it is too sudden, it slows down whatever I was doing." (P4)

5.2.2 *Effect on context switch overhead*

Context switch overhead for the text entry task is defined as the time taken by the users to resume their text entry task once they have re-authenticated. The context switch overhead is represented by the time interval between the dismissal of the re-authentication prompt and the first key press on the text entry task once the re-authentication prompt has disappeared. It was not possible to compute this metric for the email task because after re-authenticating a user would complete reading the email text visible on the screen before interacting with the device. Our expectation was that a visual of the user task in the background would reduce the context switch overhead. To confirm this, we conducted a one-way between subjects ANOVA to compare the effect of the four configurations on the context switch overhead. However, the results indicate no significant differences across the four configurations *(F(3,116)=1.15, p=0.33)*.

Discussion: While no statistically significant differences were observed, during the interviews, most users found the Imm-Dark-Imm-Lock configuration to be abrupt and reported that it was difficult to resume their task after re-authentication:

> "I lost my place [context] on what I was doing before [the lock appeared], so it is my least favourite. It would be too frustrating for me for everyday use, so I would rather take the risk." (P9)

> "You can't prepare for what's going to come. It takes more time to pick up after unlock" (P10)

5.2.3 *Effect of grace period*

We allowed a grace period of four seconds for the Grad-Dark-Grad-Lock configuration. During the grace period the participants could continue working on their task for four seconds before getting locked out. We observe that all participants took advantage of this grace period by continuing their work during the text entry activity. The average task completion time for the Grad-Dark-Grad-Lock configuration was 13 seconds and we found that on average users entered 38% of the text during the four second grace period with some users entering up to 60% of the total text in the grace period. A similar trend was observed for the email task where 23% of the swipe events occurred during this period (average time to complete the email task for the Grad-Dark-Grad-Lock configuration was 41 seconds).

We find that the inter-key intervals (time interval between two consecutive key presses) of the users reduced significantly for the Grad-Dark-Grad-Lock configuration during the grace period. The average inter-key interval of users reduced by almost 60% during the grace period when compared to the average inter-key interval during the task (see Figure 3). A paired t-test was conducted to compare the inter-key interval between the grace and non-grace period for the same text entry activity for each user. The results show that inter-key intervals are significantly different between the grace and non-grace period *($t(29) = 2.1$, $p = 0.04$)*.

Discussion: Our results indicate that participants took advantage of the grace period by attempting to quickly complete the text entry activity. They typed faster than their normal speeds during the grace period.

5.2.4 *Effect on task error rate*

In case the input of the users mismatched the displayed text for the text entry task, we counted it as an error (with at most one error per task). Our results indicate that users made errors in 77 out of 600 text entry tasks. However, a one-way between subjects ANOVA for the task error rate across the four configurations and BASE_ROUND indicates no significant differences *($F(4,145)=1.51$, $p=0.2$)*. Similarly, while participants made errors in 43 out of 240 email tasks, the differences were not significant across the different configurations *($F(4,28)=0.28$, $p=0.84$)*.

Discussion: The task error rate among the configurations were comparable. Though the inter-key interval of the users during the grace period reduced significantly, it did not affect the task error rate compared to the other authentication configurations.

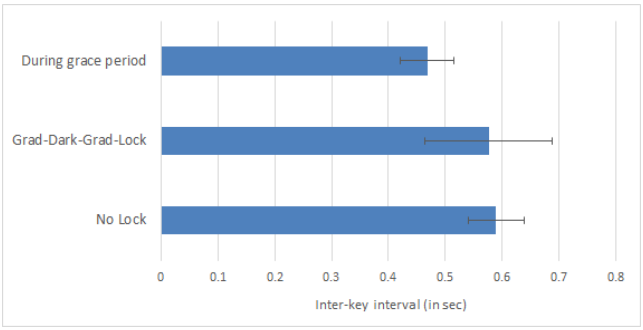


Figure 3: Inter-key interval for the text entry activity (error bars represent 95% confidence interval). The top bar represents the inter-key interval for the Grad-Dark-Grad-Lock configuration during the grace period.

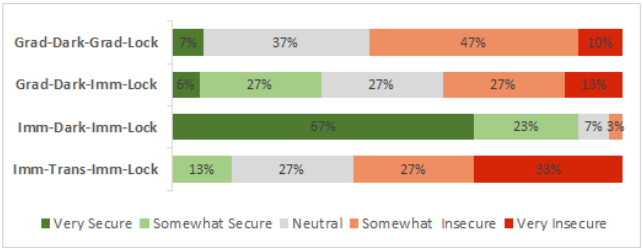


Figure 4: User perceptions of the security of the four re-authentication prompt configurations.

5.3 Qualitative Feedback

For the apps evaluated in this work, 100%, 73%, 60% and 30% of the participants considered the banking, email, photo and contacts app to be sensitive, respectively. The responses to the pre-study question regarding the adversaries that the participants (who used protection) wanted protection against indicate that different scenarios require different levels of protection. Almost all users wanted protection against strangers, which corresponds to the bus scenario. Corresponding to the office scenario, 54% of participants wanted protection against co-workers. On the other hand only 11% of participants considered that they needed protection against family members, which corresponds to the home scenario.

We now present the findings from the feedback of the participants regarding the usability and security perceptions of the configurations for each app in the different usage scenarios.

5.3.1 *Security perceptions*

Figure 4 shows the security perceptions of the participants for each re-authentication configuration. Significantly more (57% more) participants thought that the Imm-Dark-Imm-Lock configuration was more secure than the other configuration ($\chi^2(3) = 151, p < 0.001$). Imm-Dark-Imm-Lock immediately hides the content on the screen to prevent the leakage of any sensitive information. Some participants indicated that they would take advantage of this increased security at the cost of usability for some apps:

> "If I am sending an important email, I do not want anybody else to look at it even for a second. It is annoying but it would be the most beneficial." (P13)

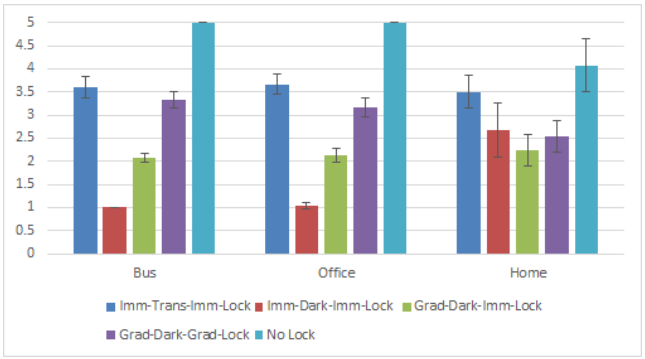


Figure 5: User preference of the configurations for the banking app in different scenarios. 1 represents the most preferred configuration while 5 represents the least preferred configuration (error bars represent 95% confidence interval).

This was followed by the Grad-Dark-Imm-Lock configuration, which was considered to be secure by 33% of the participants. We found that only 13% and 7% of the participants considered the Imm-Trans-Imm-Lock and Grad-Dark-Grad-Lock configurations to be secure. As expected, the visible task in the background is perceived negatively by most users in terms of security. The Grad-Dark-Grad-Lock configuration provides access to the device for a short period of time and participants felt that their content was vulnerable during this period. We now explore whether the configurations that were perceived to be less secure were considered appropriate for some usage scenarios.

> "I liked the idea that how the lock appears at the start [during Grad-Dark-Imm-Lock], so if it is someone else, they can't enter any text message and they can't send anything compared to the last scheme [Grad-Dark-Grad-Lock] where they can do anything if they are fast enough" (P4)

The Imm-Dark-Imm-Lock configuration was perceived most secure and all participants indicated that they would only consider using this configuration for their banking app on a bus and at the office (see Figure 5). On the other hand, for the home scenario, users had different preferences. 40% of the users indicated that they would still only consider using the Imm-Dark-Imm-Lock configuration for the banking app at home while 23% of the users indicated that they would prefer using the Grad-Dark-Imm-Lock configuration instead. Some of the user comments shed more light on the user preferences for the banking app:

> "Banking would be very sensitive, so I want it to get dark as quickly as possible." (P9)
>
> "Even with my partner, I won't feel completely secure with my banking app opened on my phone that is why I would prefer immediate dark." (P4)

The feedback from the users was inconclusive for the email app and there is no one configuration that users significantly prefer over the other for the different usage scenarios. On the other hand, for the photos app, the majority of

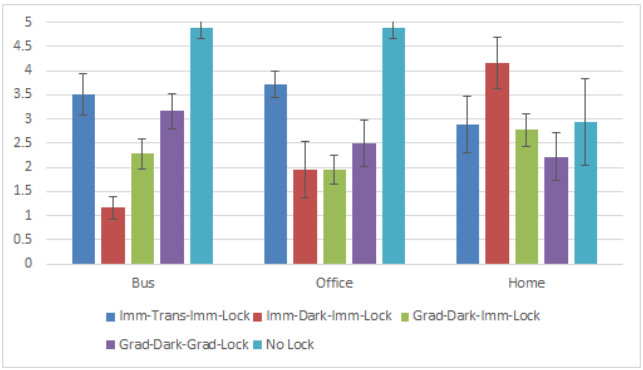


Figure 6: User preferences for the configurations for the photos app in different scenarios. Only users who consider the photos app as sensitive are included (N=18). 1 represents the most preferred configuration while 5 represents the least preferred configuration (error bars represent 95% confidence interval).

the participants who considered the photos app to be sensitive preferred the Imm-Dark-Imm-Lock configuration for the bus scenario (Figure 6). For the office scenario, the participants who were very concerned about protecting their photos preferred configurations that obscured or gradually obscured the app, preventing it from being accessed by their co-workers:

> "I won't care about my photos with respect to a stranger but in office where its more professional environment with the people I know, I would increase the security of the scheme." (P12)
>
> "I have a lot of photos that are very personal and I don't want them [strangers] to see any part of them." (P6)
>
> "I might have already shared a lot of photos with my partner, so I would prefer a comfortable lock scheme." (P6)

For the contacts app, the participants were willing to use configurations that provided device access for a period before locking them out. They wanted it so because this would allow a stranger to contact them in case they lost their device. The participants were less concerned about securing their contacts at home or office because they felt that they shared contacts with individuals at these locations.

> "If someone picked up my phone and they are looking at my contacts, they could try to return it to me through someone in my contacts, so I would choose something except the one that turns dark immediately." (P7)
>
> "For contacts, now there is an issue of privacy because these are people which they [office colleagues] might also know, so it is important that I protect their information but at the same time I don't want it to be very inconvenient for me when I look at the contacts." (P2)

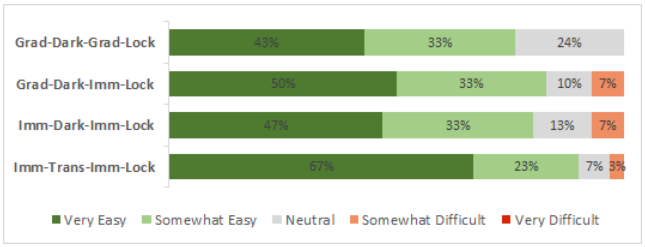


Figure 7: User perceptions on how easy it was to use the evaluated configurations.

The configuration preferences in terms of the percentage of users willing or not willing to use a particular configuration for various scenarios are presented in Appendix C.

Discussion: The participants considered the Imm-Dark-Imm-Lock configuration to be the most secure out of all four configurations. The inclination of the users while selecting the configurations are location- and app-based. While they prefer the Imm-Dark-Imm-Lock configuration to protect their banking information, they prefer to protect access to the photos app only at unknown locations. Users feel comfortable while browsing their device at home, and care less about using a more secure configuration except for the banking app.

5.3.2 Usability perceptions

Our main goal while designing these configurations was to reduce the usability issues with re-authentication reported by Khan et al. [19]. To this end, our configurations provided the users a visual of their tasks or a grace period to continue their work without disruption. We now present the perceived usability of these configurations.

We asked the users to rate the configurations in terms of ease of use. Figure 7 summarizes the responses of the users. We found that all configurations received a high rating in terms of ease of use and there were no statistically significant differences among the four configurations. In addition to a positive reception of the fade-in effect in Grad-Dark-Grad-Lock, users utilized the grace period to input data. Some of the users' comments include:

> "It helps you to continue typing and get your thoughts out. It didn't allow you to access the app though [after sometime] so it is a good balance between usability and security." (P16)

> "If I was in a rush to send an email to a client or my boss, I wouldn't want it to immediately get dark, I would want that buffer time to carry on my thoughts." (P4)

We also asked users how obstructive and annoying they thought each configuration was. Their responses (see Figure 8) indicate that significantly more participants considered the Imm-Dark-Imm-Lock configuration as more obstructive ($\chi^2(3) = 96, p < 0.01$). Similarly, Figure 8 shows that significantly more participants considered the Imm-Dark-Imm-Lock configuration was more annoying ($\chi^2(3) = 71, p < 0.01$). In terms of obstructiveness, 70% of the

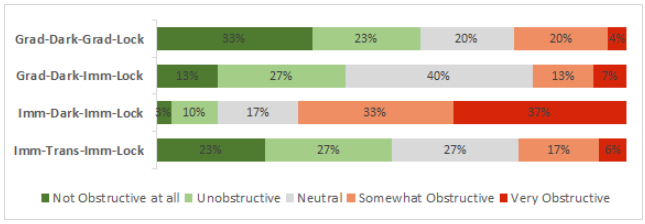


Figure 8: User perceptions regarding obstructiveness of the configurations.

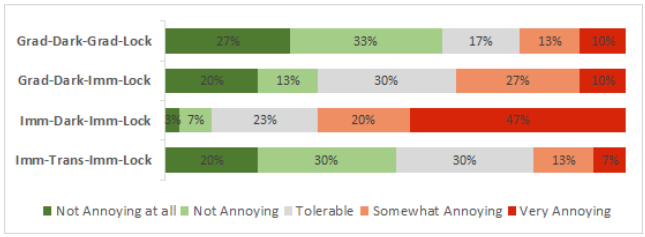


Figure 9: User perceptions regarding annoyance of the configurations.

participants rated the Imm-Dark-Imm-Lock configuration as somewhat or very obstructive and 67% of the participants rated it as somewhat or very annoying (Figure 9). This explains why Imm-Dark-Imm-Lock was the least preferred configuration for the users for email (47%), photos (52%) and contacts (47%) apps for the home scenario. On the other hand, users positively perceived the gradual fading of the screen transparency and the delay of the authentication prompt. User comments that reflect these findings are:

> "I lost my place what I was doing before [the lock appeared], so it is my least favorite. It would be too frustrating for me for everyday use, so I would rather take the risk." (P9)

> "I found it [Imm-Dark] very annoying because it was really an abrupt interruption to me, others were not abrupt." (P8)

> "When you were explaining to me, I thought it would be difficult to wait for the lock but I guess it was nice to not lock right away, so you can continue what you are doing and wait for it to come up." (P12)

Discussion: While the Imm-Dark-Imm-Lock was considered most secure and was preferred for sensitive apps and risky scenarios, it annoyed the users. On the other hand, the less secure configurations were perceived to be more usable and users preferred those for less sensitive apps and for medium- and low-risk scenarios.

5.3.3 Overall Perceptions

We found no significant difference when users were asked to rank the four configurations in the order of their preference while considering both the security and the usability of the configurations. Our results suggest that the users generally find it hard to select a particular configuration as their most preferred configuration and their choices are influenced largely by their perceived levels of the sensitivity of

the apps they are using and their perceived security of the surrounding environment.

6. DISCUSSION

In this section we discuss our findings from the semi-structured interviews and suggest future directions.

Annoyance due to the fade-in effect: While the majority of users responded positively to the modified re-authentication prompt configurations, six participants found the fade-in effect to be annoying. During the interviews, these participants indicated that the cause of this annoyance was the wait for the authentication prompt to appear:

> "I would rather deal with the lock as quickly as I can so I can get back to using the phone." (P9)

One participant suggested that the source of annoyance was its resemblance to the interruption on the web for subscription-based content:

> "I don't like it at all because it reminded of those websites, where you are scrolling and it stops letting you read the content and that kind of is obstructed and annoying." (P7)

We now outline the alternates that were suggested by these and other participants.

Participants' suggestion on how to re-authenticate: We sought suggestions from the participants during the semi-structured interview on how the re-authentication should be performed or improved. They proposed displaying a small timer at the top of the screen to indicate the time left before the users would be re-authenticated. Their comments were:

> "Maybe it can prompt you to type out a pattern on your phone without the visual obstruction, maybe like a small notification. It will warn you that it is going to lock and you can dismiss it by providing the secret." (P9)
>
> "Maybe instead of gradual fading, you can have a small timer up there on the screen near the status bar so that I should be expecting to get a lock screen." (P15)

Other comments regarding the design and display of the re-authentication prompt suggest that the delay before the appearance of the re-authentication prompt and the colour of the screen during the fade-in effect should be customizable.

Future design implications: Participants' responses show that the evaluated configurations are more usable albeit less secure than the Imm-Dark-Imm-Lock configuration. More specifically, in terms of participants' ratings, Section 5 showed that the Imm-Dark-Imm-Lock configuration favored security at the cost of usability whereas, all other configurations favored usability at the cost of security. Participants' feedback suggests that no particular configuration provides an optimum trade-off between usability and perceived security for re-authentication across all scenarios.

Furthermore, while most participants of our study had similar security preferences in terms of the three scenarios evaluated in this study, there was disagreement regarding the security preferences for the four apps. Therefore, re-authentication schemes need to provide users with a control to define these security preferences. A comment by a participant demonstrates the need for this:

> "You can have three different levels of security [depending on security preferences] and group your apps into those levels depending on the security you want for each app." (P9)

Similar to the findings of research efforts on primary authentication schemes, our findings indicate that future experiments on user re-authentication should leverage app sensitivity and location information to ease the re-authentication burden. For instance, an enterprise email client can use a more usable configuration to re-authenticate when the user is within the office building. Similarly, a banking app, which is providing additional security through an app-level IA mechanism [18], should use the Imm-Dark-Imm-Lock configuration.

7. LIMITATIONS OF THE STUDY

Similar to other human subject experiments, our participants were limited to those willing to participate. The feedback given by the participants was subjective in nature and therefore represents only the results of a limited sample of the population. Each participant had a different perception of the security level of the apps and the scenarios presented to them. For instance, for all apps (except for the banking app), the same app was rated by some participants as 'very sensitive' and by others as 'not sensitive at all'.

Another limitation is the smaller portion of participants (13%) who did not use any authentication mechanism on their smartphones. The usability and security perceptions of the configurations may have been different if more users perceived primary authentication schemes as inconvenient. Since participation in the study was voluntary, we had little control over preventing this disparity. Furthermore, the majority of our participants were students which may limit the generalization potential of our results. For instance, working professionals may have more sensitive data on their devices and they may have different security preferences.

For re-authentication purposes, the authentication prompt was presented to the participants in the center of the screen. This placement may have negatively affected the context switch overhead. An evaluation of other placement options, including a split screen configuration where the authentication prompt shares the screen with the user activity (see § 3.1) is a potential area of study in the future. We did not counterbalance the order of the configurations across the participants, which may have introduced bias.

A lab-based evaluation was performed because it was sufficient to achieve our objectives. However, we acknowledge that our participants were not subjected to real attacks and only considered hypothetical scenarios to evaluate the configurations. While performing experiments on the user device would have reduced issues due to user's unfamiliarity with the device and may have emphasized the need to protect their sensitive data, for the purpose of this study, we used synthetic tasks on a Nexus 5 device that was provided by the researchers. The synthetic tasks were used to take

measurements and a researcher provided device was used to avoid bias due to different screen sizes and type of devices.

8. RELATED WORK

Researchers have extensively investigated the usability issues with primary authentication schemes [5, 15, 34, 36] and have shown that these issues prevent users from using these schemes [9, 15]. Our research focus is to investigate different configurations of a subset of these schemes (PIN and pattern-lock) for re-authentication purposes and not to address previously uncovered usability issues issues (e.g., time consuming, considered unnecessary for some cases [15]) with these schemes.

To mitigate the usability issues, several research proposals have been put forth that reduce the authentication overhead of the users by leveraging user behaviour [21, 32, 37], device context [16, 24, 25] or the sensitivity of launched apps [17]. We provide a brief overview of these schemes in § 2.1. During the usability evaluation of a behaviour-based scheme, Khan et al. [19] observe the usability issues arising from re-authentications due to false rejects. They also list some suggestions by their participants on how the negative usability effects of re-authentications can be mitigated. One suggestion was to not interrupt the user and instead send an email alert or take a picture of the perpetrator. Another, more secure suggestion that inspired this work was to authenticate the user in a smaller portion of the screen in parallel and to offer the user a grace period before the device locks out.

Another line of research has focused on addressing the usability issues with existing primary authentication schemes by proposing alternate mechanisms, including gesture-based authentication [1, 7, 31] or graphical passwords [20, 30]. Users have reported positive experiences during preliminary evaluations of these schemes [1, 30]. We considered using different configurations of these schemes for re-authentication in our study; however, the usability perceptions of the participants would have been biased due to their missing experience with these schemes. Instead, participants evaluated different configurations of an authentication scheme that they are already familiar with in our study.

Another related work is SnapApp [3], which is a primary authentication mechanisms that provides a trade-off between security and usability. It presents a user with two unlock methods on the device screen — a PIN for secure access to all the device and a simple slide gesture for fast yet temporary access (30 seconds or less) to the device. Similar to our work, SnapApp favors usability at the cost of security; however, it is not a re-authentication scheme. To the best of our knowledge, our paper performs the first ever evaluation of modified primary authentication schemes for re-authentication scenarios.

9. CONCLUSION

We have proposed two modifications to the default authentication prompts of two primary authentication schemes (PIN and pass-lock) to make them more suitable for re-authentication scenarios: a transparent authentication prompt and a time delay before the authentication prompt appears. In terms of task performance, the proposed configurations perform as well as the default configuration however, the proposed configurations were perceived to be more convenient and less annoying by the users. We observe that user preferences of the configurations are largely context-based and there is no particular configuration that users want to use at all times. In terms of preference, while users want to use the default configuration (which obscures the app content) for highly sensitive apps, their choices for medium and less sensitive apps are influenced by their perception of the security of the surrounding environment and users preferred the proposed configurations for most of the less risky scenarios.

In terms of future work, a field study needs to be performed to understand the real-world performance of these configurations. Furthermore, since smartphone users who do not configure authentication on their devices are potential users of novel authentication strategies (such as IA), an evaluation study needs to be performed with such participants. Finally, our experiment suggests the need to design new re-authentication strategies that satisfy the unique usability and security requirements of re-authentication.

10. ACKNOWLEDGMENTS

Thanks to our shepherd, E. von Zezschwitz, and the anonymous reviewers for their valuable comments. We also thank Google, NSERC and the Ontario Research Fund for their support.

11. REFERENCES

[1] M. T. I. Aumi and S. Kratz. Airauth: evaluating in-air hand gestures for authentication. In *16th International Conference on Human-computer Interaction with Mobile Devices & Services*. ACM, 2014.

[2] J. Bonneau, E. W. Felten, P. Mittal, and A. Narayanan. Privacy concerns of implicit secondary factors for web authentication. In *SOUPS Workshop on "Who are you"*. ACM, 2014.

[3] D. Buschek, F. Hartmann, E. von Zezschwitz, A. De Luca, and F. Alt. Snapapp: Reducing authentication overhead with a time-constrained fast unlock option. In *CHI Conference on Human Factors in Computing Systems*. ACM, 2016.

[4] H. Crawford and K. Renaud. Understanding user perceptions of transparent authentication on a mobile device. *Journal of Trust Management*, 1(1), 2014.

[5] A. De Luca, A. Hang, E. von Zezschwitz, and H. Hussmann. I feel like i'm taking selfies all day!: Towards understanding biometric authentication on smartphones. In *33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015.

[6] A. De Luca, E. Von Zezschwitz, and H. Hußmann. Vibrapass: secure authentication based on shared lies. In *SIGCHI conference on Human factors in computing systems*. ACM, 2009.

[7] A. De Luca, E. Von Zezschwitz, N. D. H. Nguyen, M.-E. Maurer, E. Rubegni, M. P. Scipioni, and M. Langheinrich. Back-of-device authentication on smartphones. In *SIGCHI conference on Human factors in computing systems*. ACM, 2013.

[8] B. Draffin, J. Zhu, and J. Zhang. Keysens: Passive user authentication through micro-behavior modeling of soft keyboard interaction. In *Mobile Computing, Applications, and Services*. Springer, 2013.

[9] S. Egelman, S. Jain, R. S. Portnoff, K. Liao, S. Consolvo, and D. Wagner. Are you ready to lock? In *ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2014.

[10] T. Feng, X. Zhao, B. Carbunar, and W. Shi. Continuous mobile authentication using virtual key typing biometrics. In *12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, 2013.

[11] D. Ferreira, J. Goncalves, V. Kostakos, L. Barkhuus, and A. K. Dey. Contextual experience sampling of mobile application micro-usage. In *16th International Conference on Human-computer Interaction with Mobile Devices & Services*. ACM, 2014.

[12] J. Frank, S. Mannor, and D. Precup. Activity and gait recognition with time-delay embeddings. In *AAAI*. Citeseer, 2010.

[13] M. Frank, R. Biedert, E.-D. Ma, I. Martinovic, and D. Song. Touchalytics: On the applicability of touchscreen input as a behavioral biometric for continuous authentication. *Information Forensics and Security, IEEE Transactions*, 8(1), 2013.

[14] C. Giuffrida, K. Majdanik, M. Conti, and H. Bos. I sensed it was you: authenticating mobile users with sensor-enhanced keystroke dynamics. In *Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 2014.

[15] M. Harbach, E. von Zezschwitz, A. Fichtner, A. De Luca, and M. Smith. It's a hard lock life: A field study of smartphone (un) locking behavior and risk perception. In *Symposium On Usable Privacy and Security*. ACM, 2014.

[16] E. Hayashi, S. Das, S. Amini, J. Hong, and I. Oakley. Casa: context-aware scalable authentication. In *Symposium on Usable Privacy and Security*. ACM, 2013.

[17] E. Hayashi, O. Riva, K. Strauss, A. Brush, and S. Schechter. Goldilocks and the two mobile devices: going beyond all-or-nothing access to a device's applications. In *Symposium on Usable Privacy and Security*. ACM, 2012.

[18] H. Khan and U. Hengartner. Towards application-centric implicit authentication on smartphones. In *15th Workshop on Mobile Computing Systems and Applications*. ACM, 2014.

[19] H. Khan, U. Hengartner, and D. Vogel. Usability and security perceptions of implicit authentication: Convenient, secure, sometimes annoying. In *Symposium On Usable Privacy and Security*. ACM, 2015.

[20] T. Kwon and S. Na. Tinylock: Affordable defense against smudge attacks on smartphone pattern lock systems. *Computers & security*, 42, 2014.

[21] L. Li, X. Zhao, and G. Xue. Unobservable re-authentication for smartphones. In *NDSS*, 2013.

[22] S. Mare, A. M. Markham, C. Cornelius, R. Peterson, and D. Kotz. Zebra: zero-effort bilateral recurring authentication. In *IEEE Symposium on Security and Privacy (SP)*. IEEE, 2014.

[23] A. Mariakakis, M. Goel, M. T. I. Aumi, S. N. Patel, and J. O. Wobbrock. Switchback: Using focus and saccade tracking to guide users' attention for mobile task resumption. In *33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015.

[24] N. Micallef, M. Just, L. Baillie, M. Halvey, and H. G. Kayacik. Why aren't users using protection? investigating the usability of smartphone locking. In *17th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 2015.

[25] M. Miettinen, S. Heuser, W. Kronz, A.-R. Sadeghi, and N. Asokan. Conxsense: automated context classification for context-aware access control. In *9th ACM Symposium on Information, Computer and Communications Security*. ACM, 2014.

[26] E. Miluzzo, T. Wang, and A. T. Campbell. Eyephone: activating mobile phones with your eyes. In *2nd ACM SIGCOMM Workshop on Networking, Systems, and Applications on Mobile Handhelds*. ACM, 2010.

[27] M. Muaaz and R. Mayrhofer. An analysis of different approaches to gait recognition using cell phone based accelerometers. In *International Conference on Advances in Mobile Computing & Multimedia*. ACM, 2013.

[28] O. Riva, C. Qin, K. Strauss, and D. Lymberopoulos. Progressive authentication: deciding when to authenticate on mobile phones. In *21st USENIX Security Symposium*, 2012.

[29] F. Schaub, R. Deyhle, and M. Weber. Password entry usability and shoulder surfing susceptibility on different smartphone platforms. In *11th International Conference on Mobile and Ubiquitous Multimedia*. ACM, 2012.

[30] F. Schaub, M. Walch, B. Könings, and M. Weber. Exploring the design space of graphical passwords on smartphones. In *Symposium on Usable Privacy and Security*. ACM, 2013.

[31] M. Shahzad, A. X. Liu, and A. Samuel. Secure unlocking of mobile touch screen devices by simple gestures: you can see it but you can not do it. In *19th Annual International Conference on Mobile Computing & Networking*. ACM, 2013.

[32] E. Shi, Y. Niu, M. Jakobsson, and R. Chow. Implicit authentication through learning user behavior. In *Information Security*. Springer, 2010.

[33] W. Shi, F. Yang, Y. Jiang, F. Yang, and Y. Xiong. Senguard: Passive user identification on smartphones using multiple sensors. In *7th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. IEEE, 2011.

[34] S. Trewin, C. Swart, L. Koved, J. Martino, K. Singh, and S. Ben-David. Biometric authentication on a mobile device: a study of user effort, error and task disruption. In *28th Annual Computer Security Applications Conference*. ACM, 2012.

[35] E. von Zezschwitz, A. De Luca, B. Brunkow, and H. Hussmann. Swipin: Fast and secure pin-entry on smartphones. In *33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015.

[36] E. Von Zezschwitz, P. Dunphy, and A. De Luca. Patterns in the wild: a field study of the usability of pattern and pin-based authentication on mobile devices. In *15th International Conference on Human-computer interaction with Mobile Devices and Services*. ACM, 2013.

[37] H. Xu, Y. Zhou, and M. R. Lyu. Towards continuous and passive authentication via touch biometrics: An

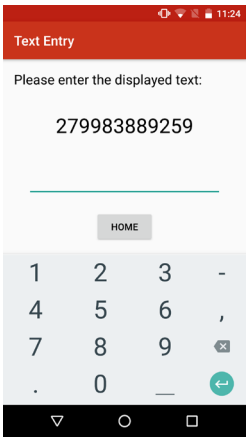
experimental study on smartphones. In *Symposium On Usable Privacy and Security*. ACM, 2014.

[38] T. Yan, D. Chu, D. Ganesan, A. Kansal, and J. Liu. Fast app launching for mobile devices using predictive user context. In *10th International Conference on Mobile systems, Applications, and Services*. ACM, 2012.

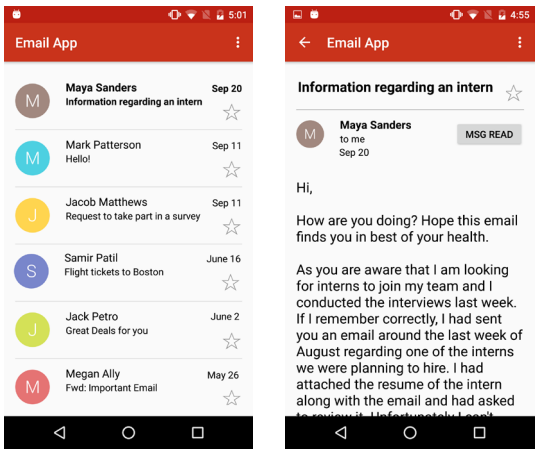
APPENDIX

A. SYNTHETIC TASK SCREENS

Figure 10 provides screen captures for the synthetic tasks performed during the user study.



(a) Text Entry Activity



(b) Email Activity

Figure 10: The activities performed by the participants during the user study. Figure (a) shows the text entry activity containing a 12-digit number, Figure (b) shows the email activity

B. SPLIT-SCREEN CONFIGUATION

Figure 11 provides screen captures for the split-screen configuration. While the *screen transparency* parameter for both Imm-Trans (Figure 11a) and Imm-Dark (Figure 11b) cases were similar to the originally proposed lock configurations, we modified the Grad-Dark configuration (Figure 11c) such that instead of gradually turning the screen dark, we used a vertical slider to gradually hide the content displayed on the top half of the screen.

C. CONFIGURATION PREFERENCES OF THE USERS

Table 2 provides an overview of the participants' re-authentication prompt preferences for the email, contacts and photos app in the bus, office and home scenarios. As mentioned in § 5.3.1, all participants preferred the Imm-Dark-Imm-Lock configuration for the banking app. For each scenario, we mention the proportion of users who are (not) willing to use a particular configuration. Users who gave a rating of 1 or 2 on a 5-point Likert scale were considered to be willing while users who gave a rating of 4 or 5 were considered unwilling to use that configuration.

D. PRE-STUDY SURVEY

Before the study, participants were asked about their security preferences. In addition, we collected demographic information from participants including their name, age group, gender, highest level of education and their current occupation.

D.1 Device Lock Usage

1. Do you currently use a lock mechanism on your phone?
   (a) Yes; (b) No
2. **If they use a lock mechanism:** Which lock mechanism do you use to lock your device?
   (a) PIN Lock (4-digit or more); (b) Password (characters and numbers); (c) Pattern-lock; (d) Fingerprint Recognition; (e) Face Recognition
3. **If they use a lock mechanism:** Who do you want to protect your smartphone access from? (choose all that apply)
   (a) Coworker; (b) Friends; (c) Spouse; (d) Own children; (e) Room-mate; (f) Other unwanted individual or stranger
4. **If they do not use a lock mechanism:** Why do you not use a lock mechanism on your phone? (choose all that apply)
   (a) It takes time to unlock the phone; (b) I don't have any data on my phone which needs to be protected; (c) No one would care what is on my phone; (d) In an emergency, others can use my phone; (e) I have never thought about it

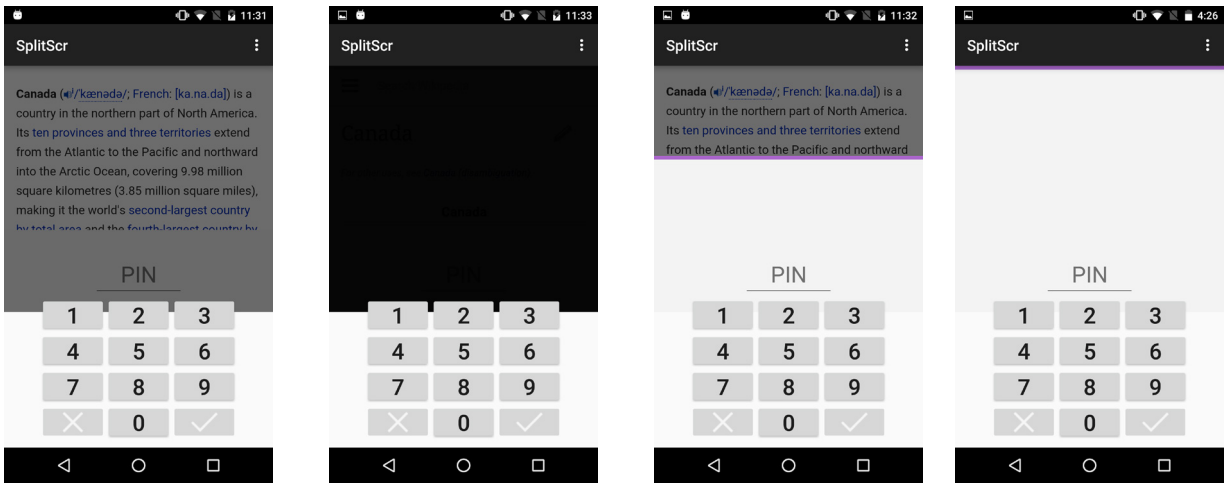
E. STUDY QUESTIONNAIRE

E.1 User perception of individual configurations

After the participants completed both activities using one of the four configurations, we asked them to give feedback on their experience with the evaluated configuration using the following questionnaire.

- Evaluate each of the following configurations that you will observe while doing the experiment. For each category, rate each configuration on a 5-point-Likert scale.

1. Immediate Dark Immediate Lock: Screen turns dark right away and PIN/Pattern appears
2. Immediate Transparent, Immediate Lock: Screen turns and stays transparent and PIN/Pattern appears right away



(a) Imm-Trans (b) Imm-Dark (c) Grad-Dark

Figure 11: The proposed split-screen configurations with varying values of the *screen transparency* parameter. For the Grad-Dark configuration, a vertical slider moves up to gradually hide the content displayed on the top half of the screen.

3. Gradual Dark, Immediate Lock: Screen slowly turns dark and PIN/Pattern appears right away
4. Gradual Dark, Gradual Lock: Screen slowly turns dark and PIN/Pattern appears after a while

(Questions to obtain users' feedback. All questions are on a 5-point Likert-type scale.)

1. Assume someone picks up your smartphone and starts reading your emails. How secure do you find the scheme to protect your data in this scenario?
   (5- Very Secure, 1- Very Insecure)
2. How easy was it to use the scheme?
   (5- Very Easy, 1- Very Difficult)
3. How obstructive was the scheme?
   (5- Not Obstructive at all, 1- Very Obstructive)
4. How annoying was the scheme?
   (5- Not Annoying at all, 1- Very Annoying)

(Once the participant evaluated and rated all four configurations, we asked them to rank them in the order of their preference.)

- Rank the schemes in your order of preference. Please take both the scheme's security and its usability into account.
  (1- Most Preferred Scheme, 4- Least Preferred Scheme)

E.2 Context-based feedback of the configurations

E.2.1 Sensitivity Ratings

Please provide a sensitivity rating of the following apps given how you use your mobile device and how sensitive you think each app is:

1. Email App
2. Contacts App
3. Photos App
4. Banking App

(5- Very sensitive, 1- Not very sensitive)

E.2.2 Scenarios

Now imagine the following scenarios and select which lock mechanism you would prefer in each case. The lock mechanism will get activated in case the system notices any suspicious activity. Please remember that since the system does not have 100% accuracy, it may assume you to be an adversary and you could encounter one of the lock mechanisms while you are using the device yourself. Assume that all of the apps below are protected only with implicit authentication and no other protection mechanism.

Bus Scenario

Imagine you riding a bus and you accidently leave your smartphone on the bus. A stranger picks your device and uses it, which gets detected by the implicit authentication protection mechanism on your device. The stranger may launch different apps on your smartphone. For each app, the implicit protection mechanism could take a different action when detecting misuse. For each of the apps listed below, rank the order of preference of the lock scheme you would prefer with 1 being your most preferred lock scheme and 5 being your least preferred lock scheme.

Please remember that even you could encounter these schemes while you are using your phone on the bus.

1. Views the emails in your inbox
2. Looks at the contacts on your smartphone
3. Views the photos stored on your smartphone
4. Accesses the banking app on your smartphone

		Bus		Office		Home	
		Would like to use?	Would not like to use?	Would like to use?	Would not like to use?	Would like to use?	Would not like to use?
Emails	Imm-Trans-Imm-Lock	27%	**53%**	27%	40%	47%	26%
	Imm-Dark-Imm-Lock	**70%**	13%	**50%**	37%	10%	**70%**
	Grad-Dark-Imm-Lock	**60%**	7%	**67%**	13%	37%	40%
	Grad-Dark-Grad-Lock	37%	33%	**50%**	23%	**63%**	13%
	No Lock	7%	**93%**	7%	**86%**	43%	**50%**
Contacts	Imm-Trans-Imm-Lock	37%	47%	37%	20%	**50%**	33%
	Imm-Dark-Imm-Lock	40%	47%	23%	**64%**	7%	**80%**
	Grad-Dark-Imm-Lock	43%	17%	**53%**	24%	27%	36%
	Grad-Dark-Grad-Lock	**57%**	20%	**70%**	10%	**57%**	13%
	No Lock	23%	**70%**	17%	**83%**	**60%**	40%
Photos	Imm-Trans-Imm-Lock	33%	**50%**	23%	**60%**	44%	33%
	Imm-Dark-Imm-Lock	**77%**	20%	**54%**	23%	17%	**80%**
	Grad-Dark-Imm-Lock	**57%**	10%	**70%**	13%	34%	23%
	Grad-Dark-Grad-Lock	23%	33%	37%	23%	**57%**	16%
	No Lock	10%	**87%**	17%	**80%**	**50%**	47%

Table 2: Configuration preferences of the participants for different apps and scenarios. Values above 50% are in bold.

Office Scenario

Imagine you are in your office and your boss calls you for a meeting. You leave your phone on your desk and one of your office colleagues starts using your phone, which gets detected by the implicit authentication protection mechanism. Your colleague may launch different apps on your device. For each app, the protection mechanism could take a different action when detecting misuse. For each of the apps listed below, rank the order of preference of the lock scheme you would prefer with 1 being your most preferred scheme and 5 being your least preferred scheme.

Please remember that even you could encounter these schemes while you are using your phone in your office.

1. Views the emails in your inbox
2. Looks at the contacts on your smartphone
3. Views the photos stored on your smartphone
4. Accesses the banking app on your smartphone

Home Scenario

Imagine you are watching television at home with your partner and you unknowingly doze off to sleep. Your partner realizes that you are asleep and starts using your smartphone, which gets detected by the implicit authentication protection mechanism. Your partner may launch different apps on your smartphone. For each app, the implicit protection mechanism could take a different action when detecting misuse. For each of the apps listed below, rank the order of preference of the lock scheme you would prefer with 1 being your most preferred scheme and 5 being your least preferred scheme.

Please remember that even you could encounter these schemes while you are using your phone at home.

1. Views the emails in your inbox
2. Looks at the contacts on your smartphone
3. Views the photos stored on your smartphone
4. Accesses the banking app on your smartphone

F. SEMI-STRUCTURED INTERVIEWS

We asked the following questions during the semi-structured interviews:

1. What was your overall impression of the configurations?
2. Would you change anything about these configurations to improve their usability or security?
3. Did you like a particular configuration more than the other?
4. Did you dislike a particular configuration more than the other?
5. Would you be willing to use any configuration on your device for daily use? Why or why not?
6. Any particular scenarios where you think that these configurations will be useful to you?

Turning Contradictions into Innovations or: How We Learned to Stop Whining and Improve Security Operations

Sathya Chandran Sundaramurthy
University of South Florida
sathyachandr@mail.usf.edu

John McHugh
RedJack, LLC.
john.mchugh@redjack.com

Xinming Ou
University of South Florida
xou@usf.edu

Michael Wesch
Kansas State University
mwesch@ksu.edu

Alexandru G. Bardas
Kansas State University
bardasag@ksu.edu

S. Raj Rajagopalan
Honeywell Labs
siva.rajagopalan@honeywell.com

ABSTRACT

Efforts to improve the efficiency of security operation centers (SOCs) have emphasized building tools for analysts or understanding the human and organizational factors involved. The importance of viewing the viability of a solution from multiple perspectives has been largely ignored. Multiple perspectives arise because of inherent conflicts among the objectives a SOC has to meet and differences between the goals of the parties involved. During the 3.5 years that we have used anthropological fieldwork methods to study SOCs, we discovered that successful SOC innovations must resolve these conflicts to be effective in improving operational efficiency. This discovery was guided by Activity Theory (AT), which provided a framework for analyzing our fieldwork data. We use the version of AT proposed by Engeström to model SOC operations. Template analysis, a qualitative data analysis technique, guided by AT validated the existence of contradictions in SOCs. The same technique was used to elicit from the data concrete contradictions and how they were resolved. Our analysis provide evidence of the importance of conflict resolution as a prerequisite for operations improvement. AT enabled us to understand why some of our innovations worked in the SOCs we studied (and why others failed). AT helps us see a potentially successful and repeatable mechanism for introducing new technologies to future SOCs. Understanding and supporting all of the spoken and unspoken requirements of SOC analysts and managers appears to be the only way to get new technologies accepted and used in SOCs.

1. INTRODUCTION

Over the years, there have been a number of research efforts focused on understanding the problems of security operation centers (SOCs). The goal of most of these efforts has been to develop useful operational tools [5, 15, 27]. Researchers have conducted interviews and, in some cases, shadowed security analysts to understand human and organizational challenges [4, 30, 31, 32] in security operations. Most of these efforts resulted in recommendations to developers building tools for SOCs. Despite the correct orientation of these efforts, a common feature of these contributions is that they suggest technical solutions to problems without considering contextual factors that may support or hinder the deployment of the solution. A consequence of the lack of a clear understanding of the operational environment is that the proposed solutions are partially successful at best.

We have been conducting an anthropological study of SOCs at two universities and two commercial corporations for 3.5 years. Our aim has been to understand real operational environments. As computer security researchers and tool builders, one of our major goals was to study the effectiveness of tools currently used in SOCs. With the help of an anthropologist, we trained five computer science students with computer security backgrounds in participant observation methods. The students then took jobs as security analysts in academic and corporate SOCs. They took detailed field notes of SOC events throughout their fieldwork. While documenting events, *e.g.*, usage of a specific tool, they also recorded related activities to establish the context for the event. Without the contextual information the intent behind the recorded actions could not be uncovered during the analysis process leaving gaps in our understanding of the event and its handling.

The motivation for any anthropological study is to obtain insights into various activities humans perform within their cultural context. Each SOC has a culture of its own and it is within that culture that the meaning of tools and processes have to be interpreted. *Activity Theory (AT)* as proposed by Leont'ev [20] and further refined by Engeström [9] is used to facilitate our understanding. At the core of AT based modeling is the notion that humans are collective beings and their activities are goal- or objective-directed. Without an objective there is no meaning to any deliberate human activity. AT also models how we use tools to achieve an objective while emphasizing the distributed nature of accomplishment. Thus, the framework proposed by AT is well suited for analyzing work in operational environments.

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22–24, 2016, Denver, Colorado.

Our most interesting discovery was the existence of tensions and contradictions within the SOC environments. In the SOC context, we found tensions between the analysts and the tools they used as well as conflicts between analysts and various operating rules. We first model SOC operations as an activity (in AT sense) and then list the multiple levels of contradictions that existed in the SOCs we studied. To the best of our knowledge we are the first to systematically identify and study conflicts within SOCs.

Based on our understanding of the systemic tensions in SOCs, our research reveals that the action-operation dynamics from AT indicate a way to resolve certain tensions, *e.g.*, building tools that automate analysts tasks that have become "operations," *i.e.*, repetitive and boring. This frees analysts to perform more creative analytical actions while also generating new tensions and contradictions in the organization and workflow. This process is on-going and tools need to be constantly adapted in a SOC environment as threats change and events evolve. Analysts move constantly between the *acting* and *operating* stages. This is the reason why "static" or inflexible tools fail in SOCs. Our success stories occur when the tools we co-create with analysts keep evolving to resolve new conflicts. It will become clear in the later sections of the paper that the tensions do not always revolve around operational tools. A tool is one component of a set of forces that interact together creating friction due to certain inherent contradictions.

We form a novel "Pentagon Model," an extension of the hierarchical structure of human activity originally proposed in Activity Theory, to capture the knowledge generation and transformation in SOCs and the proper roles of tools in SOC operations. It provides a novel framework within which developers for SOCs can elicit requirements for their tools. We show that identifying and resolving contradictions is a prerequisite not just to building a useful tool but to implementing any novel idea in a SOC. A tool is part of the larger context of SOC workflow and becomes involved in complex interactions that impact multiple dimensions and domains within the SOC. In this way, a tool is not "just a tool" and must be understood within this broader context.

A 3.5 year journey and a substantial amount of data analysis was required to reach these conclusions. In the rest of the paper, we use one story about building an incident response portal for a SOC to illustrate this journey, and explain rationales behind any methods we used in the research and models/results formulated from the analysis.

2. THE STORY OF THE INCIDENT RESPONSE PORTAL

The incident response portal was built for the first SOC we studied, one managing security for a public university in the United States. It consists of a team of 3 to 4 analysts headed by a manager. Each analyst specializes in tasks such as firewall management, incident response, PCI compliance, *etc.* Due to the small team size, the analysts often have to perform non-routine tasks usually done by other analysts. During our fieldwork the students worked as analysts performing these operational tasks. Before continuing, we need to explain our core anthropological research method, *participant observation.*

2.1 Participant Observation

Understanding security operations requires access to operational SOCs and the cooperation of the analysts who work in them. This access is not easy to obtain for reasons that include:

- *The sensitivity of the data handled.* Analysts deal with exploits that can result in loss of valuable information, compromise the privacy of users, or physical damage to infrastructures. A degree of paranoia seems to come with the job. With the academic research literature's current focus on discovery and public disclosure of vulnerabilities, researchers are seen as untrustworthy outsiders. Gaining the subjects' trust is a first step towards performing useful research. Management support is necessary, but not sufficient.

- *The problem of tacit knowledge.* The job of a security analyst is highly complex and decisions are made based on intuitions and hunches that are not documented [26]. In many cases, analysts are unable to articulate what they know or describe clearly the basis for a conclusion or action.

- *The workload.* SOC analysts are always confronted with more incidents than they can resolve. Any process that requires additional efforts but does not directly help the analysts' job is resented.

These factors limit the utility of traditional research methods such as interviews, questionnaires, and passive observation.

Cultural anthropology is a branch of anthropology aimed at studying human beings in their natural settings. The research method employed by cultural anthropologists is *long-term participant observation* in which researchers traditionally spend a year or more within an indigenous population as a member of the community. They take part in the day to day activities and follow the practices of the population. This allows them to obtain an increased understanding of local practices beyond common assumptions about such practices. As they pull themselves deeper into local practices they come to feel and experience the world and may eventually be able to approximate the *native* point of view, in other words, understand how an insider perceives their own culture. This leads to the researcher understanding the symbols, artifacts, and activities as they are perceived by the members of the subject community. Without this understanding, an observer tends to process every event performed by the subjects using the observer's own cultural bias. Such a bias does not lead the researcher to the true reason behind the observed activities [14]. Viewing or attempting to view the activities from the native's point of view is the best one could do in understanding another culture.

The idea of attaining the native point of view resonated very well with our goal of studying security operations because of the well defined closed culture of the SOC. We sought and obtained the cooperation of the SOC management. Our team anthropologist trained five computer science students having a computer security background in participant observation methods which included the observation and note taking that would occur during the fieldwork process.

Over a period of 3.5 years our students occupied positions as security analysts in *four* different SOCs, two in universities

and two in major corporations, a deployment that continues part of the ongoing research effort. The student researchers have worked as level-1 & 2 analysts, incident responders, software-developers, and forensic analysts. They have helped in training security staff and designing security policies, becoming something like "natives" in the SOC cultures, while also keeping detailed notes about their experiences and ongoing SOC activities.

2.1.1 Ethics and Participant Safety

In our research the security analysts and the managers were considered as human subjects. The research was reviewed and approved by the Institutional Review Board (IRB) and analysts completed informed consent forms that explained the research objectives and the voluntary nature of participation. We addressed any concerns expressed, with a detailed description of the nature and expected outcomes of the research. We used aliases when referring to analysts and their managers during discussion and data analysis to preserve their anonymity.

2.2 Why Build the Tool

Early on in our research we observed that the bulk of the analysts' time is spent responding to security incidents reported by external third party entities. The most common of those incidents is malware trying to connect to its command and control (C&C) server. The third party provides the university with information containing the type of malware, the IP address on which the malware activity was observed, usually that of the external interface of the NAT firewall, and the time at which the activity was detected. All this information is sent as an alert via email messages. The responding analyst has to follow the following steps in sequence.

- Identify the internal IP of the infected client from the firewall NAT logs.
- Use the internal IP to identify the MAC address of the infected host from DHCP and/or ARP logs.
- Look up the identity of the user of the infected device using the MAC address from the authentication logs.
- Determine the point of contact (POC) for the incident based on the location of the user (*e.g.*, a department).

Once the analyst obtains all or most of the information, he recommends a potential remediation measure (*e.g.*, format the host disk and re-install the OS), and then puts all the information into a ticket and sends it to the POC. The owner of the infected device also gets a notification about the infection and the recommended remediation steps.

This seemingly simple task is laborious and time consuming. No single tool available at the SOC can provide the direct answer to the question "who is the owner of the infected device," even though the correlations from the multiple logs are straightforward. The deployed security information and event management (SIEM) solution was very slow even for searches on a single week's data. Discovering correlations in the data within the SIEM was almost impossible due to its unacceptably slow performance. The analyst had to manually inspect multiple logs for each of the alerts and it took 10 minutes (on average) to correlate the logs and file a single ticket. The SOC received approximately 15 such alerts per day. It was obvious to our student researchers that the analyst got *burned out* by this repetitive task as did the student researcher tasked to do the same job. He felt that his time was spent on meaningless activity and that he was doing nothing interesting. Further aggravating the situation was the manager's insistence on detailed documentation of the manual method (by the student) so that anyone could perform it.

2.2.1 Reflection on the Process

At this point the student became frustrated by the repetitiveness of his SOC job. This is the moment at which he started to gain the *native point of view* as an analyst. Just as our student researcher was feeling that he had lost the direction of his research, he and the whole research team engaged in a *reflection process*, where the field worker discussed his problems with the rest of the research team. Through this process, we realized that these specific problems can be addressed by building a custom tool for responding to this type of incident. It was clear that this insight arose because the student had reached an essential native point of view unattainable through other means such as interviews. At the same time, it was clear that the student brought uncommon skills, *i.e.*, tool building, to the analyst position.

2.3 How the Tool Worked

In the reflection process, we identified steps in this repetitive process that could be automated. For the malware incident described above the task of a security analyst could be decomposed to answering the following set of questions.

- *What* - Type of threat reported.
- *Who* - Users, IP address, security personnel, *etc.*
- *When* - Time the threat was reported and other temporal information.
- *Where* - Location of the infected device in the network/organization.
- *How and Why* - Context that could have raised the alert, perhaps the most important and interesting.

The analyst was stuck in this process because he was spending more time gathering the basic information such as *who and where* rather than on establishing the context – *how and why*. Our realization was that tools must gather and deduce information along the four basic dimensions of information (what, who, when and where) so that the analysts could spend their cognitive effort along the analytical dimensions (how and why). This insight helped us build the incident response tool.

2.3.1 Automated Incident Response

Together with the analysts we built an incident response portal based on this insight. We used a database to store log information and collected and parsed logs using periodically executed scripts, making the process more efficient. The database also contained a relationship between net blocks and the POC that allowed the notification of the responsible incident response personnel.

The tool has a web interface through which the analyst enters: (1) the external facing IP address and port number where malicious activities were reported; (2) the remote IP address and port number involved in the activities; (3) the timestamp and time zone when the activities were observed. The tool correlates this information and presents the analyst with a filled-in incident ticket with all the required information such as the user of the infected device and the POC.

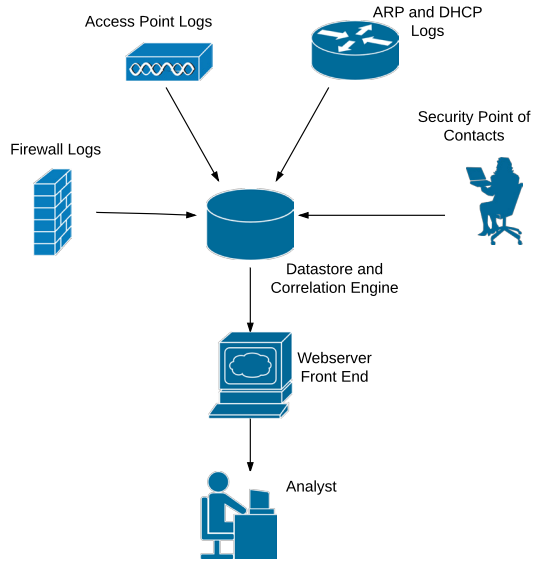


Figure 1: The Incident Response Portal

The analyst then performs the analytical steps answering *how* and *why* the incident might have occurred in the first place. He then suggests possible remediation measures and submits them to the ticketing system. **Using our tool the whole incident response process was reduced from 10 minutes to 10 seconds.** The time saving is due to the automation of the old tasks of manually searching the various logs to establish the *who* and *where* aspects of the incident, now done through automated database queries using the information entered into the web interface. Figure 1 shows the basic workflow of the tool. **This appears similar to a SIEM workflow yet none of the SIEM products that we found in the SOCs provides the automation provided in our incident response portal.**

This shows a major problem in the design methods used for security products. Without understanding the workflow of a SOC and where the friction points are, a tool is useless. Our tool was quickly adopted by our SOC analysts. It not only resolved a major bottleneck in the SOC's workflow, but also broke a major trust barrier for our student fieldworkers. After this tool was successfully built and used by the SOC analysts, the analysts immediately became more open to discussing other challenges in their work to our fieldworkers, and sought our help in building other tools that ease their job. This tool co-creation process was our first major finding in our 3.5 years' anthropological study [26].

2.4 What Happened Afterwards

After this initial success we identified a number of other problems in the SOC that can benefit from automation. The research team developed a number of tools to automate those recurring analyses. The tools were well received and the SOC process was more efficient than before.

Our research went on and we conducted fieldwork at three additional SOCs – another university SOC and two corporate SOCs. Unlike the university SOCs, the corporate SOCs were highly hierarchical. Analysts in one corporate SOC are classified as level-1 (L1, lowest level), level-2 (L2), and incident response (IR, highest level). In this SOC, one of the students worked as L1 and IR analyst while at the same time developing some forensic analysis tools. The other corporate SOC outsourced its L1 tasks to a third party and our student fieldworker took the role of L2 analyst. The corporate SOCs had more analysts (around 22 L1s, 2 L2s, and 5 IRs in one SOC) compared to the university SOCs. Analysts in the corporate SOCs had well-defined roles while in the university SOCs they always had to engage in cross-training and wear multiple hats due to small team sizes.

Through this additional field work we identified the cause of burnout in SOCs using Grounded Theory [25]. We identified the vicious cycles among a number of human, technical, and managerial factors that lead to burnout. We also found a few cases where the vicious cycles were turned into virtuous ones thus mitigating the burnout. In some of those cases the automation of repetitive tasks resulting from tool co-creation was the key enabler.

When the student researchers returned to the first university SOC after a few months, they found that the incident response portal had been rarely used in their absence. We realized that lack of support for the tool was the cause for concerns. New requirements kept emerging and the analysts in the SOC analysts had neither time nor the skills required to customize the tools as the requirements evolved. We then realized that there was more to the success or failure of the tools beyond their technical features.

3. FURTHER ANALYSIS OF THE FIELD WORK DATA

Our experience with the incident response portal encouraged us to return to our field notes and dig deeper to further understand the role of tool building in SOCs and whether there is a guiding principle that could allow us to replicate the success we had in terms of building successful tools to help SOC operations. After six months of analysis, we discovered that an adapted version of a well known model called Activity Theory can form the cornerstone of this guiding principle.

3.1 Activity Theory

The origin of Activity Theory (AT) is found in the works of the Russian psychologists Leont'ev [20] and Vygotsky [28] during the 1970s and 1980s. AT has a proven record of helping researchers comprehend various challenges in work environments. For example, it has been used to study the use of technology in educational environments, to understand the changes brought on by introducing new technology (laptops) into teaching practices [7], and to study the differences between the teachers' beliefs and actual practice when a new tool is introduced in learning [6, 22, 24]. Researchers used AT to understand the effect of new tools on learners, especially their resistance to newly introduced technology for learning, and on highlighting how old habits impede the adoption of new tools [2].

The AT model in Figure 2 is adapted from Engeström [9]. Elements of the original model are shown in parentheses and in red font. Un-parenthesized elements result from our application of the model to SOC operations. Engeström defines an activity system as *object oriented, collective, and culturally mediated human activity* [12]. The fundamental

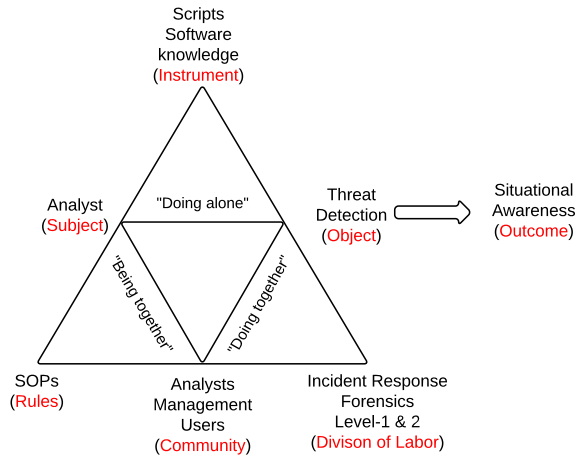


Figure 2: Activity Theory Model of Security Operations

idea of AT is that humans perform tasks to achieve an objective. Without that objective the task has no meaning. The inner downward-pointing triangle symbolizes the interactions of individuals and the collective in achieving an objective. Each edge in the downward triangle represents the relationship between the three nodes [9]: (a) an individual does certain tasks to achieve an objective, (b) an individual is part of a social structure represented by the community node, and (c) the community of which the individual is a part of acts together to achieve an objective. Furthermore, the three relationships are mediated by three different aspects – instrument, rules and division of labor, forming the encompassing upward-pointing triangle. In trying to accomplish their objective humans use certain tools or in AT terms *instrument*. The tools can be physical, such as a hammer when breaking rocks, or symbolic such as language for communication. AT further states that human beings do not act in isolation but within a community. There are certain rules that govern interactions among the members of the community. In order to achieve their objective, people take up different roles (division of labor) based on their expertise.

According to AT, tool mediation – design, use, and preservation of physical and symbolic instruments – is seen as a major distinguishing factor between human and animal activities [9, 17]. The two triangles in the AT model together represent three different types of mediated interactions [17]: (a) subject-object interaction is mediated by Instrument, (b) interaction of subject with their community is governed by Rules, and (c) a community achieves their objective by taking up specific roles corresponding to Division of Labor. The three different mediations arise due to social, cultural, and cognitive aspects of human life.

A SOC can be modeled as an activity system where the subjects are the analysts and their objective is to monitor/mitigate threats and provide situational awareness. To achieve this objective they use tools such as SIEM, homebrewed software and scripts, and their knowledge in computer security. The community they interact with includes other analysts, management, and end users. The traditional rules governing the communication between analysts and other stakeholders are the so-called standard operating procedures (SOP). SOPs recommend course of action for every incident type guiding the analyst in drafting a communication and mitigation plan in response to a security incident. Analysts also assume roles on the operations floor, *e.g.*, level-1 (junior) analyst, level-2 (journeyman) analyst, incident responder, forensic analyst, *etc.* Under this interpretation, it is easy to see that a SOC work environment fits nicely within the AT framework.

AT has been successful in understanding distributed human activity ranging from primeval hunting to modern day work environments. So it is natural that SOC operations can be successfully captured by the AT model. AT also sheds light on the use of tools by humans in achieving their goals in collaborative activities. Since one of our goals were to obtain insights on the role of tools in SOC operations, it further convinced us to use AT to drive further analysis.

3.2 Analysis Methods and the First Result

Throughout the 3.5 years of fieldwork spanning 4 SOCs we observed many recurring patterns and similarities in their problems. Due to the large amount of field note data, a systematic approach is needed to ensure the objectivity and comprehensiveness of the analysis.

Our analysis of field note data is both inductive and deductive. It is inductive in the sense that we look for patterns in data *without* any preconceived hypothesis. As we formulate theories to explain the patterns we found in one part of the data, we also test those theories on the other parts of the data. In this sense our analysis is also deductive. To facilitate this type of analysis, we leveraged a qualitative data analysis technique called *template analysis*.

3.2.1 Template Analysis of Data

Template analysis is a qualitative data analysis technique developed by Nigel King [18]. It is useful when the researcher has a partial understanding of the concepts to be identified in the data. This technique starts with an *a priori* set of codes or themes that the researcher is interested in and the codes evolve as the analysis is performed. The technique is flexible in that the researcher starts with some preconceived concepts but can also identify and add new concepts as they are discovered. Below are the steps in the template analysis process.

Define a priori themes A set of themes are developed based on the concepts the researcher is interested in identifying in the data.

Transcribe and familiarize The researcher reads through the field notes and familiarizes herself with the data she is going to analyze.

Initial coding Parts of the field notes that are relevant to the research questions are identified. Then the *a priori* codes are attached to those parts of the data wherever they are applicable. When a section of fieldnote data matches the research question but no existing code could be applied, a new code is devised or an existing one is broadened to cover it.

Produce initial template Once a subset of the data is coded a set of themes is generated. These form the initial template. The template might have a hierarchy of codes within each of the themes.

Develop the template The initial template is applied to the entire data set repeatedly. Modifications to the template are performed whenever a text does not fit

into the template. This iterative process refines the code set and a final template is produced.

Interpretation At this point, the researcher has coded the entire data using the developed template and writes up her findings based on the final template.

Quality checks and reflexivity The researcher periodically consults with an expert team that includes fellow researchers on the project to ensure quality of the analysis she performs. The coding researcher must also perform frequent reflections to make sure her own personal beliefs and biases do not affect the interpretation of the collected data.

A study by Frambach *et al.* exploring the effect of globalization on medical education provided the inspiration for combining AT with template analysis [13]. Following this work, we began by looking for the basic elements of the AT model in our fieldwork data and found that the model provided substantial explanatory power for understanding work carried out in SOCs. We then applied more concepts from the AT theory to further understand the data. Thus we first developed a list of codes based on the AT model and performed data coding. New codes were added as new themes emerged. This continuous application of template analysis eventually resulted in **one of our main discoveries in this paper: the existence of contradictions in SOC operations and its key role in preventing SOCs from doing an effective job.**

4. CONTRADICTIONS

A key feature that arises when using AT to study work environments is the notion of *contradictions*. From AT perspectives, contradictions are defined as "a misfit within elements, between them, between different activities, or between different developmental phases of a single activity" [19]. Some researchers have referred to contradictions as *systemic tensions* [1]. Other definitions include "unintentional deviations from the script [which] cause disco-ordinations in interaction" [11] and "problems, ruptures, breakdowns, clashes" in activities [19]. Engeström [10] even recognized contradictions as "the motive force of change and development" [12]. In a typical scenario when contradictions arise, individual(s) begin to question the established norms and start to deviate from the rules. A positive outcome is that individuals get together and develop a new course of action that resolves the original contradiction leading to a better workflow [10].

4.1 Primary Contradictions

A tension that exists within a single node in the AT model (Figure 2) is called a primary contradiction [9]. In a work environment, these tensions arise due to the dichotomy between the "professional logic" of the employees and the "commercial logic" imposed by their organization [3]. The professional logic of security analysts (subject) dictates that they constantly improve their skills and be efficient in detecting and mitigating security threats. On the other hand, SOCs are under constant pressure to demonstrate their value to the parent organization. This results in a number of metrics being defined to measure the performance of SOC analysts. Ultimately, the job of the analysts is skewed very much towards generating those defined metrics. This creates a conflict within them. **They are confounded with two non-identical objectives – doing the *right thing* versus the *required thing*.**

Returning now to the incident response portal story, the analysts' frustration was caused by a conflict between their desire to continuously improve their skills and thus wanting to handle more intellectually challenging incidents, and the fact that SOC management emphasizes metrics such as number of resolved incidents instead of the complexity or subtlety of the incidents. As an analyst one has to tend to both these objectives which are often in conflict with each other. The analyst can choose to close a high quantity of easy tickets (thereby scoring high marks on managerial metrics) or attend to more complex incidents that may be more fulfilling. This leads to frustration and eventually burnout. This contradiction is faced by the analysts within themselves; that is, it is a contradiction that exists inherently in the "Subject" node of the AT triangle of Figure 2.

We went back to our field notes to find more examples of such primary contradictions. Following the template analysis methodology, we coded our data with the initial goal of identifying contradictions in the SOC's operations. The initial template generated as a result of coding a subset of the data is shown in Table 1 in Appendix A. The initial template was then used to code the entire field notes, resulting in the final template which was used to interpret the results. Below we illustrate some findings from the analysis.

4.1.1 Primary Contradiction within Subject (Analyst)

In addition to the frustration we witnessed in the first SOC, this primary contradiction within analysts is observed across the SOCs we studied. One analyst in a corporate SOC noted:

> "I wanted to work in an environment where there will be continuous learning and I have started to feel that I am not learning anything new in my current job. In fact, I took the current job hoping to analyze malware every day and learn more in that process. I feel that the SOC currently is not doing any real threat detection which in turn is limiting my opportunities for learning. I have decided in my life, to spend a significant amount of time for the improvement of my career. Now I feel bad that my commitment is not paying off."

In another instance a SOC manager asked his analysts to work towards generating metrics:

> "There will be metrics collected for all analysts from the case management tool (CMT) so that a report can be generated and shown to the upper management. If the team has to scale, handling a number of cases, we need to produce numbers to show to upper management. So far this is being done through success stories and this does not scale as it looks very general. Some part of the management is also interested in knowing the impact our team has on the infrastructure. Go over the metrics and say which ones make sense and do not. You have to live with it and get involved. If you do not get involved now then when the change is made into CMT you will have to provide the data. I do not want to push it out

there without questioning and for the sake of doing it. I also want to measure the fidelity of the incident. Features in CMT that do not lead to any metric must be removed."

4.1.2 Primary Contradiction within Instrument (Tools)

Security analysts use a number of tools to perform their job. Some of them are physical such as software and scripts, while others are cognitive, such as knowledge and training. In an ideal case tools will help analysts become efficient in their job. From the professional-logic perspective this is the true purpose of a tool. Interestingly, the tools in operation floors are purchased instead due to reasons not aligned with efficiency. Typically the most expensive product in a SOC, SIEMs are purchased because they are considered information security "best practice." Ironically, most of the SIEM solutions we saw deployed at the SOCs were not up to the task of basic event correlations necessary for incident analysis, as illustrated in our incident response portal story. **Here the commercial logic for having the tool is compliance not operational efficiency, resulting in this primary contradiction.**

> In one of the corporate SOCs, the management decided to use a particular case management system (CMS) due to the support it provided with the existing SIEM solution. While the integration seemed helpful at the beginning, the CMS turned out not to fit the workflow of the SOC. The CMS was never replaced, which subsequently lead to secondary contradictions with the analysts (Section 4.2.2).

4.1.3 Primary Contradiction within Rules (SOPs)

As we noted earlier, the rules in SOCs are the standard operating procedures, or SOPs. The purpose of SOPs is to make sure for a given incident every analyst will respond in a similar way. In other words, they ensure predictability in operations. However, there is a fundamental conflict that SOPs face which is between expected behavior and creativity of analysts. Security operation is a distributed activity involving a number of analysts. If they are encouraged to act their own way *all the time* there will be chaos. On the other hand, one does not know when to deviate from the norm and try out new techniques. This inflexibility hinders detecting and mitigating threats which are constantly adapting. This dualism is at the core of the conflict that exists within the SOPs used in operations.

For example, an analyst encountered an operational scenario where he had to email a member of a business units to validate an alert but was very hesitant to proceed. After waiting for a while he contacted a senior analyst and asked him for advice on how to proceed. The junior analyst specifically said that he did not know how to proceed as this scenario was not covered by any of the procedures. This example demonstrates a familiar problem we encountered throughout our fieldwork. While SOPs can empower an analyst within limits, the same SOPs can dis-empower the analyst from acting beyond them.

4.1.4 Primary Contradiction within Division of Labor

In work environments, the division of labor is achieved by assignment of roles to employees. In a SOC typical roles include level-1&2, forensics, incident response, and content development engineer. The role assignment ensures that people have the right skills and expertise for the assigned task. There exists a dualism within division of work that leads to efficiency problems. The very specific role assignments to analysts leads to analysts working in silos; thus they often lack empathy for other analysts. On the other hand, analysts have to constantly work with their colleagues in other roles; the lack of empathy creates barriers in this collaboration, thus fundamentally defeating the purpose of division of labor.

For example, a level-1 analyst was frustrated about the high volume of events generated by a rule written by a level-2 engineer:

> "The engineering team is very stubborn. Jack (name changed) thinks that he knows everything and does not understand the frustration of analysts."

Likewise, upper-level analysts become frustrated by those in lower levels. Level-1&2 analysts escalate incidents to incident response teams whenever they require assistance. One day the incident response team members complained that they were getting too many escalations. Having worked at both teams the fieldworker found the two teams to be completely unaware of the priorities, problems, and concerns of each other.

4.1.5 Primary Contradiction Within Objective

Finally, there is also a primary contradiction within the objective of the SOC itself. The primary objective of the SOC as commonly understood is to detect and mitigate security threats for their parent organization. Perversely, the better a SOC gets at detecting/preventing threats the harder it becomes to show their value to the organization.

In one of the corporate SOCs alerts that were insignificant were deliberately left unoptimized as optimization would reduce the number of alerts in the stream. Fewer alerts would then mean that management would perceive that the SOC team could do their job with less number of analysts and the parent organization would then put pressure on the SOC management to reduce the team size by laying off some of their analysts. As a result analysts have to deal with a large number of useless events and eventually get worn out.

4.2 Secondary Contradictions

The existence of primary contradictions will also create conflicts *between* elements of the AT model. In AT these are called *secondary contradictions* – tensions that exist between any pair of nodes in the AT triangle of Figure 2. They are a manifestation of the inherent primary contradiction within the single nodes [3]. Our template analysis revealed a number of pair-wise contradictions in SOCs.

4.2.1 Subject - Rules

Throughout our study we observed a constant tension between the analysts and the standard operating procedures (SOPs) they are required to follow. A security analyst wants to solve intellectually challenging security incidents. This requires using novel analysis methods that are not in the SOPs. The SOP rules do not provide enough freedom for an analyst as there is a written down procedure for every

type of incident. The mundane nature of executing procedures time and again hinders creativity. The rules define the tasks of the analyst based on the opinion from the management. The SOC management wanted the SOPs on the argument that SOPs help ensure predictable performance of the SOCs (commercial logic). But at the same time this prevents the analysts from being creative in their jobs (professional logic), and thus prevents them from being more efficient in operations. The secondary contradiction, *i.e.*, the conflicts between the subject (analysts) and the rule (SOPs), is a manifestation of the primary contradictions inherent in the analysts and in the SOP which we discussed before. This secondary contradiction is also a main cause of the frustration at the first SOC we worked at that eventually led us to develop the incident response portal to help address.

As one analyst in another SOC complained:

> "The procedures were turning us into robots. The procedures were so detailed at some point that all the analysts were doing was to click and fill in data."

If not tended to, this contradiction has been found to cause adverse effects such as analyst burnout leading to frequent turnovers, as pointed out in our prior work [25]. Periodic review of rules to identify patterns that could be automated is one way to mitigate the effects of this contradiction, but this is not done often enough (or at all) in most SOCs.

4.2.2 Subject - Instrument

From the perspective of technology transfer, this contradiction is the most interesting to explore as it involves interaction of analysts with technology. The SOCs we studied did not have the right tools to help their analysts as most of the tools were developed without proper understanding of the analysts' workflow. A top-down decision was made by the management on the type of tools to be procured for the SOC. This is essentially a manifestation of the primary contradiction within the tools (Section 4.1.2). As a result SOC tools have suffered from a number of shortcomings.

One of the major concerns about the tools in SOCs pertains to *poor attribution*. To make the best decision a security analyst must be provided with all the temporal and spatial information related to an alert. The purchased tools were designed with no knowledge of operational workflows and thus completely missed this aspect. Analysts were provided with partial information making it hard to attribute the alert to an owner or a device. In another case we observed, analysts were not able to query the wireless domain controller to extract the authenticated user IDs along with the device host name because the vendor had not anticipated this need and decided not to provide that feature. Such shortcomings result in analysts spending most of their time performing low-level data processing tasks to gather the missing information, rather than creative investigation.

4.2.3 Division of Labor - Instrument

A SOC is comprised of analysts with specific role assignments. In order to achieve the goal of division of labor, where analysts perform the tasks they are good at, it is imperative that they have the right tools to assist them. The preference for features in a tool depends upon the role and technical expertise of the analyst. A forensic analyst might like to use a Linux desktop and might be comfortable using a command-line interface. A compliance analyst whose primary task is to check for conformation of systems to rules might be comfortable only with a graphical user interface (GUI). Tools are oftentimes purchased based on the managerial logic that interferes with the preferences and requirements of the analysts (Section 4.1.2). As a result, analysts in different roles could not accomplish their tasks and the purpose of dividing work based on expertise is defeated.

This contradiction is well illustrated by our story with the incident response portal. After the tool was built, the process of responding to the malware incidents was simplified to the point that it could be handed off to the Network Operations Center (NOC) of the university. The NOC analysts were less skilled compared to the SOC staff and their job was to handle cognitively less intensive tasks. Our tool, however efficient in handling malware alerts, was not ready to be used by the NOC staff simply because it used a command line interface. The conflict our tool ran into was between Division of Labor (skill set of analysts) and Instrument (tools they had to use). The incident response portal exposed an interface that required more cognitive work than the NOC analysts are comfortable with. As a result, the SOC's effort to transfer this task to NOC did not happen for a long time, and the more skilled SOC analysts were still stuck performing the mundane ticketing task for malware incidents (though more efficiently than before).

We resolved this contradiction by providing an alternate web interface to the portal in addition to the command line access for SOC staff. The web interface abstracted away a number of technical tasks and pushed them into the background. The NOC staff were then able to file malware tickets at the push of a button. Clearly, the same tool needs to have multiple interfaces depending on the type of analysts who will be using it. Otherwise one cannot get the expected benefit of distributing work among analysts.

It is important to note that this is also an example of how an attempt to resolve one contradiction may create a new contradiction. The tool was originally designed to improve the work of SOC analysts, but it ultimately had an impact on the division of labor, being accepted as a tool for the NOC. But here the tool failed because it had been designed with a command-line interface for the SOC. This highlights the fact that **conflicts will keep emerging in a SOC no matter how much you can do to improve its process. Such conflicts must be resolved on a continuous basis.**

5. FROM CONTRADICTIONS TO INNOVATIONS

The previous section discussed the contradictions we identified in SOCs during our anthropological study. Each contradiction requires a different course of action to be resolved. Some measures are technical while others are managerial. The rest are influenced by economic considerations. This leads to a question of particular interest to the audience of this conference:

Can technologists do something to turn some of the contradictions into innovations? If so, how?

Contradictions are at the heart of Activity Theory and they are the potential triggers for workplace innovations [9, 12]. When we looked back at our fieldwork data we realized that

it was by identifying and resolving certain contradictions that we succeeded in bringing an innovation to security operations.

Let us return once again to the incident response portal story. The analysts were stuck performing a high volume repetitive task. Neither the analysts nor the field workers could invest time in any creative security projects because the repetitive malware incidents had to be taken care of as high priority. The analysts would get penalized if they did not close the malware tickets in a timely fashion as required by their manager. They have to balance between two conflicting motives of their job: engage in creative security analysis, and resolve the constant stream of incoming security alerts. The presence or lack of the right tool will either reconcile or aggravate the two contradictory motives. The incident response portal we built resolved/mitigated a number of contradictions manifested in this story.

Our tool was built in the context of the SOC environment and hence fits the operational workflow. Our tool development process is *analyst-developer co-creation*. In this model the fieldworkers are also analysts themselves, and they engage in developing tools that aid in analysts' work. As fieldworkers, we switched hats between developer and analyst to enable co-creation within ourselves. This addressed the secondary contradiction of tools falling short of analysts' expectations (Subject - Instrument in Section 4.2.2). The incident response portal reduced the ticketing time from 10min to 10sec, allowing the analysts to close the immediate incidents more quickly. As a result they will have more time for creative analysis. Therefore the incident response portal mitigated the primary contradiction within the analysts (Section 4.1.1), since they can now more easily balance the two conflicting objectives of their job. The tool also mitigated the primary contradiction within the tools (Section 4.1.2). While the SIEM used by the SOC (considered a must-have due to "best practice") was not up to the task, the incident response portal bridged this gap by introducing some real value (helping analysts in their job) into the SOC's tool box.

We continued to conduct template analysis on the field notes to revisit all the cases when we built tools for SOCs. Every one of them confirmed to us that **the reason the tools we built were adopted by a SOC and became useful was because they all helped resolve some contradictions in the SOC. They will keep being useful and used by the SOC as long as we continue updating the tool to resolve new contradictions as they emerge (including contradictions that emerge in part due to the tool itself). If we stop the process of identifying/resolving contradictions, the tool will stop being used in the SOC.**

After combing through all the success and failure stories of our tools in the SOCs we studied, we further realized that the process of resolving contradictions in a SOC can be placed in the proper perspective by looking at another important aspect of activity theory – the dynamic nature of activity.

5.1 Human Activity Dynamics

Humans performing an activity operate at multiple cognitive levels in achieving their objective. We use an example by

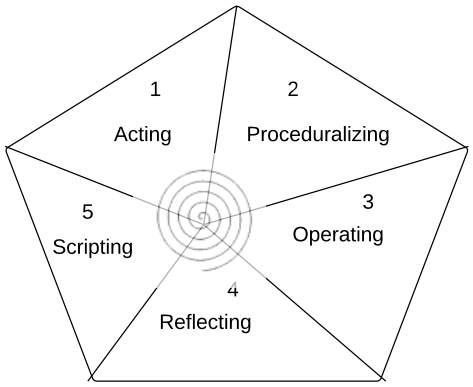


Figure 3: Pentagon Model for Knowledge Transformation

Kaptelinin *et al.* to give further insight into this hierarchy of activity [17]. The example sheds light on the non-stationary nature of the hierarchy, *i.e.*, the hierarchy evolves over time and the importance of specific actions shifts. Consider the activity of learning to drive a car. For the first few days, the learner consciously performs tasks such as changing lanes, looking in the mirrors, and shifting gears. Each of these in AT terms is called an *action*. Broadly, human tasks that require explicit attention are categorized as actions. The high level of cognitive effort required by each activity prevents the learner from multitasking during the learning period. With practice and continued instruction, the actions become second nature and can be performed subconsciously. At this point, they become internalized and are now called *operations*. The cognitive effort needed to perform operations is almost negligible, thereby enabling multitasking. The ability to perform operations persists even after years of non use. One never forgets how to ride a bicycle[1]. We refer interested reader to a more detailed discussion in Appendix B.

We now look at the process that we carried out when turning some of the contradictions in security operations into innovations, through the lens of the hierarchical model of human activities.

5.2 Activity in Security Operations

We found the action-operation dynamics to be applicable to tasks performed by security analysts. Steps such as log analysis, filing incident tickets, and communicating with stakeholders, when performed consciously by an analyst can be categorized as *actions*. After repeated applications these steps can be internalized within an analyst and be performed with very minimal cognitive effort, at which time they become *operations*. Our template analysis revealed that the action-operation transition in SOCs involves some interesting aspects of knowledge transformation. Specifically, **our analysis identified three additional stages in this transition that are *not* present in the traditional AT literature**. Figure 3 shows what we call the *pentagon model* for knowledge transformation in SOCs. The five stages of activity repeat as a cycle; each stage is described below.

Acting Analysts in the acting stage are handling a new se-

[1]There is a neurological explanation for this. See `http://www.abdn.ac.uk/news/3275/`.

curity incident, *e.g.*, a zero day or previously unidentified incident. A new incident does not have an SOP or other written procedures for its handling. As a result, the analysts have to consciously perform each step of the investigation. This stage requires a creative mindset and demands a high cognitive effort from analysts.

Proceduralizing Once analysts understand the incident, they develop a procedure for handling similar incidents. Documentation needs to be written describing the procedure. This ensures that other team members are aware of the new incident handling process and preserves the knowledge. This is one of the newly identified stages of the activity hierarchy. Because documenting the procedure usually requires multiple iterations and is a cognitive activity distinct from handling the original incident, it deserves its own place in the hierarchy.

Operating The operating stage occurs when the procedure for handling the new incident is mature and predictable enough for the analyst to perform it subconsciously. There is a self-contradictory nature to this stage. On the one hand, the cognitive effort needed to perform the procedure has become minimal or nonexistent. On the other hand, when the analysts are in the repetitive operating mode (for periods of days) they do nothing creative. This can lead to severe problems such as burnout [25] and partially explains the high turnover rate among SOC analysts, unless a separate set of people with suitable personalities are tasked with this job.

Reflecting This is the second of the three new stages of the SOC activity we identified. Reflection is a process whereby analysts identify aspects of the operational tasks that have become repetitive and require little or no cognitive effort. These are candidates for automation or for delegating to a lesser skilled organization. In a highly efficient SOC, this is performed as often as once a month. We have observed operational environments where no reflection takes place. Analyst burnout and a high turnover are more common in these environments.

Scripting In the scripting stage analysts, either themselves or by working with a development team, automate aspects of incident handling that have been identified as candidates for automation in the reflection process. Usually these are scripts written in rapid development languages such as Python or Ruby. However, implementation can also be done via long-term developmental efforts using web frameworks or coding in a lower level language. This is the third new stage we identified in the SOC activity.

5.3 Automation and Conflict Resolution Revisited

Every new analytical task starts being performed consciously by an analyst (*acting*). The task then, after some stabilization, is documented as an SOP (*proceduralizing*). The stabilized task is eventually internalized by analysts (*operating*). Most SOC managers and analysts stop at this stage. As explained in the previous sections, this will result in primary and secondary contradictions within and between analysts, their tools, and SOPs, leading to frustration and burnout. Let's look back at the contradictions we saw in the incident response portal story. The analysts got frustrated and burned out because they were stuck in the operation stage and did not have any time to think about new threats and problems. Automation of the repetitive operations resolved this contradiction and allowed the analysts to move from the operation stage to the acting stage. This also allowed for the analysts to be more prepared to deal with new threats.

Unfortunately, our fieldwork finds that the process of incremental automation in SOCs is predominantly reactive. Scripts are written only in response to high workload, such as when the volume of an alert stream is too high. We propose that senior analysts and managers should conduct periodic reviews of analytical tasks and identify those that have been *operationalized* within the analysts. In other words, the review should focus on identifying aspects of SOPs that have become cognitively repetitive for the analysts. Those tasks could then be automated *proactively* by either the analysts or software developers with the requirements provided by the analysts. Our incident response portal is an outcome of such a process. Tools created this way will fit well within the cognitive analytical process of analysts and free them to perform more creative tasks.

The pentagon model is also well aligned with the nature of detecting and responding to cyber threats. The variety of security threats evolve rapidly these days demanding creative analysis. Analysts must remain in the conscious *acting* stage as much as possible to be effective. Tools developed following the pentagon model are not static. The constraints that determined the requirements of the tool might change creating new conflicts. This will first push the tasks back to the acting stage demanding manual intervention by analysts and developers. Using the co-creation process, the tool can be adapted to resolve the new conflicts by going through the reflecting and scripting stage again.

Implications of Pentagon Model for Analyst Burnout

The net effect of the cycle in the pentagon model is to recognize that a new incident serves as a potential harbinger for a flood of similar incidents to come. Converting its mitigation from a challenging cognitive task to something that can be offloaded or automated, frees the more capable analysts to meet the next challenge. Thus the cycle repeats. There is another potential problem that we identified in the model – *the rate of transition from the scripting to the acting stages.* If the arrival rate of new incidents exceeds the rate of the cycle time in the model, burnout may occur despite the cognitive challenges, due to the lack of time to automate the operation. If the arrival rate of new incidents is much lower than the rate of the cycle time, burnout may be supplanted by boredom which also leads to a high turnover.

6. TOOLS AND BEYOND

The incident response portal was part of a broader workflow innovation process. The tool would have no meaning if one removed the objective the SOC wanted to attain using that tool. The SOC wanted to implement a hierarchy in the operational workflow. Its staff is composed of highly skilled analysts but a small team. They wanted their job to be-

come investigating *novel* incidents and devising mitigation plans to deal with similar events in the future. They could then write down an SOP document listing out the steps that should be taken to respond to each of the novel incidents. Once the response steps have become stable enough and highly repetitive, they can then transfer it to teams composed of less skilled analysts such as the Network Operations Center (NOC). This ideal did not happen until our fieldworkers helped the SOC identify and resolve a number of contradictions in their workflows by building the incident response portal.

It is within this background that the development and deployment of operational tools must be viewed. Hence it is appropriate to say that resolving contradictions is a prerequisite for not just developing successful operational tools, but to implement any novel idea in SOCs. And due to the complex activity system in which tools and new ideas are deployed, they must be continually updated and re-adapted to address new and emergent contradictions, some of which are created by the innovation itself.

6.1 Conflict Resolution is a Sensitive Process

Identifying and turning contradictions into useful innovations is a challenging task. The chance of a contradiction becoming a useful workflow improvement depends largely on first acknowledging the contradiction [23]. Many contradictions go unnoticed due to a variety of factors including lack of management support or denial by those affected. During the fieldwork we observed many contradictions that were never spoken of by L-1 analysts fearing repercussions. It has been observed that turning a contradiction into an innovation does not happen only at an individual level. A collective effort by the community is needed and tools used by the community may need to be transformed together to enable the innovation [29]. The incident response portal required collaborative effort from the analysts and fieldworkers who acted as analyst/developer. The tool's development required the approval of the SOC manager who allowed the analysts to spend their work time in the co-creation process. Due to different roles and objectives within the activity system, it may be difficult to achieve sufficient consensus around an innovation. Sometimes contradictions are not openly discussed because they are just embarrassing [8]. SOC analysts frequently encounter security breaches; discussing the problems in handling security incidents with other people will put them in a bad light.

In our work, the use of anthropological methods helped us earn the trust of analysts in discussing embarrassing or otherwise undiscussable contradictions. We worked as analysts ourselves and hence were able to experience the contradictions first hand. **It becomes clear that building *trust* among analysts and between various SOC teams is a key enabler for acknowledging and discussing contradictions, and is thus a pre-requisite for bringing about useful innovations to SOCs.** SOC managers must view friction in operations as opportunities for making things better rather than simply reprimanding the analysts. Above all, managers should earn the trust of their analysts and be a participant in the conflict resolution process as they are the authoritative persons to bring actual changes to operations.

6.2 Conflict Resolution is a Continuous Process

As mentioned in Section 2.4, we returned to the SOC where the incident response portal was deployed after a brief hiatus of a few months. To our surprise we found that our tool was shelved and not used by SOC or the NOC staff. As we renewed our fieldwork, which involved continued co-creation, our tool once again was adopted into daily operations by the analysts. Reflecting back on this experience, our incident response portal was temporarily out of operations due to the hiatus in conflict resolution when the fieldworkers were absent in the SOC. This led us to the realization that **successful tools must address contradictions on a continuous basis for their continued usefulness**. This explains why the SIEM solution at this SOC (and at other SOCs we studied), which was essentially a static tool, was barely functional. In short, human activity is a dynamic system. If a tool is to be and remain effective, it must also be dynamic.

7. DISCUSSION

Our conclusion that useful tools for SOCs must help resolve the various contradictions in the work environment on a continuous basis seems to be at odds with how security product vendors produce technologies these days. Many vendors still view this as a "build-once-sell-to-everyone" market, without much understanding of the variations in the workflows and contradictions that may arise within the various SOCs they tend to sell the products to. Our research results imply that tools built this way will not work effectively to help SOC analysts. It seems to follow that useful security tools for SOCs may best be built within SOCs, by people who can identify and understand the contradictions within the work environments. Our experience in the anthropological study shows that to achieve this understanding, it takes a person becoming an analyst and doing the job in the SOC.

Our pentagon model highlights the importance of the "reflecting" and "scripting" stages in SOCs. Unfortunately oftentimes SOC management does not understand the importance of automation and does not allocate enough work force to ensure analysts have time to perform reflection and automation. As a result the analysts are stuck in operation mode, leading to burnout. On the other hand, when the event rate is low, simulation-based approaches could be used to generate events that turn analysts to the acting mode when there are not enough real interesting events.

The ability of analysts to transition to acting stage in the pentagon model depends on their skill set to do rapid software prototyping. In our work the student fieldworkers were skilled programmers, and at the same time security analysts. This allowed them to develop tools that automate the operations. We found that a typical analyst has two problems when it comes to developing quality tools. The first issue arises from a lack of time to write code. In operations, priority is given to handling incidents and responding to tickets. A large number of events per analyst means that analysts do not get the right amount of time to write software, and are not even encouraged to do so. The second issue is that some analysts just do not have the skills to program. As discussed above, good tools can be written only when you actually do the job. This implies that the analysts may be the right people to develop the required tools, which begs

the question of whether programming ability should be a desired qualification for SOC analysts.

8. LIMITATIONS

The main limitation of our work is the subjectivity of the researchers in collecting and analyzing fieldwork data. To address this limitation the collected data was anonymized and shared with the entire research team and extensively discussed. The results presented here are based on our collective study of four different SOCs by five student fieldworkers and a number of senior researchers. We acknowledge that in order to further generalize the findings we need to expand our study to even more SOCs. However it is also important to point out that in our experience thus far SOCs can be very different and overall generalization may be unwarranted and misguided. It might instead be fruitful to pursue a more particularist approach, in which each SOC is studied within its own terms and in an effort to understand its own tensions and contradictions. In this regard, the generalizable aspect of our work is the approach in the work. After further study of more SOCs it may become apparent that the primary and secondary contradictions identified here are evident in all SOCs. Or further study of several SOCs might eventually result in the creation of a typology that can identify different types of SOCs with different sets of tensions and contradictions. Furthermore, we hope to expand our analysis to explore tertiary and quaternary contradictions – contradictions between different activity systems and business units within broader organizations. This is an ongoing effort and we hope to conduct similar studies to gather more insights. Notwithstanding these limitations, we would like to emphasize that conducting long-term anthropological study for SOCs is a process that yields perspectives that are otherwise unobtainable.

9. RELATED WORK

The use of anthropological methods to study SOCs and the idea of co-creation as a means to develop usable operational tools was first reported by us in our prior work [26]. Continuing our anthropological study, we then studied the problem of burnout among security analysts [25]. The work identified multiple vicious cycles between a number of human, organizational, and technological factors to be the primary reasons for burnout and high turnover of analysts in SOCs. The work presented in this paper uncovered a more fundamental principle when it comes to understanding SOC work efficiency and tool building. We present an activity theory model to explain the burnout and tool building in SOCs, yielding insights that were not obtained in our prior work.

There have been prior efforts in studying security operations mainly focused on tool development. Jaferian *et al.* [16] used activity theory to model challenges in reviewing access control policies in organizations. They design a tool that enables easy decision making for access control. Others used interviews and focused on providing guidelines for developing operational tools [5, 15, 27]. There have also been research efforts focused on understanding human, organizational, social, and other factors such as communication in the context of security operations [4, 30, 31, 32]. The main limiting factors of these prior works is the limited time spent in SOCs. From our own experience, it takes time to gain the trust of analysts and their management which is key to understanding the real problems causing inefficiency in security operations. We earned the trust of analysts by working alongside with them. We spent between 6 months and a few years in each of the SOCs, enabling us to understand problems as they evolved over a longer period of time. We believe that the insights we obtained are much deeper than if we had used short term methods such as interviews and questionnaires.

10. CONCLUSION

In this paper we present an activity theory (AT) model to explain the inefficiency in security operation centers (SOCs) and how tool building can help bring useful innovations. We analyzed field notes from our 3.5 year long anthropological study of four academic and corporate SOCs. The analysis revealed a number of primary and secondary contradictions in operational environments that manifest as conflicts. A concrete list of contradictions is presented by modeling SOC operations within the AT framework. Success or failure of technology solutions to improve SOC efficiency depends on acknowledging and mitigating these contradictions. By studying the resolved conflicts we understand why our tools were adopted into operations and became successful. With the reason in hand it becomes possible to reproduce it to solve similar other problems for SOCs. We further found that for a tool to be useful and usable in an operations floor it has to constantly resolve new conflicts that emerge. We leverage the hierarchical structure of human activities proposed in AT and extend it to a Pentagon Model for knowledge generation and transformation in SOCs. This model can be used by SOC managers and developers to identify tasks that could be automated periodically, resolving contradictions and improving SOC efficiency. Finally, the framework presented in this paper can be used to not just build tools, but for other positive changes that improve analysts' efficiency in general.

11. ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for the constructive comments that helped us in revising the paper, and Allison Woodruff for shepherding our paper. We thank the four SOCs who opened their doors for this research, and the analysts who worked together with us. This research is supported by the U.S. National Science Foundation under Grant No. 1314925. It is also partially supported by the Department of Homeland Security Science and Technology Directorate through DHS S&T/HSARPA/CSD BAA 11-02 and the Air Force Research Laboratory, Information Directorate under agreement number FA8750-12-2-0258. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

12. REFERENCES

[1] S. Barab, M. Barnett, L. Yamagata-Lynch, K. Squire, and T. Keating. Using activity theory to understand the contradictions characterizing a technology-rich introductory astronomy course. *Mind, Culture, and Activity*, 9(2):76–107, 2002.

[2] F. Blin. Call and the development of learner autonomy: Towards an activity-theoretical perspective. *ReCALL*, 16(02):377–395, 2004.

[3] C. Bonneau. Contradictions and their concrete manifestations: an activity-theoretical analysis of the intra-organizational co-configuration of open source software. In *Proceedings from EGOS Colloquium, Sub-theme*, volume 50, 2013.

[4] D. Botta, K. Muldner, K. Hawkey, and K. Beznosov. Toward understanding distributed cognition in IT security management: The role of cues and norms. *Cognition, Technology & Work*, 13(2):121–134, 2011.

[5] D. Botta, R. Werlinger, A. Gagné, K. Beznosov, L. Iverson, S. Fels, and B. Fisher. Towards understanding IT security professionals and their tools. In *Proceedings of the 3rd symposium on Usable privacy and security*, pages 100–111. ACM, 2007.

[6] J. Buell. COWs in the classroom: Technology introduction and teacher change through the lens of activity theory. *Unpublished manuscript. University of Illinois at Urbana-Champaign*, 2003.

[7] J. Buell. Learning to teach with laptops: A case study of teacher change. In *Society for Information Technology & Teacher Education International Conference*, pages 1984–1985, 2004.

[8] P. Capper and B. Williams. Cultural historical activity theory (CHAT). In *American Evaluation Association Conference*, pages CHAT–1 – CHAT–23. American Evaluation Association, November 2004. From a workbook entitled *Enhancing evaluation using systems concepts*. Available as of March 1st 2016 at `http://www.bobwilliams.co.nz/Systems_Resources_files/activity.pdf`.

[9] Y. Engeström. *Learning by Expanding: An Activity-Theoretical Approach to Developmental Research*. Orienta-Konsultit Oy, 1987. A second edition was published by Cambridge University Press in 2014.

[10] Y. Engeström. Expansive learning at work: Toward an activity theoretical reconceptualization. *Journal of Education and Work*, 14(1):133–156, 2001.

[11] Y. Engeström, K. Brown, L. C. Christopher, and J. Gregory. Coordination, cooperation, and communication in the courts: Expansive transitions in legal work. In M. Cole, Y. Engeström, and O. A. Vasquez, editors, *Mind, Culture, and Activity. Seminal Papers from the Laboratory of Comparative Human Cognition*, chapter 28, pages 369–388. Cambridge University Press, Oct. 1997.

[12] Y. Engeström, R. Miettinen, and R.-L. Punamäki. *Perspectives on activity theory*. Cambridge University Press, 1999.

[13] J. M. Frambach, E. W. Driessen, and C. P. M. van der Vleuten. Using activity theory to study cultural complexity in medical education. *Perspectives on Medical Education*, 3(3):190–203, June 2014. On line at `http://doi.org/10.1007/s40037-014-0114-3`.

[14] C. Geertz. From the native's point of view: On the nature of anthropological understanding. *Bulletin of the American Academy of Arts and Sciences*, 28(1):26–45, 1974.

[15] P. Jaferian, D. Botta, F. Raja, K. Hawkey, and K. Beznosov. Guidelines for designing IT security management tools. In *Proceedings of the 2nd ACM Symposium on Computer Human Interaction for Management of Information Technology*, page 7. ACM, 2008.

[16] P. Jaferian, H. Rashtian, and K. Beznosov. To authorize or not authorize: Helping users review access policies in organizations. In *Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 301–320, Menlo Park, CA, July 2014. USENIX Association.

[17] V. Kaptelinin and B. Nardi. Activity theory in HCI: Fundamentals and reflections. *Synthesis Lectures Human-Centered Informatics*, 5(1):1–105, 2012.

[18] N. King. Template analysis. In G. Symon and C. Cassell, editors, *Qualitative Methods and Analysis in Organisational Research: A Practical Guide*. Sage Publications Ltd, London, 1998.

[19] K. Kuutti. Activity theory as a potential framework for human-computer interaction research. *Context and consciousness: Activity theory and human-computer interaction*, pages 17–44, 1996.

[20] A. N. Leont'ev. The problem of activity in psychology. *Soviet psychology*, 13(2):4–33, 1974.

[21] A. N. Leontjev. *Problems of the development of the mind*. Progress, Moscow, 1981.

[22] C. P. Lim and D. Hang. An activity theory approach to research of ICT integration in Singapore schools. *Computers & Education*, 41(1):49–63, 2003.

[23] C. P. Nelson. *Contradictions in learning to write in a second language classroom: Insights from radical constructivism, activity theory, and complexity theory*. PhD thesis, The University of Texas at Austin, Austin, TX, 2002.

[24] D. L. Russell and A. Schneiderheinze. Understanding innovation in education using activity theory. *Educational Technology & Society*, 8(1):38–53, 2005.

[25] S. C. Sundaramurthy, A. G. Bardas, J. Case, X. Ou, M. Wesch, J. McHugh, and S. R. Rajagopalan. A human capital model for mitigating security analyst burnout. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 347–359, 2015.

[26] S. C. Sundaramurthy, J. McHugh, X. Ou, S. R. Rajagopalan, and M. Wesch. An anthropological approach to studying CSIRTs. *IEEE Security and Privacy Magazine*, Sept/Oct 2014.

[27] N. F. Velasquez and S. P. Weisband. Work practices of system administrators: Implications for tool design. In *Proceedings of the 2nd ACM Symposium on Computer Human Interaction for Management of Information Technology*, page 1. ACM, 2008.

[28] L. S. Vygotsky. *Mind in society: The development of higher psychological processes*. Harvard university press, 1980.

[29] E. Wardle. Can cross-disciplinary links help us teach 'academic discourse' in FYC? *Across the Disciplines*,

1, July 2004. Found as of March 1st 2016 at `http://wac.colostate.edu/atd/articles/wardle2004/index.cfm`.

[30] R. Werlinger, K. Hawkey, and K. Beznosov. Security practitioners in context: Their activities and interactions. In *CHI'08 Extended Abstracts on Human Factors in Computing Systems*, pages 3789–3794. ACM, 2008.

[31] R. Werlinger, K. Hawkey, and K. Beznosov. An integrated view of human, organizational, and technological challenges of IT security management. *Information Management & Computer Security*, 17(1):4–19, 2009.

[32] R. Werlinger, K. Muldner, K. Hawkey, and K. Beznosov. Preparation, detection, and analysis: the diagnostic work of IT security incident response. *Information Management & Computer Security*, 18(1):26–42, 2010.

APPENDIX

A. SNAPSHOT OF TEMPLATE ANALYSIS

Table 1: Snapshot of Initial Template after Coding a Subset of Data

Theme	Sub-themes	Examples
Primary contradiction	Subject	Metrics define the job.
Secondary contradiction	Subject - Rules	Hinders creativity. Unreasonable.
	Subject - Instrument	Poor attribution. Lack of customization. Lack of analyst perspective. Wrong assumptions. Long tuning process. Lack of visibility into tool functionality. High learning curve. Poor documentation.
	Subject - Community	Misaligned priorities. Pushback.
	Division of labor - Object	Inflexible role assignments. Lack of peer visibility.

B. HIERARCHICAL NATURE OF ACTIVITY

According to AT, human activity can be organized into a hierarchy of levels. This idea is often illustrated using a classical example from Leont'ev [21]. He differentiates between two different types of objects that come into play when people are engaged in socially distributed activities. Usually there is a *motivating object* that inspires the people to perform a particular activity and there is a *directing object* that is more immediate and guides them towards the motivating object. He explains this distinction using the example of hunting. When hunting together, people are divided into two groups: one that scares the animals by beating the bushes. These are called *the beaters*. The other group, called *the ambushers* (or *shooters* in current terminology) waits for the scared animals to come towards them so they can kill them. The original motivating object for the collective activity was food. An outsider positioned to examine only the activities of one group would find them difficult to fathom. The game is often well in advance of the beaters and might not be visible to an observer following them. The ambushers appear to be waiting idly, as they must be in position before the beaters start their drive. It is only when the observer discerns the relationship between the two groups that the hunt becomes apparent.

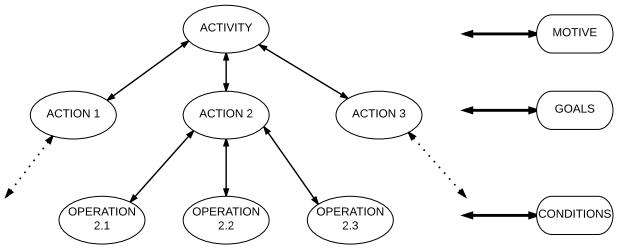


Figure 4: Activity Hierarchy

Figure 4 shows three levels in the hierarchy of human activity [17]. This abstraction can be adapted to fit any context. At the top level is the *activity* itself which is guided by the *motive*. The activity is broken down into sub-units called *actions*. The actions are motivated by *goals* that, seen in isolation, may appear to have nothing to do with the overall motive of the activity *e.g.*, the action of beaters may appear to have nothing to do with the overall motive of hunting. Each action is then decomposed into further smaller units called *operations*. Operations are in fact actions that have been customized to the environment under which they are carried out. The distinction between an action and an operation is that one may be aware of the fact that they are performing an action while an operation is a subconscious routinized task.

Productive Security: A scalable methodology for analysing employee security behaviours

Adam Beautement, Ingolf Becker, Simon Parkin, Kat Krol and M. Angela Sasse
University College London
{a.beautement, i.becker, s.parkin, k.krol, a.sasse}@cs.ucl.ac.uk

ABSTRACT

Organisational security policies are often written without sufficiently taking in to account the goals and capabilities of the employees that must follow them. Effective security management requires that security managers are able to assess the effectiveness of their policies, including their impact on employee behaviour. We present a methodology for gathering large scale data sets on employee behaviour and attitudes via scenario-based surveys. The survey questions are grounded in rich data drawn from interviews, and probe perceptions of security measures and their impact. Here we study employees of a large multinational company, demonstrating that our approach is capable of determining important differences between various population groups. We also report that our work has been used to set policy within the partner organisation, illustrating the real-world impact of our research.

1. INTRODUCTION

In order to express their preferences and requirements, security managers in organisations typically declare a centrally-managed security policy. This is then applied to all IT systems and individuals operating within the domain of the organisation. These policies are informed by the expertise, recommendations and regulatory requirements of the practitioner community, but ultimately must also fit to the working practices of the business itself. Effective security management must therefore tailor policies to both the operational and organisational contexts. For example, a commercial organisation aiming to maximise business opportunities will have a different security profile to a military organisation with a strong preference for confidentiality over availability. Likewise, policies must take in to account not only that the daily working life of an employee is not just about security [15], but also that employee populations are not homogeneous. A policy that is effective in theory may not translate into secure behaviour in practice, if it is not aligned with the productive processes of the organisation and the goals and capabilities of the employees to whom the policy applies. This creates a need for security managers to empirically assess and compare the policies under their control, in order to determine how well they meet these goals [23]. It is toward this end that the Productive Security project has worked for the last four years.

While technically-focused sources of data – such as system logs – are commonly used to support analysis of policies, they do not provide an insight into employees' thought processes. Security systems are not just the sum of their technical components – user co-operation plays a critical role in providing organisational security, which highlights the need to consider the relationships between people, process, and technology [14]. In addition, an over-reliance on technical solutions can hinder an organisation's capacity to support employees in their productive tasks [30]. Behavioural data is therefore an important factor for effective security management, and a goal of this work has been to create a set of repeatable metrics capable of assessing employee attitudes and behaviour around security. In particular, we develop a methodology capable of identifying areas in which the security policy itself creates incentives for negative behaviour. Rigid systems can force compliance with policy but promote disgruntlement [6]. Where conflict exists between security systems and productive tasks, friction results. Workarounds and 'circumvention strategies' [1] are then likely to develop as users take advantage of system flexibility to modify how technology and procedures work. This reduces security effort but often introduces security vulnerabilities as a side–effect (e.g., using the same password for a number of accounts across both work and personal life). Managers may even be complicit in supporting workarounds if secondary tasks (such as security) stand in the way of business continuity [26]. Different threat models exist within different areas of life, so the vulnerabilities in one space can weaken security in others (e.g., carrying unencrypted USB data devices in transit between work and home) [5].

Balancing the demands of primary, productive tasks and secondary tasks – such as security – introduces cost-benefit dilemmas in which individuals are forced to choose between security and productivity. In particular, security that overburdens the user and is not aligned with their working practices can become less effective [6]. Security is presented to employees as being for their own good, but can introduce externalities, burdening the individual with indirect costs (e.g., changing an increasing number of passwords at regular intervals) [16]. Individuals may, rationally, perceive the personal cost of compliance as greater than the security be-

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22–24, 2016, Denver, Colorado.

nefits gained. As a result, detecting instances where security and business processes are in conflict is critical for both an organisation's security and productivity.

Our methodology is a multistage process designed to elicit realistic responses from employee populations at scale. As it is necessary for our data to support the decision–making process of different organisations of all sizes, both of these points are of great importance. Data that does not closely represent the operational reality of the organisation cannot be used to drive decision–making, as it is not a reliable predictor of future states and outcomes. Likewise a data collection method that is overly time-consuming or does not scale (potentially up to tens of thousands of participants) quickly becomes impractical for larger organisations. Both of these concerns are addressed by the Productive Security (ProdSec) methodology.

A programme of research as heavily based in the operational context of organisations as this could not be undertaken without the partnership of organisations representative of its target audience. Our research project has been fortunate to have such partners and this work could not have taken place without their support and cooperation.

There follows a review of the related literaturesection 2. In section 3, we introduce the ProdSec methodology, which leads us to the results of our study in section 4. Discussion and Conclusion follow in sections 5 and 6.

2. RELATED LITERATURE

A number of existing works use surveys and/or interviews to explore the relationship between an organisation's information security policy and employee behaviour. These works examine the impact of attitudes and perceptions on behaviour and consider both intrinsic and extrinsic influencing factors. For example, Pahnila et al. [18] found that attitudes towards security and the habits of individuals can have a significant effect upon the intention to comply with security policies. They also assert that the social environment around an individual will have an effect upon their propensity to comply with policy.

Expanding on intrinsic motivators, Rhee et al. [25] use social cognitive theory to model the influence of experience with security incidents upon self-efficacy, and the role of self-determination upon the outcome of security-related scenarios. A large-scale survey completed by ˜400 students found that individuals with high self-efficacy used more security tools and were more vigilant to security, and experience of security compromises negatively impacts self-efficacy.

The notion of competence was also investigated by Workman et al. [33], who explored the "knowing-doing" gap in individuals who have appropriate security skills and knowledge, but who do not apply these skills consistently. Based on the results of a survey in which 588 members of a technology services company participated, the paper concludes that security technology should be user-centred to avoid a tension between assessing threats and use of coping responses.

Siponen et al. [27] utilise Protection Motivation Theory (PMT) to reason about employee compliance with information security policies. The work considers component parts of PMT, namely threat appraisal and coping appraisal (where this includes response costs). A survey was conducted with 917 employees of Finnish companies. Amongst the findings, threat appraisal was found to have a significant impact on intention to comply with information security policies. Employee beliefs about their ability to adhere to policy influence their intention to comply. This finding stresses the importance of perception; the authors assert that if policies are not perceived as relevant by an employee, adherence to policy will be diminished.

Perception was also the focus of work by Bulgurcu et al. [11], who infer that the perceived costs and benefits of compliance (or non-compliance) are formed by the perceived consequences. The authors find that intention to comply is heavily influenced by attitude, beliefs and ability to comply. The relationships between these factors are explored using a survey of 464 employees across a number of organisations. The study identified three belief classes relating to consequences of compliance decisions – benefit of compliance, cost of compliance, and cost of non-compliance.

The prevalence of attitude and perception as themes throughout these works strongly influenced our survey design. However, these surveys all rely on some sort of rating (e.g., Likert) scale, or a sliding scale (e.g., keeping information safe is beyond, or within, a person's control). We build on these themes but opt to take a more immersive scenario-based approach.

A related approach is taken by Albrechtsen and Hovden [4], utilising the differences in skills, perceptions, and interpersonal relationships to characterise the 'digital divide' between information security managers and end-users. The researchers analysed interviews with 11 managers and 18 employees alongside complementary web-based surveys exploring how 87 managers and 151 users assess security threats and vulnerabilities. The study acknowledges that users prioritise other work tasks, that policy is potentially impenetrable and hard to find for the non-expert, and that security provisioning is often one-way. We extend this approach by grounding survey questions in interview outcomes, towards greater resonance with real-world user experiences.

Other methods of constructing scenario content have been attempted. Both D'Arcy et al. [13] and Parsons et al. [19] generate survey questions by drawing on existing literature and interviews with experts. While this makes good use of general information, it does not allow for surveys to be tailored to the specific context of deployment. Darcy et al. use their survey to explore links between stressful information security demands and intentional violation of security policies, to identify workplace factors which contribute to policy violation, including overload, complexity, and uncertainty. Stressful conditions contribute to security coping strategies, as behaviours are adapted in response to stress factors, which than have a knock-on effect on productivity. Where security requirements are perceived as overloading, complex and uncertain, users then become disengaged, implying that high-effort policies can themselves promote insecure behaviour. The inclusion of productivity as a consideration is of particular interest here, mirroring our goal of 'Productive Security'.

Counter to D'Arcy et al. [13], Guo et al. [15] propose a model of what is referred to as 'Non-Malicious Security Violation (NMSV)', validated by a survey, delivered in both paper

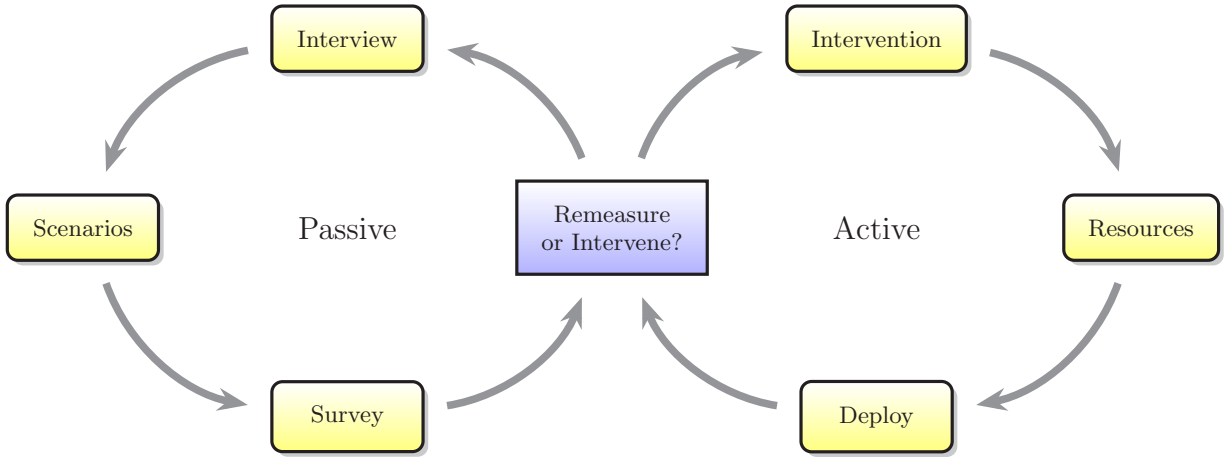


Figure 1: Overview of processes in our methodology

and web formats, of employees and their working conditions. The authors look beyond visible behaviours and instead examine the role of attitudes toward security policy violations, such as the productivity advantage of non-compliance, perceived risks, and workplace norms. As in D'Arcy et al.'s work, scenarios are developed based on related literature and interviews with security practitioners and experts, where end-users of policy are not directly engaged. Results imply that job performance advantages, perceived security risk and workgroup norms are key predictors of intention to engage in NMSVs, with users favouring business tasks. The study also found that attitudes toward security policy itself were not significant in driving non-compliant behaviour, in contrast to Bulgurcu et al. [11]. The authors recommend a user-centred security management strategy, where employees can satisfy productivity goals while also maintaining security.

A scenario-based approach is also taken by Blythe et al. [9], who study how individual and organisational factors in the workplace impact secure behaviours, using interviews based on 16 'vignettes'. These cover security behaviours identified from information security policies. Vignettes were effectively used as a device for building a rapport with participants and eliciting attitudes and beliefs relating to a specific subject, an approach reflected here in our interview technique. Results suggest that research should focus on individual security behaviours rather than beginning and ending with policy compliance, and that participants accepted responsibility for some elements of security, while leaving others to their organisation.

Building on the existing literature, the work presented here uses an immersive scenario-based survey, moving away from severity-based questionnaires to situate surveys in the environment in which they are deployed. Scenarios are derived from the results of a semi-structured interview process, based on common areas of friction. We discuss this methodology in more detail in the following section.

3. METHODOLOGY

The goal of the ProdSec methodology is to provide researchers studying organisations with a repeatable, scalable data gathering process that allows them to better understand the security-related issues facing their employees, and the behaviours and attitudes they adopt in response.

The full ProdSec methodology consists of two independent, iterative processes. Figure 1 illustrates the steps involved. The two cycles represent a *passive* data collection and monitoring phase on the left, and an *active* intervention phase on the right. This paper will focus on the passive cycle, although an overview of the active cycle is presented here for contextual clarity.

The passive cycle identifies the predominant security behaviours and attitudes within an organisation, along with specific points of friction between the business and security processes. In order to support real-world decision-making, data collection must both accurately represent the real-world environment where it is applied, and be sufficiently scalable so as to be of use to potentially very large, multi-national organisations. These two goals are to some degree in conflict. Rich, in-depth data capable of accurately representing a real-world context can require a greater investment of time and effort to collect, making it problematic at scale. ProdSec tackles this by utilising a two-stage method.

Firstly, semi-structured *Interviews* are conducted with a vertical cross-section of the organisation to capture attitudes and behaviours across as many roles, physical locations and demographic groups as possible. We discuss this aspect in section 3.1. Based on interview findings, we carefully craft a scenario-based survey that reflects dominant security-related issues. By tailoring our survey to each operational context, we ensure that survey questions are relevant and recognisable to participants, with the aim of eliciting more realistic and genuine responses.

Once this cycle is complete, security practitioners then have the choice of monitoring the situation over time by repeating the measurement cycle at some future interval (e.g.,$\approx$ 6 months later), or actively engaging with any uncovered problems. The right half of figure 1 describes the active *Intervention* phase. Based on the conclusions drawn from the passive cycle we work with the organisation to prioritise the issues identified, and design and deploy optimal intervention(s), taking into account business as well as socio-technical factors section 3.8. Direct collaboration is important as interventions also need to be suitably centred around the human.

Visualising the methodology as cyclical is essential for understanding its intent, reflecting the notion that security is a process and not a fixed state. It is our experience that organisations often see the implementation of an intervention as the final step, whereas we consciously position this step as part of an ongoing process. The passive cycle can therefore be used to track changes in attitude and behaviour over time, as a consequence of the organisation's evolution or in response to specific interventions. Ongoing monitoring can inform decision-makers as to whether interventions are having the desired effect, and indeed whether interventions have themselves influenced security processes. Likewise, no one set of interventions will provide a 'silver bullet' solution, requiring repetition of the active cycle.

The analysis in this paper focuses on the data collection and analysis stages of the passive cycle. As this research spans three years, with two main phases of data collection, lessons have been learnt along the way. As such, we refined our methodology between the two rounds of data collections, both at multinational companies. In the interests of clarity, the methods given below are those used in the second, or most up to date, version of the methodology. Where lessons learnt between the two phases are relevant and of interest, they have been included in the discussion.

Sections 3.2 and 3.3 describe the methodology from semi-structured interviews to scenarios and the subsequent design of the scenarios. As a last step in the passive cycle we study the responses of employees to the scenarios, by way of the survey. We discuss the analytical approach in section 3.7 and boxout 4. The results from the survey and related analysis are described in section 4.

3.1 Semi-structured interviews

Interviews are resource-intensive to both conduct and analyse, which limits their use on a large scale. Interviews do however have several advantages, most notably that they provide the rich contextual information needed to ensure that more scalable forms of data gathering, such as surveys, can capture realistic and relevant data. The interactive nature of interviews also allows the researcher to probe more deeply on topics of interest during the data gathering process.

Although we have provided a set of questions in Appendix C, it should be noted that semi-structured interviews are not restricted to a fixed set of questions. In our work, the researcher attempts to build a rapport with the interviewee and guide the discussion through topics of interest, from the perspective of the individual. Each interview would last an hour and cover a range of security-related topics including: security awareness, data sharing, password management, laptops and removable media, remote working, clear desk policy, physical security, reporting and training.

We aim for $\approx$ 100 interviews to be conducted at each organisation, making it impossible for a single researcher to conduct them all. We recognise that this introduces problems of consistency – different interview techniques yield different levels of insight, engagement, and expression from participants. Interviewer training involving all interviewers focused on where and how to ask more follow-up questions through careful "probing". We also attempted to standardise topic areas and question phrasing. This was adopted explicitly in our second round of data collection, having learnt from the first round that while we were able to identify where problems existed within the organisation, we did not have a good grasp on how frequently those problems occurred, or how widespread they were. The interview topics and probing questions can be found in section C.

Participant recruitment was managed in conjunction with the partner organisation. Our goal was to speak with a vertical cross-section of the organisation. Security problems are not confined to any one employee group; to develop a full understanding of the current organisational security culture, interviews should ideally be conducted within a variety of departments and with employees across a range of roles. This was not always possible due to internal pressures within the partner organisations. For example, in the company for which results are presented here, we were only able to interview within their Operations division as other divisions had been involved in a different survey process close to that time and there were concerns about data collection fatigue within the organisation.

In each case, we ask volunteers to take part in an interview, incentivising participation with – in the case of the survey presented here – a raffle prize. Making the process voluntary rather than mandatory carries both advantages and disadvantages. The success of semi-structured interviews is heavily dependent on the level of rapport the interviewer can develop with the participant – a participant that opens up to the interviewer is likely to give more honest and detailed responses. This is particularly true in the case of security interviews, given that discussion can potentially touch upon self-reporting of transgressions, rule-breaking and circumventions. A good rapport is then necessary to build the trust that is necessary between interviewer and participant.

The interview study was promoted in an item in the company newsletter, asking for employees to talk about security issues. Volunteers responding to the item may well have an agenda of their own. If an individual is keen to talk to the interviewer or otherwise be open about their views on security, this in itself may not represent a balanced view of the security culture in the organisation. It is then difficult to ascertain if their view is typical of the wider population or not. We favoured using volunteers as our methodology includes a survey stage, in part to ascertain the prevalence of any problems identified during the interview stage. We regarded it as preferable to ensure responsive and in-depth interviews, and aim for more representative participation at the survey stage.

3.2 Interview analysis

For interview findings to be of use throughout the data collection methodology, key insights must be extracted from the interviews and made available in a more usable form distinct from the original transcripts. This was achieved by a process of thematic analysis [10] conducted by three researchers. As with the interviews, we recognised the importance of consistency. To this end, we collaboratively developed a codebook that allowed us to systematically apply codes across the interviews. Initial coding was conducted by each interviewer on separate transcripts, with codes being added as necessary in order to capture new concepts as they arose. Regular coding meetings were then scheduled to merge and

Boxout 1: Expert- and Employee-driven Scenarios

The strict demand for the methodology to base survey questions on what emerges from the interviews derives from our experience with the first round of data collection with another company. Here, we included scenarios in the survey which were based on what in-house security experts asserted was happening. However, the results from these scenarios were less coherent and tended to be more polarised than those developed strictly from interview data, implying that they were less recognisable to participants. As the aim of the work is to improve both security and productivity, our survey with the second organisation was built solely on how security is perceived by employees, relative to their primary tasks.

prune these code sets with a view to avoiding instances of duplicate or very similar codes. Codes were also grouped into code families at this time, covering the major topic areas observed within the interview transcripts. After several iterations of this process, we arrived at a largely stable code set of approximately 120 codes that was flexible enough to accommodate the range of topics found in the interviews.

Once this coding process was complete we were then able to tally up the codes to identify the most common security-related issues. These then formed the foundation for a more scalable data collection method that was nevertheless grounded in real-world situations. It is important to note that this process is necessarily bespoke for each context. Interviews conducted in each organisation will reveal different problems, cultures, and technologies. Although we did find problems common to both organisations we assessed (see boxout 2) there were still many differences between how these issues were expressed and the contributing factors that surrounded them.

3.3 Online scenario-based survey

Our approach to scalable data collection was driven by an online survey. In order to efficiently reach large numbers of people, it was necessary to allow them to take part in the data collection exercise from any location, in particular from their usual work environment. Not only does this increase the response rate by minimising demands on participation, but it also furthers our aim of making data collection as naturalistic as possible, as the collection environment matches the operational environment being assessed. As our primary method of recruitment and communication with participants was through company newsletter emails, we embedded a link to the survey in an issue of the newsletter.

As described in section 2, many surveys present short questions with either a multiple choice or Likert scale-style answers. It is our view that this approach is unlikely to engage participants, in particular due to a widespread fatigue with questions of this style. Participants are likely to skim through the survey and apply little thought to their answers. Also, short questions would not allow us to utilise the full value of the interview information. As such, we elected to build scenario-based survey questions, in which participants were presented with one of the common situations identified via our interview analysis.

Boxout 2: Scenario Commonality

Although the companies that we worked with operate in different sectors there was some overlap in the issues that arose as the result of our interview analysis. Clear desk policies, tailgating through physical security and file sharing were present as issues in both environments. This suggests that as more companies are assessed, a database of scenarios could be developed that over time reduces or minimises the need for the interview stage. However, despite the similarities in topic area it was still necessary to alter key details in the text accompanying each scenario, in order to present to each participant set a scenario that approximated the reality of their environment. For example, one company observed tailgating through security doors, the other through turnstiles in their foyer. Accurately representing these details increases the realism of the scenario, with an aim to encourage honest responses.

Once a topic was selected the scenario was written, using organisation-specific details and terminology from the interviews (see boxout 2). For each scenario, we also created four possible answers or outcomes, again drawing on the interview data to craft these so they appeared familiar and plausible to the participants. How these options were then used will be covered in more detail in section 3.6.1. Our goals here were to:

- Present scenarios to the participants that seemed both realistic and familiar,
- Offer answer options that were likewise realistic and familiar,
- Gather as much implicit data as possible to maximise the benefit of the survey while minimising the time taken per participant.

A potential problem with survey – especially one covering a sensitive topic such as security – is that participants attempt to give the answers they feel are expected of them, or are correct, rather than an honest reflection of their own thoughts and views. These deviations fall into two main categories, response bias and demand characteristics. We addressed these in different ways.

Response bias refers to biases introduced by the participant being influenced by sensory inputs and cognitive processes when answering the question, and thus unintentionally altering their response. For example, how a question is phrased can alter the response (in particular if it is a leading question). Aside from eliminating instances of leading or priming language we also took care to phrase our attitude scenarios from the point of view of a fictitious colleague, as participants are often more comfortable reporting on the actions of others. So rather than asking, 'what would you do in this situation?', we asked, 'what would Jessica do?', intending this to counter some aspects of the response bias. This

helps us obtain an accurate representation of the maturity levels of the employees.

Demand characteristics refer to the fact that research participants might speculate about the purpose of the research and give responses that they think align with what the researchers are trying to find out. This is something we were particularly concerned about, as security is a sensitive topic with potentially significant outcomes. As we were essentially asking people to report on their own rule breaking, the potential for participants trying to give 'right' answers was high. To tackle this we made sure that in each scenario there were no obviously correct options — each possible answer involved some difficulty or transgression.

We split the scenarios into two types, based on our interview analysis. When participants reported incidents it was either in the form of something they did themselves, or something they observed their colleagues doing. Our scenarios followed the same approach, and were divided into Behaviour and Attitude scenarios.

3.4 Behaviour-type scenarios

These scenarios present the actor with a situation that puts the requirements of the primary business process in conflict with some aspect of the security policy. Typically this involves the actor in the scenario needing to complete a specific task, the completion of which is being slowed down or prevented by a security process or mechanism. Four options were then given that presented courses of action that would resolve this conflict. Each of these options contained an element of non-compliance so as to avoid participants seeking to give the 'right' answer.

In pursuit of our goal to capture rich behavioural data, we linked each answer option as closely as possible to one of four behavioural risk types. This meant answer choices also allowed us to monitor the prevalent behaviour types. Crossler et al. [12] posit that cultural theory can be used as a predictor of the impact the norms of an organisation can have upon perceptions of security-related risks. The behavioural risk categories used in our surveys stem from Adams [2], and their characteristics are given below:

Individualists rely on themselves for solutions to problems,
Egalitarians rely on social or group solutions to problems,
Hierarchists rely on existing systems or technologies for solutions to problems,
Fatalists take a 'naive' approach to solving problems, feeling that their actions are not significant in creating outcomes.

Individualists may for instance feel less loyalty to others in the organisation and to policy, but may be more likely to report what they see as inappropriate behaviour to others [12]. The topics covered in the Behaviour-type scenarios were:

- Password manager,
- Use of the company VPN for remote working,
- File storage, and
- Conducting a credit check in a retail location.

Full texts for these scenarios can be found in appendix A.

3.5 Attitude-type scenarios

To further explore the security culture in an organisation, we complement the behaviour-type scenarios with attitude-type scenarios. Rather than presenting participants with a task, the actors described in the attitude-type scenarios observe an instance of non-compliance in their environment – such as finding a screen unlocked – and respondents are asked to indicate how they would react. The four options in this case represent distinct responses, such as to confront the transgressor, or dismiss the incident as commonplace. As with the behaviour-type scenarios, each response contained an element of non-compliance or an implicit cost. While it may seem like confronting a transgressor is an obvious right answer, there is in fact a high social cost associated with doing so (we find for instance that typically security-conscious individuals are regarded as paranoid by their peers).

The answers here were linked not to behavioural risk types but to a model of cultural maturity that has been developed in support of this work. The model considers security competence relative to an individual's business tasks. Other works describe the need for competence in repeatable tasks which can form good security habits [31]. Here, we also consider the capacity to embody policy (where it is clear) and adapt it to new or complex situations that require a conscious response, a distinction that has been explored by Reason in the realm of safety [22].

Our model contains a series of levels (see section B) which attempt to articulate the maturing relationship between the individual and security policy. Those at the lower levels engage with security only as absolutely necessary, while those at the higher levels champion security in their local environment. The levels linked to the answers in the survey are as follows:

Level 1: Is not engaged with security in any capacity.
Level 2: Follows security policy only when forced to do so by external controls.
Level 3: Understands that a policy exists and follows it by rote.
Level 4: Has internalised the intent of the policy and adopts good security practises even when not specifically required to.
Level 5: Champions security to others and challenges breaches in their environment.

Although the model includes a level 1, practically speaking individuals at this level will not be found in an organisational environment, as there is typically infrastructure in place that at the least requires employees to have a registered username and password to facilitate access to IT resources. As such our survey utilises level 2 and upwards. Level 2 assumes that compliant behaviour must be imposed upon individuals to ensure that routine tasks remain secure, and so in turn the IT infrastructure constrains behaviour. This is analogous to 'basic hygiene' as described by Stanton et al. [29], acting to manage the 'dangerous tinkering' and 'naive mistakes' which might otherwise happen. However, organisational security is complex and technology-based solutions alone cannot anticipate and manage all situations. Level 3 assumes that employees have enough security knowledge to make some in-situ decisions, whereas toward level 5 employees know enough to apply their knowledge and skills to unforeseen situations,

Boxout 3: Assessing Non-Compliance

The first round of data collection through interviews gave us insights into why participants decided to break policy in order to complete a business talk. However we were unable to establish the prevalence of this behaviour, leading us to augment the survey. We added in an additional question to the business type scenarios that asked participants if they found it acceptable to prioritise the business process in this way.

as well as to articulate workable solutions to those around them.

The topics covered by the attitude-type scenarios were:

- ID badges,
- Clear desk policy,
- Tailgating through physical barriers, and
- Secure disposal of confidential hardcopy.

Full texts for these scenarios can be found in appendix A.

3.6 Scenario tasks

For each scenario, a survey question would have two phases – participants would be asked to select their preferred option, and also complete a rating task.

In the case of the behaviour-type scenarios, the ranking task would involve asking participants to rank all four options in order of how likely they would be to take a particular course of action. For the attitude-type scenarios, this ranking would ask for participants to indicate how strongly they agreed with the response of the actor within the scenario.

With the ranking exercise complete, participants were asked to complete a rating task. Here a participant would be asked to rate on a Likert scale of 1 to 5 how severe a breach of security the behavioural-type options were to them, with 1 being *not severe at all* and 5 being *very severe*. In the case of the attitude-type scenarios, a participant would be asked to rate on a scale of 1 to 5 how acceptable the options were, with 1 being *not acceptable at all* and 5 being *very acceptable*. It should be noted that the survey software was set up to not allow participants to backtrack and change previous answers. This was an intentional design choice so participants could not adjust previous answers to align with subsequent ratings, allowing us to detect discrepancies between the two tasks as a way of highlighting areas significant friction between employees and security. We identify these areas by performing a statistical correlation between the rating and ranking scores for each scenario. We describe this analysis in boxout 4.

The behaviour-type scenarios had an additional question that preceded the ranking and rating tasks. The participant is asked to evaluate the severity of the scenario presented to him, allowing us to assess the participants willingness to trade of the completion of the business task and the preservation security.

Boxout 4: Hotspots

We refer to instances of ranking scores being positively correlated with severity rating scores, or negatively correlated with acceptability rating scores, as *'hotspots'*. Where these correlations are detected, it indicates that participants are favouring the use of options that they know carry high risk and which represent unacceptable forms of behaviour. Hotspots represent significant areas of concern for the organisation, as they show areas in which employees report that they have to choose (knowingly or unknowingly) insecure practices.

3.6.1 Scenario selection and distribution

For each organisation, 8-10 scenarios were created. However, for several reasons we did not wish for each participant to complete the entire set of scenario questions. First, it would be too time-consuming; our partner organisations were generally concerned about the productivity impact of large-scale data collection involving any number of employees over a short period, and so we wished for our scenario survey to be completed in 10-15 minutes. Second, as the scenarios were tailored to specific topics not all of them would be relevant to all parts of the business. For example, giving a question about a retail environment to an engineering division will yield data based on guesswork rather than experience. During deployment, we would request a range of demographic information – including business role (the options for which were drawn from the company's structure) – at the beginning of the survey, then deploying a subset of 3-4 scenarios to each participant based on their responses. This is an example of where it is important – and necessary – to engage with a partner organisation at the right level, to ensure that survey tools etc. can be managed in a way that fits naturally with activities within the business.

3.7 Survey tasks – scoring method

As discussed in sections 3.4 and 3.5, each of the options given with each scenario was linked to either a behaviour type or a maturity level. In order to determine the prevalence of each categorisation, a scoring method linked to the ranking task (see section 3.6) was used. The position of the option in the ranking task determined the score of the associated type. This score was cumulative over the scenario questions. For example, if the first option was linked to Behaviour Type then the ranking of this option determined the score given to Type 1. The scoring was as follows:

Rank 1: 4 points
Rank 2: 3 points
Rank 3: 2 points
Rank 4: 1 point

As each participant answered a maximum of two behaviour and two attitude questions, these scores were normalised for each participant, enabling statistical analysis. This scoring system was also used to determine the popularity of the scenario options themselves.

3.8 Selecting interventions

Organisations may enact 'interventions' in order to influence a change in the regular security behaviours of employees. The rich picture of employee behaviours and attitudes provided by the interview and survey process means that the methodology described in this paper can support a more systematic and informed approach to the identification of interventions. While this paper does not cover the outcomes of this step in detail (the right-hand side of Figure 1), the intended use of the data and survey results are included here in the interests of completeness.

The interview and survey results provide security managers with information on the most pressing problems – the 'hotspots' (see boxout 4) – encountered by employees in their own organisation's IT environment, as well as an idea of the factors that underpin an issue. Interventions can then be targeted at the motivating factors of an issue, rather than the symptoms or the elements of it which most relate to specific regulatory expectations. It is intended that researchers would engage with the organisation to determine which category of intervention is optimal, drawing on system (re)design, awareness and training, or technical controls where appropriate. Means of addressing tensions between security and productivity are further discussed in [6]. Having some sense of the scale of a policy hotspot is also useful (how many employees regularly enact an insecure behaviour), as organisations can then invest resources proportionally, and crucially consider the intended scale of an intervention to ensure that it is properly implemented and does not introduce problems of its own (for instance by updating only a subset of the awareness materials which employees are directed to use, which can in turn introduce inconsistencies).

3.9 Research ethics and data handling

The study successfully went through an ethics approval process at our institution (approval number: 3615/002) and was registered with the Data Protection Act (registration number: Z6364106/2012/11/08). We had a written agreement with management – which was distributed with the recruitment email – that employees would not face negative consequences for policy violations they reported. The audio-recordings were transcribed by an external company under NDA. Transcripts were redacted to remove any identifying information such as names of people and locations. The original audio recordings were deleted.

4. STUDY RESULTS

In this section, we present the analysis of the survey data collected at the second large company we studied. We focus on the analysis of maturity scores (section 4.2) and behaviour types (section 4.3) between the different groups of the organisation: business division (sections 4.2.1 and 4.3.1), age group (sections 4.2.2 and 4.3.2) and location (sections 4.2.3 and 4.3.3).

4.1 Response rate

In total, 641 complete survey responses were recorded. The briefing document informed participants that any surveys completed in less than 5 minutes (minimum reading time in our pilot study) would not be included, which left us with 608 responses for analysis.

For the purposes of our study, the organisation is split up across 7 business divisions as well as a number of locations. The majority of responses originated from the *Sales & Services* (292), followed by *Operations* (152). The remaining divisions were all significantly smaller, ranging from 11 to 47 responses. Participation was more equally divided between the business locations surveyed, with locations *1: HQ* and *5* (a large regional office) being the largest ones with 118 and 130 responses respectively. Further we analyse trends across 8 age groups. Survey respondents were approximately normally distributed across the age groups, with the age group $30-34$ representing the largest share with 124 participants. The edge cases of < 25 and ≥ 55 were nonetheless sufficiently large with 53 and 22 responses respectively, to allow for potentially statistically significant results across all age groups.

The number of responses were sufficient to allow a factor analysis by business division (sections 4.2.1 and 4.3.1), age group (sections 4.2.2 and 4.3.2) and location (sections 4.2.3 and 4.3.3), with 8 factors each. A full factor analysis with 512 factors is outside the scope of this methodology at present as it would a sample size several orders of magnitude larger.

4.2 Maturity levels analysis

Each participant responded to at least one maturity type scenario (section A), by ranking the four options presented in order of preference as well as assigning an acceptability score on a Likert scale to each option (see section 3.3). A comprehensive statistical analysis was then carried out on these responses, with the results in table 1. In total, three such tables have been produced, but only the first one is shown here for brevity. The remaining diagrams are included in the supplementary material (see section 6.2).

The last line of table 1a shows the full organisation's maturity level properties (please refer to the caption of table 1 for the details of the statistical analysis carried out). The ranking and acceptability score of each of the maturity levels are all statistically significantly separated and increasing with the maturity score. Level 5 has an average rank of 3.30 and acceptability score of 4.51. These ranks are high – a perfect score would represent an average maturity rank of 1, 2, 3 and 4 for levels 2 to 5 respectively. Further, there is a strong positive correlation between rank and acceptability score: the more acceptable the option, the more likely the participant is to choose it.

4.2.1 Maturity by division

Table 1a illustrates the relationship between maturity levels and business divisions. The data is shown in terms of variations from the organisation's mean in order to facilitate comparisons across the business divisions. There are a number of interesting deviations from the organisational mean. Only the *Sales & Services* division ranks maturity level 5 statistically significantly above level 4, where the *Finance & Prof. Services* division ranks maturity level 4 highest and statistically significantly higher than level 5. The participants from the other divisions did not discriminate between level 4 and 5 options. Participants from *Sales & Services* opted for responses corresponding to level 5 statistically significantly more often than any other division in the organisation with a mean level 5 rank of 3.64.

Business Division	Level 2		Level 3		Level 4		Level 5		τ
	Rank**	Accept**	Rank**	Accept**	Rank**	Accept**	Rank**	Accept**	
Business	0.33**	0.29*	−0.23*	−0.32*	0.10	0.33$^{*}_{**}$	−0.21$^{**}_{**}$	−0.14*	0.61**
Finance & Prof. Services	0.37**	0.24**	−0.12*	−0.23$_{*}$	0.13$^{*}_{*}$	0.37**	−0.38$^{**}_{**}$	−0.34$^{**}_{**}$	0.59**
Human Resources	0.23**	0.18	−0.21	−0.47	0.14	0.20$_{**}$	−0.16$_{**}$	−0.15	0.66**
Marketing & Consumer	0.43**	0.37**	0.12$_{*}$	−0.15	−0.10	0.29	−0.45$^{**}_{*}$	−0.42$^{**}_{**}$	0.50**
New Business	0.33*	0.32	−0.32*	−0.23	0.19	0.47$_{**}$	−0.21$_{*}$	0.04	0.55**
Operations	0.23*	0.14	0.25$^{**}_{**}$	−0.34$^{**}_{**}$	−0.13$^{**}_{**}$	−0.23$_{**}$	−0.34	0.06$^{**}_{**}$	0.45**
Other	0.30**	0.30**	−0.33**	−0.42**	0.25**	0.27*	−0.22$^{**}_{**}$	−0.38$^{**}_{**}$	0.65**
Sales & Service	−0.31**	−0.23**	−0.03$_{**}$	0.34$^{**}_{**}$	−0.00$_{**}$	−0.06$^{**}_{**}$	0.34$^{**}_{**}$	0.12$^{**}_{**}$	0.76**
mean	1.48	1.50	2.13$_{**}$	2.19$_{**}$	3.08$_{**}$	3.98$_{**}$	3.30$_{**}$	4.51$_{**}$	0.62**

(a) Maturity level rankings and acceptability score split by business division.

Business Division	Scenario	Individualist		Egalitarian		Hierarchist		Fatalist		τ
	Sev**	Rank**	Sev**	Rank**	Sev**	Rank**	Sev**	Rank**	Sev**	
Business	0.52**	0.17	−0.10$_{**}$	0.59**	−1.18**	−0.53*	0.29	−0.23	0.10	−0.22**
Finance & Prof. Services	0.38**	0.34$^{**}_{*}$	−0.24	0.50**	−0.76**	−0.67**	0.29	−0.16	0.15	−0.19**
Human Resources	0.62*	0.53*	−0.13	−0.09	−0.90**	−0.16	−0.41	−0.29	−0.73	0.08
Marketing & Consumer	0.84**	0.32	−0.74**	0.86**	−0.84**	−0.47*	0.34$_{*}$	−0.71**	0.18	−0.22**
New Business	0.58	0.59*	−0.58	0.62	−0.39	−0.80*	−0.11	−0.41	0.01	−0.39**
Operations	0.03**	−0.02$_{**}$	−0.33$^{**}_{**}$	−0.34**	0.38$^{*}_{**}$	−0.40$^{**}_{**}$	−0.39$^{**}_{**}$	0.76$^{**}_{**}$	−0.96**	−0.48**
Other	0.24	0.35**	−0.46*	0.04	−0.53**	−0.26	0.11	−0.13	−0.16	−0.28**
Sales & Service	−0.28**	−0.18$^{**}_{**}$	0.37$^{**}_{**}$	−0.06	0.25**	0.48$^{**}_{**}$	0.10*	−0.24$^{**}_{**}$	0.50**	−0.17**
mean	2.24	2.68$_{**}$	3.49	2.02	3.76$_{**}$	2.80$_{*}$	3.20	2.50$_{**}$	3.44$_{**}$	−0.20**

(b) Behaviour types rankings and behaviour severity score split by business division.

Table 1: The values in each cell of the tables above describe the variation from the mean in their column, with the mean being shown at the bottom (the mean is the value for the organisation as a whole). Based on the scoring in section 3.7, higher ranks imply more popular choices. Similarly, the higher the Accept/Sev score, the more acceptable/severe the participants take the option to be. In the second row, the $^{**/*}$ after Rank/Accept/Sev show statistical significant variations from the median rank or acceptability or severity score respectively based on the Kruskal-Wallis H-test for independent samples at $p < 0.01/p < 0.05$ confidence respectively. If this Kruskal-Wallis test shows statistical significance, for each subgroup a two-sided Mann-Whitney rank test between this subgroup and the union of all other subgroups is carried out; the results of these tests are shown by further $^{**/*}$ at each number, showing statistical significance at $p < 0.01/p < 0.05$ confidence respectively.
Further, the colours show the order of mean Rank/Accept/Sev for each of the groups (i.e., ranking them horizontally). The largest mean is given the darkest colour, and the colour changes to a lighter shade if there is a statistically significant difference between the distribution of ranks/scores of the current mean and the next largest mean, based on a one-sided paired Wilcoxon rank test. This statistical test is further shown by $_{**/*}$ at the value of the higher cell, showing $p < 0.01/p < 0.05$ confidence respectively. If more than one cell contains the same colour, there is no statistical significant variation between the rankings/scores for these options.
Lastly, the rightmost column τ lists Kendall's τ correlation coefficients between the rank and the acceptability/severity score respectively for each of the groups. Kindall's τ ranges from -1 (perfect anti-correlation) to 1 (perfect correlation). $^{**/*}$ signifies rejecting the null hypothesis of independence (i.e. $\tau = 0$) with statistical significance at $p < 0.01/p < 0.05$ confidence respectively.

The acceptability scores demonstrate a similar trend. Only *Operations* and *Sales & Service* discriminated between level 4 and 5. Yet none of the divisions inverted the ranking. *Operations* are noteworthy since they clearly distinguished between level 2 and 3 maturity as well as acceptability scores.

4.2.2 Maturity by age

There are three age groups that did not discriminate between level 4 and 5 maturity levels: $35-39$, $50-54$ and $55+$. The $35-39$ group also shows statistically significant lower average level 5 rank than the other age groups, but ranks level 4 statistically significantly higher than the other age groups. All age groups ranked the acceptability of the options according to the maturity levels.

4.2.3 Maturity by location

Responses from location *1: HQ* rank maturity level 4 higher than level 5 as well rank level 2 significantly higher than all other locations. This is also evident in the acceptability score: level 5 is perceived as statistically significantly less acceptable and level 2 as more acceptable than at all other locations. Employees at locations 4, 5 and at minor offices

were unable to distinguish between levels 4 and 5. Location 3 achieved the highest average level 5 rank of 3.6, statistically significantly higher than the average.

Acceptability scores only varied significantly for staff at the minor offices, which collectively scored level 5 with an extremely high score of 4.91. Further, the level 3 score was significantly lower, with a mean of 1.48, making it indistinguishable from level 2's score.

4.3 Behaviour types analysis

The answer options of the four behaviour scenarios map to the four behaviour types. The participants were asked to rank the options in the order they would consider enacting them themselves as well as assign a severity score on a Likert scale to each answer option and to the scenario in general. The statistical analysis that follows is similar to the analysis of maturity levels described above. Again, we show only one analysis table here for brevity (table 1b).

The last line of table 1b shows the analysis of behaviour types for the organisation as a whole. The ranking of the behaviour types is Hierarchist (2.80 mean ranking), Individualist (2.68), Fatalist (2.50) and Egalitarian (2.02). All pairwise differences are statistically significant (see table 1b). The ranking of the severity of the options for each to the behaviour types is less clear as they can only be divided into 3 statistically distinguishable categories (as indicated by the use of three shades of colour only), although the Egalitarian option is seen as statistically significant most severe at 3.76. Further, there is a statistically significant negative correlation between severity score and behaviour type, implying that the employees rank less severe options higher, as may be expected.

It should be noted that there is no inherent ordering between the behaviour types (as it was the case between maturity levels), hence when analysing the data and table 1b, care has to be taken not to infer a ranking of the types themselves relative to each other, but rather work with the ranking of the types by the participants.

While at the level of the whole organisation there are is a statistically significant ordering of the preferences of the behaviour types, this changes considerably when analysing across different subgroups as discussed in sections 4.3.1 to 4.3.3, where there are in many cases only 2 statistically different groups.

4.3.1 Behaviour types by division

In the *Business* division the Egalitarian and Hierarchist are ranked statistically significantly higher and lower, respectively. This is also the case in *Finance & Prof. Services*, but foremost the Individualist type is ranked highest here. *Marketing & Consumer* also agrees on the Egalitarian and Hierarchist differences, but here the Fatalist option is statistically significantly lower ranked than in the organisation as a whole. *Operations* are by far the most Fatalist: they rank this option statistically significantly highest and Egalitarian lowest, and are also much less Egalitarian and Hierarchist than the organisation generally. The *Sales & Services* team agree with the organisational ranking of the types, but they gave a significantly higher score to the Hierarchist option than any other division by at least 0.64.

The *Human resources* division represents the first Hotspot (see boxout 4), as the division shows a non-negative correlation between the option's severity score and rank. This implies that employees choose which option to prefer independent of the severity they assign to that option.

Analysing the severity scores, the *Operations* division is alone in perceiving a full ordering of the options, ranking the fatalist score third most severe and statistically significantly much less severe than the rest of the organisation. This is in stark disagreement with *Sales & Services*, who perceive the Fatalist option much more severely with a ranking difference of 1.44.

4.3.2 Behaviour types by age

There are no statistically significant variations between the different age groups for the Individualist and Egalitarian behaviour types. All the differences occur when considering the Hierarchist and Fatalist types: both the age groups $25-29$ and $30-34$ are statistically significantly more Hierarchist than all other age groups. The age group $50-54$ shows the opposite, they are significantly less Hierarchist. When examining the Fatalist type, the picture changes: The $30-34$ group is significantly less Fatalist, the $50-54$ and the 55+ are significantly more so. In fact the the $50-54$ group rank Fatalist highest, followed by a statistically significant difference by the Hierarchist – an opposite order to the organisation at whole and unique to this group. It is interesting to note that the middle three age groups from 35 to 49 (as well as the under 25 group) have little or no preference between the behaviour types and also rank them nearly equally on the severity scales.

Between the different age groups there are no statistically significant variations of the severity scores for any of the behaviour types.

4.3.3 Behaviour types by location

The predominant behaviour types vary widely by business location. Both locations *1: HQ* and *Homeworker* rank the Individualist options highest, in the case of *1: HQ* because it ranks the Individualist and Hierarchist types statistically significantly higher and lower than the other locations, respectively.

Locations *5* and *Minor Offices* rank Fatalist first; this is followed by a statistically significant lower score for the Hierarchist type at these locations. Locations 2, 3, 4 and *Other* show an opposing trend, ranking the Hierarchist type higher than other locations and the Fatalist type lower. It is worth noting that the *Other* category represents mostly retail workers spread across the company's various sites.

Interestingly, there are also a large number of statistically significant variations in the severity scores, with all four types rejecting the null-hypothesis of equal distribution of the Kruskal-Wallis test. This is also reflected in strong variations in the severity score of the behaviour scenarios across the locations. Employees at location 3 saw all four options as significantly more severe, increasing the severity scores of each option by over 20%, but the scenario's severity score remains unchanged. The opposite effect is portrayed by *Homeworkers*, who rate the scenarios 0.51 more severe than the average, but show no variations for any of the behaviour type severity scores.

There is also a second hotspot present in this comparison: location 2 shows no statistically significantly negative correlation between severity scores and rank, implying that employees at this location choose which option to take independent of the severity they assign to the option.

5. DISCUSSION

Our research applied a scenario-based survey to assess both security maturity levels and self-reported security behaviours, and employee understanding of how risky certain behaviours are. A statistical analysis of the results of the survey conducted at the company allows us to draw several key conclusions regarding the security culture within the organisation. In line with the existing literature, we found that assessing attitudes provides a solid approach to understanding how employees interact with security policy. However, our scenario-based survey approach allows us to go further and detect intra-population differences within the organisation, showing that there are significant differences between different employee groups in how they respond to security-related challenges in the workplace. The salient outcomes are discussed below.

Our analysis of the survey has shown that the organisation in general has a very positive security posture: the majority of employees are at maturity level 5 and there is a downwards gradient of the ranking of the lower maturity levels. This combines well with a founded understanding of the acceptability and severity of the options presented to the employees of the organisation; employees in general choose what are in their opinion the more acceptable and less severe options. This strength is based on the willingness of the majority of the organisation's employees to engage actively with security. The predominant attitude within the company is to adopt good security practices, even when not specifically required to by technology or policy prescriptions. Many members of the organisation reported that they would challenge any breaches of policy they observe in their environment, with older employees being less likely to do so. Where friction exists between the business and security processes, employees take a predominantly Individualist approach to conflict resolution, meaning they rely on their own skills and knowledge. This echoes the results of Rhee et al. [25] and Siponen et al. [27] who both recognise the role of self-efficacy in decision making. Individually-derived approaches to security, driven by personal perception of what constitutes secure practice, can also manifest when policy and support is not known or visible to the individual [17].

The ranking of the behaviour types is also positive, but the differences in their respective rankings are weaker. Hierarchists are unlikely to challenge the existing structures, and while they may follow security policies to the letter, the Individualist that innovates may identify and solve new challenges before they become problematic [17]. Some CISOs might think that it is desirable if all employees were Hierarchists, but it could be argued that it would be counterproductive for an organisation to be exclusively one behaviour type, as there are many benefits in diversity. From a productivity point of view, the organisation requires diversity and even from a security point of view, variation has benefits. A diverse mix of behaviour types may even be essential, as security issues are embedded in, and deeply influenced by, social context such as corporate and national culture [21]. In this sense, these issues have to be understood and addressed before any successful intervention program can be introduced.

Based on the results presented in the previous section 4, we will now discuss a number of areas of the organisation that are of particular interest. These areas hint at where interventions could be focused, or otherwise lessons learned and further studies conducted.

5.1 Targeting interventions

The *Sales and Service* division stands out by having significantly stronger maturity levels. They are also able to accurately assess the severity of the acceptability of the options presented. This is further accentuated by the extremely high rank of the Hierarchist type at the Sales and Services division. This alludes to a highly security competent division that is comfortable in its organisation structures. It should be an exemplary part of the organisation that should be able to provide a benchmark for the rest of the organisation.

Conversely, the *Finance and Professional* division rank maturity level 4 highest and further, this division ranks the Individualist type first. It is an interesting case of a combination of less mature security combined with an Individualist approach to security: interventions could focus here first, as this combination has the potential to create problems in the future. The *Operations* division is most Fatalist, and assesses the options as much less severe than all other employees, while maintaining a high maturity level. This suggests that many Operations employees may have given up trying to achieve their tasks using the organisational structures and policies, and instead attempt to fulfil their business goals as easily as possible. Their classification of the Fatalist options as much less severe implies that there are no negative effects of sidestepping the organisational structures. This division represents the 'disillusioned' section of the organisation. Their security maturity is in line with the organisation as a whole, but they feel that the organisation has ignored their needs. They are a primary target for engagement and it is paramount to find ways to make security fit better into their work.

The *human resources* division turned out to be a hotspot (section 4.3.1) that represents an interesting example of this organisation's security structure. While the employees choose highly mature options that were also most acceptable (with a very strong positive correlation), their choice of behaviour type is independent of the severity of each option. This may seem contradictory at first, but could stem from a diverse set of willful employees who are equally present in all four behaviour types and stand to their decisions. It may be argued that this is in fact a desirable property in a human resources division.

There are interesting variations between the different age groups of the employees. The young (25-34) are more Hierarchist, whereas older employees (50+) are more Fatalist. The middle age groups are split between all behaviour types. This could be interpreted as indicating that younger employees see the benefits of the organisation's structures and support, and might rely on them due to their lack of experience. Most younger people in the company are in the *Sales & Services* division where they experience fraud more directly. At the same time, older employees have diversi-

fied their beliefs and the oldest "have seen it all" and might have become disillusioned with their lack of influence and progress in the organisation. This is irrespective of their maturity ranking, as the differences in maturity ranking are only minor between age groups. In order to ensure that this trend is not repeated in the future, the organisation needs to take particular caret to ensure that the voices of their young employees are heard and respected, especially as the older and therefore usually more respected employees portray a more challenging behaviour type that may often choose to ignore regulations. If breaking the rules is seen as a sign of seniority, it is toxic since older employees should be role models. Compliance should not be something that is only for those lower in the company's structure.

There are clear cultural differences between the business locations. This manifests in strong variations in maturity levels as well as maturity types. The HQ uniquely ranks maturity level 4 higher than level 5. Separately, it also ranks the Individualist option first amongst the behaviour types. There is a strong absence of Hierarchists at this location. Interestingly, the organisation implemented a hot-desking policy here, which may explain the strong individualism that is present. While the low security maturity present at this location is suboptimal, the unique distribution of behaviour types may be positive and act as a foundation for involving all employees in security in diverse ways. For example, it is in the HQ where the organisation needs to constantly reinvent itself through use of new products and services, and a large number of Individualists may support this.

More worrisome for the organisation's well-being are the employees at locations 5 and *minor offices*, who rank the Fatalist option highest. Further investigations may be required to find a solution to improve the distribution of behaviour types at these locations.

Lastly, location 2 represents a hotspot (section 4.3.3) that is similar to the *human resources* division mentioned above. Here, there is no correlation between an option's rank and its severity score; that is, employees choose what option to take independent of how severe they perceive this option to be. This hints at an organisational site where the employees are well aware of the security impact of their options, but are inclined to choose the options that they have learned will work at their locations. While further investigations at this site may be a prudent course of action, interventions may be fruitful particularly as the Fatalist type is uncommon and maturity levels are above the organisation's mean.

We were able to present our findings to the company at board level; as a result the security managers restructured security spending for the following year to target the locations, divisions and age groups we had identified as giving the most concern. Specifically, managers set targets to improve communication with these groups, and in particular to promote the need for leadership and to enable non-confrontational challenging of policies, amongst employees aged 25 to 45. Location 1 (HQ) was also targeted for special attention in this regard. This outcome showcases the real-world impact our methodology is capable of creating.

5.2 Limitations

We divide our discussion of limitations into those relating to methodology and findings. Each engagement with an organisation is time-consuming, involving interviews which are used to generate scenarios specific to the organisation. We envisage that the cost will decrease with further iterations of the methodology, but may present a high barrier of entry.

We would like to emphasise that from an organisational point of view however, employing our methodology is worthwhile because it creates a benchmarking tool that the organisation can use to re-evaluate and monitor over time to compare to previous iterations. As researchers working with many organisations, we envisage that the organisations where we conduct interviews yield a library of questions that we may be able to reuse for other organisations that are broadly similar. We also acknowledge that being able to benchmark a company's security posture would feed the particular desire by some organisations to compare themselves with other organisations in the same sector. We would be hesitant to use our methodology to compare multiple organisations because it is difficult to obtain meaningful results as companies are complex and unique.

Acknowledging these demands, the authors are variously involved in efforts to simplify the extraction of meaningful results from interview data through additional tools (e.g., [7] and [8]).

These mappings from scenario option to maturity level and behaviour type need more extensive validation. As yet the mapping from scenario option to maturity level and behaviour type are not thoroughly validated; but we have attempted to make them match to [2] and section B as closely as possible. As the scenarios and options were created specifically for the organisation and the survey had to be conducted reasonably quickly after the interviews, a thorough validation was not possible. Validating these mappings is part of our ongoing work.

A large proportion of the respondents were from the *Sales & Services* division. While the statistical tests for table 1 have accounted for this, a large proportion of the *Sales & Services* staff worked at location *Other*, and fit in the younger age groups. These employees have more contact with customers and are more exposed to fraud and since they are younger, they tend to be more receptive to training. We were careful to make sure that this did not bias the results, but a perfect study may have sampled the organisation's populations more carefully.

Our survey did not capture many contributing factors to the participants responses that may have helped to explain their answers. The respondents background (e.g., computer literacy, previous jobs, other relevant experience) would have provided hints at a number of other relationships worthwhile studying, and potentially allow us to tailor interventions even more specifically. We collected free-text responses at the end of the survey that we will analyse as part of future work, they might help us shed more light on employees' reasoning and justification for their choices. Data collection does not stop once the intervention phase is reached – the methodology presented here supports decision-makers to identify broad employee categories and hotspots to target for improvements. A follow-up intervention may in itself involve data collection to identify contributing factors to particular behaviours.

6. CONCLUSION

The methodology presented here allows organisations to take steps towards empirically assessing the security culture, as well as gaining an understanding into the predominant behaviours and attitudes found within the organisation. We address the issue of scalability by deploying a scenario-based survey that employees can complete in 10-15 minutes but can therefore be deployed to a large enough fraction of the organisation to be representative. We ground all the scenario details, and answer options, in information gathered from a series of semi-structured interviews with employees of the organisation. We demonstrate that this approach allows us to detect statistically significant differences between employee groups that can inform targeted interventions. Business area, age, and geographical location all provide axis of differentiation. Giving an organisation an understanding of these details can potentially allow them to plan their future training, communication, awareness and policy making strategies more effectively. Enabling targeted interventions that focus on particular employee groups can save employees from both being involved in non-targeted interventions and needing to determine if they apply to them. Targeted interventions are then a good step towards reducing the draw on employees' compliance budget [6].

6.1 Future work

Tailoring our diagnostic tools to the operating context and working practices of the organisation provides meaningful results. Security awareness material can similarly be crafted to resonate with the experiences of employees in weaving security into their productive tasks. Tsohou et al. [32] discuss ways of interpreting cognitive and cultural biases – such as those described in the behaviour-type scenarios – to produce effective security awareness material. Awareness should be a two-way street: security specialists should use the understanding of what drives individuals' behaviour to engage with those individuals and be receptive and find collaborative solutions to conflicts between security and business processes.

Siponen and Vance [28] define conditions for field studies of policy violations – another avenue for future work could compare employee behaviour to the declared information security policy of an organisation. This will expose gaps in policy, and help to identify policies which are routinely ignored or misinterpreted, or communicated badly. For instance, Renaud and Gaucher [24] note a distinction between an intention to behave in a secure manner, and enacting a secure behaviour in practice – if an intention to comply is not supported by the infrastructure of the organisation the solution will not lie in the production of awareness material [3].

6.2 Acknowledgement

We would like to thank our partner company for making this research possible. Many thanks to Iacovos Kirlappos and many others for their help in data collection. Adam Beautement, Simon Parkin and Angela Sasse are supported by EPSRC and GCHQ, grant number: EP/K006517/1 ("Productive Security"). Ingolf Becker and Kat Krol are funded by EPSRC's grant to the Security Science Doctoral Training Centre, grant number: EP/G037264/1.

6.3 Supplementary material

The additional tables for section 4 (i.e., two more sets of tables similar to table 1 and the ipython notebook containing related statistical tests) can be accessed at `http://dx.doi.org/10.14324/000.ds.1496888`.

References

[1] A. Adams and M. A. Sasse. Users are not the enemy. *Communications of the ACM*, 42(12):40–46, 1999.

[2] J. Adams. Risk and morality: Three framing devices. *Risk and morality*:87–106, 2003.

[3] E. Albrechtsen. A qualitative study of users' view on information security. *Computers & security*, 26(4):276–289, 2007.

[4] E. Albrechtsen and J. Hovden. The information security digital divide between information security managers and users. *Computers & security*, 28(6):476–490, 2009.

[5] A. Beautement, R. Coles, J. Griffin, C. Andronis, B. Monahan, D. Pym, M. A. Sasse and M. Wonham. Modelling the human and technological costs and benefits of USB memory stick security. *Managing information risk and the economics of security*:141–163, 2009.

[6] A. Beautement, M. A. Sasse and M. Wonham. The compliance budget: managing security behaviour in organisations. In *New Security Paradigms Workshop (NSPW)*, 2008, pages 47–58.

[7] I. Becker, S. Parkin and M. A. Sasse. Combining qualitative coding and sentiment analysis: deconstructing perceptions of usable security in organisations. In. Learning from Authoritative Security Experiment Results (LASER). IEEE, San Jose, California, US, 2016.

[8] O. Beris, A. Beautement and M. A. Sasse. Employee rule breakers, excuse makers and security champions: mapping the risk perceptions and emotions that drive security behaviors. In. NSPW. ACM, Twente, Netherlands, 2015.

[9] J. M. Blythe, L. Coventry and L. Little. Unpacking security policy compliance: the motivators and barriers of employees' security behaviors. In *Symposium On Usable Privacy and Security (SOUPS)*. USENIX Association, Ottawa, 2015, pages 103–122.

[10] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[11] B. Bulgurcu, H. Cavusoglu and I. Benbasat. Information security policy compliance: an empirical study of rationality-based beliefs and information security awareness. *MIS quarterly*, 34(3):523–548, 2010.

[12] R. E. Crossler, A. C. Johnston, P. B. Lowry, Q. Hu, M. Warkentin and R. Baskerville. Future directions for behavioral information security research. *Computers & security*, 32:90–101, 2013.

[13] J. D'Arcy, T. Herath and M. K. Shoss. Understanding employee responses to stressful information security requirements: a coping perspective. *Journal of management information systems*, 31(2):285–318, 2014.

[14] G. Dhillon and J. Backhouse. Current directions in is security research: towards socio-organizational perspectives. *Information systems journal*, 11(2):127–153, 2001.

[15] K. H. Guo, Y. Yuan, N. P. Archer and C. E. Connelly. Understanding nonmalicious security violations in the workplace: a composite behavior model. *Journal of management information systems*, 28(2):203–236, 2011.

[16] C. Herley. So long, and no thanks for the externalities: the rational rejection of security advice by users. In *New Security Paradigms Workshop (NSPW)*, 2009, pages 133–144.

[17] I. Kirlappos, S. Parkin and M. A. Sasse. Shadow security as a tool for the learning organization. *ACM SIGCAS Computers and Society*, 45(1):29–37, 2015.

[18] S. Pahnila, M. Siponen and A. Mahmood. Employees' behavior towards IS security policy compliance. In *Annual Hawaii International Conference on System Sciences HICSS*. IEEE, 2007, 156b–156b.

[19] K. Parsons, A. McCormac, M. Butavicius, M. Pattinson and C. Jerram. Determining employee awareness using the human aspects of information security questionnaire (HAIS-q). *Computers & security*, 42:165–176, 2014.

[20] M. Paulk, B. Curtis, M. B. Chrissis and C. V. Weber. Capability Maturity Model^{SM} for Software. Version 1.1. CMU/SEI-93-TR-024. Technical report. 1993.

[21] S. L. Pfleeger, M. A. Sasse and A. Furnham. From weakest link to security hero: transforming staff security behavior. *Jhsem*, 11(4):489–510, 2014.

[22] J. T. Reason. *The human contribution: unsafe acts, accidents and heroic recoveries*. Ashgate Publishing, Ltd., 2008.

[23] K. Renaud. Blaming noncompliance is too convenient: what really causes information breaches? *Security & privacy, IEEE*, 10(3):57–63, 2012.

[24] K. Renaud and W. Goucher. The curious incidence of security breaches by knowledgeable employees and the pivotal role a of security culture. In, *Human Aspects of Information Security, Privacy, and Trust*, pages 361–372. Springer, 2014.

[25] H.-S. Rhee, C. Kim and Y. U. Ryu. Self-efficacy in information security: its influence on end users' information security practice behavior. *Computers & security*, 28(8):816–826, 2009.

[26] N. Röder, M. Wiesche, M. Schermann and H. Krcmar. Why managers tolerate workarounds–the role of information systems, 2014.

[27] M. Siponen, S. Pahnila and A. Mahmood. Employees' adherence to information security policies: an empirical study. In, *New Approaches for Security, Privacy and Trust in Complex Environments*, pages 133–144. Springer, 2007.

[28] M. Siponen and A. Vance. Neutralization: new insights into the problem of employee information systems security policy violations. *MIS quarterly*, 34(3):487, 2010.

[29] J. M. Stanton, K. R. Stam, P. Mastrangelo and J. Jolton. Analysis of end user security behaviors. *Computers & security*, 24(2):124–133, 2005.

[30] M. A. Tariq, J. Brynielsson and H. Artman. The security awareness paradox: A case study. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2014, pages 704–711.

[31] K.-L. Thomson and R. von Solms. Towards an information security competence maturity model. *Computer fraud & security*, 2006(5):11–15, 2006.

[32] A. Tsohou, M. Karyda and S. Kokolakis. Analyzing the role of cognitive and cultural biases in the internalization of information security policies: recommendations for information security awareness programs. *Computers & security*, 52:128–141, 2015.

[33] M. Workman, W. H. Bommer and D. Straub. Security lapses and the omission of information security measures: a threat control model and empirical test. *Computers in human behavior*, 24(6):2799–2816, 2008.

A. BEHAVIOUR AND ATTITUDE SCENARIOS

Below, we provide the texts of the behaviour and attitude scenarios used in the study. Labels are included to indicate which option related to which behaviour and attitude type, but these were not displayed to participants in the study and the order of the options was randomised as well.

A.1 Scenario C (Behaviour) – Password Manager

Hina, a member of the Operations division, has recently been required as part of her job to use a new piece of software about once a week. This requires her to log in to the service using a new username and password combination. Unfortunately the password manager does not work correctly with this new software and fails to store or enter her password. Because of the lack of support Hina is worried about being able to use the service as she struggles to remember infrequently used passwords.

Assuming that Hina decides to continue using the service without the support of the password manager, if you were Hina, what would you do in these circumstances?

Individualist: Store the password using a method of your own devising – you can be trusted to keep it safe.
Egalitarian: Share your password with a trusted member of your working group so that if you forget it they can remind you.
Hierarchist: Stop trying to remember the password and just use the password reset feature to generate a new password each time you need to use the service.
Fatalist: Re-use a password from another service that you have committed to memory.

A.2 Scenario D (Behaviour) – VPN

Robert, an analyst in the Operations team, has a set of logs from secure company hardware that he needs to upload to the manufacturer's website for analysis. He is working from home and unfortunately while connected to the VPN, he is unable to utilise the upload function on the manufacturer's site. It is necessary that the logs are analysed each day so he cannot wait until he is next in the office if he is to successfully complete this task.

Assuming that Robert decides to upload the logs via a different method, if you were Robert, what would you do under these circumstances?

Individualist: Make a local copy of the logs, disconnect from the VPN and upload the logs over your home connection.
Egalitarian: Give the password to the server to a trusted colleague not working from home and ask them to download the logs from the server before uploading to them to the manufacturer.
Hierarchist: Email the logs directly to the manufacturer's customer support email, and ask them to conduct the analysis and send the file back.
Fatalist: Email the logs to a colleague not working from home and see if they can upload the logs via a direct LAN connection.

A.3 Scenario F (Behaviour) – File Storage

Concerned about the safety of his current work, Shamal decides to back up his data, some of which is confidential. As he uses his own laptop under the 'bring your own device' scheme, he usually stores all his work on his drive on the central server but he wants to have a second copy just in case something happens or he loses connectivity to the company network. He thought about using one of the common drives but none of the ones he regularly uses have sufficient space.

Individualist: Create a local copy on the hard drive of your BYOD laptop, it is the only machine you work on so you know it will be safe and this ensures you will always have access to it if needed.
Egalitarian: Use a common drive that you used for an old project and still have access to, as your credentials were never revoked. It has enough space although you do not know who manages it now.
Hierarchist: Use an online service, such as Dropbox, to store the data as it is more under your control.
Fatalist: Back your work up onto a USB stick – you have ordered an encrypted one but while you wait for it to arrive you use a personal stick you have to hand.

A.4 Scenario H (Behaviour) – Credit Check

Karina works as a Sales Assistant in a company store. Her manager has asked her to increase her sales, in order to meet the store's monthly target. In her experience, customers can be put off by the need for credit and ID checks, and sometimes fail them altogether. She knows of a few unofficial ways of making the checks seem less of a problem, or to increase the chance of customers passing them.

Individualist: Attempt multiple credit checks in quick succession in order to try to figure out which details are causing the problem and amend them.
Egalitarian: Give information about the credit check to a few of your personal contacts so that they can prime potential customers on what they need to do to beat the system before referring them to the store.

Hierarchist: Use your employee discount to offer the customer a more attractive deal.
Fatalist: Give the benefit of the doubt when encountering IDs with indicators of possible fraud, such as dates of birth that do not seem to align with the apparent age of the customer, or addresses in different cities.

A.5 Scenario A (Attitude) – ID Badges

Jemima is a member of the Operations team working in Location 1. While sat working at her desk, she notices someone she doesn't recognise walk past without a visible ID badge. This prompts her to do one of the following:

Level 2: Nothing, the security badges are only used for accessing the building and once you are in serve no other real purpose.
Level 3: Nothing, although security badges are meant to be visible at all times it is a formality and it is the job of the security guards to check not hers.
Level 4: Make sure that her own ID badge is visible, seeing someone without theirs reminds her that she should have hers on display.
Level 5: Go and talk to the person and ask if they have a badge. If they have, remind them to have it on display, if not then politely escort them to security.

A.6 Scenario B (Attitude) – Clear Desk Policy

When leaving his desk to go for lunch with some colleagues Darren, a member of the HR team, notices that one of them has left his screen unlocked. The rest of the people he is with don't seem to have noticed, or seem to be OK with leaving it as it is. Darren got into the habit of locking his screen some years ago while working in a different company. As his colleagues start to walk away he decides to:

Level 2: Do nothing, there is no risk here as no-one could get into the office without passing through security. The screen locks are there just as a formality.
Level 3: Do nothing, the screen will automatically lock after a few minutes and this will keep things secure.
Level 4: Lock the screen himself.
Level 5: Quickly find out whose desk it is from the group and ask them to lock it before they leave for lunch.

A.7 Scenario E (Attitude) – Tailgating

Jessica is heading toward an access controlled entry door and notices a man she does not recognise gain entry by following close behind someone else who had tagged in at the door. The two men are walking close together although they do not appear to obviously be in conversation. The second man is holding a cup of coffee in one hand and his laptop in the other. His ID badge is not immediately visible. Jessica decides to:

Level 2: Return to her desk, she sees this sort of thing quite regularly and it is probably because his hands were full that he did not swipe through himself.
Level 3: Do nothing, if he is up to some mischief the security guards will catch him later on.
Level 4: Find a security guard at one of the manned turnstiles and tell them what happened.
Level 5: Follow the man and ask to see his ID badge.

A.8 Scenario G (Attitude) – Secure Disposal

John works as a Sales Advisor in a company store in London. During a busy period of the day he notices that a customer, served by one of his colleagues, has left their paperwork behind. John's colleague grabs the paperwork and throws it into a wastepaper bin under the desk. Seeing this John decides to:

Level 2: Carry on serving customers in the store, all the rubbish will be thrown out at the end of the day anyway so it is no big deal, and using the shredder in the back area, locked by a keypad, is inconvenient when the bin is right there.
Level 3: Make a note to check with his manager what the appropriate action would be, as it has been some time since he took the Data Protection training module and he cannot clearly remember the details.
Level 4: Go and grab the paperwork out of the bin when he has a spare moment and take it to the shredder in the back of the store.
Level 5: Go over immediately and ask his colleague to take the paperwork out of the bin and put it in the shredder, having documents lying around exposes both the store and the customer to the risk of identity theft.

B. MATURITY MODEL

This model expresses the maturity of the security culture within an organisation in terms of how aligned with the policy employee behaviour is, and also how integrated the policy is with the primary business process of the organisation. Most critically the model does not represent a checklist of required behaviours for employees, but aims to reinforce the synergy and co-operation required between employer and employee to deliver effective security. As such it is not possible to reach the highest levels of the model in an environment with an inefficient or poorly implemented policy that is in conflict with the primary process of the organisation. Thus the model is capable of guiding change both for the organisation and the individuals that work for it.

This model is based on the Carnegie Mellon Capability Maturity Model [20]. This model expresses the degree of formality associated with various processes. What we need from our security behaviour model is a characterisation of what represents effective employee security behaviour, as observed by the organisation. This will then act as a scale against which progress can be measured, as well as a tool for identifying the current state of security behaviour. The CMM consists of five levels, moving from unplanned/unmanaged through a managed state to one of optimisation through incremental innovation. These levels are listed below with definitions for reference.

Level 1 – Initial (Chaotic) It is characteristic of processes at this level that they are (typically) undocumented and in a state of dynamic change, tending to be driven in an ad hoc, uncontrolled and reactive manner by users or events. This provides a chaotic or unstable environment for the processes.

Level 2 – Repeatable It is characteristic of processes at this level that some processes are repeatable, possibly with consistent results. Process discipline is unlikely to be rigorous, but where it exists it may help to ensure that existing processes are maintained during times of stress.

Level 3 – Defined It is characteristic of processes at this level that there are sets of defined and documented standard processes established and subject to some degree of improvement over time. These standard processes are in place (i.e., they are the AS-IS processes) and used to establish consistency of process performance across the organization.

Level 4 – Managed It is characteristic of processes at this level that, using process metrics, management can effectively control the AS-IS process (e.g., for software development). In particular, management can identify ways to adjust and adapt the process to particular projects without measurable losses of quality or deviations from specifications. Process Capability is established from this level.

Level 5 – Optimizing It is a characteristic of processes at this level that the focus is on continually improving process performance through both incremental and innovative technological changes/improvements.

When considering a Security Behaviour version of this model we must consider how to convert these organisational indicators to indicators of personal behaviour. One approach is to consider how the individual is managing or motivating their own behaviour – what factors they are considering when planning their security actions. At the highest level, they will be actively working toward an improved and improving security culture. At the lower levels employees will be following the policy by rote (possible reluctantly, ineffectively or incompletely) or simply taking actions as they see fit, based on their own internal security model with no input from the organisation. The following levels represent this range of behaviours.

Level 1 – Uninfluenced At this level, user behaviour is mediated only by their own knowledge, instincts, goals and tasks. Their actions will reflect only the needs of their primary task and will only deviate from that where their internal security schema conflicts with those actions. While some members of the organisation may be sufficiently knowledgeable to act securely it is expected that employees at this level will introduce a range of vulnerabilities in to the system. In practice this level can only exist where employees are working on non-organisational systems, as even the act of logging in to a managed network means that organisational security is exerting an influence.

Level 2 – Technically Controlled Employees at this level act as in level 1 except where technical controls exist that enforce policy on a case-by-case basis. Technically controlled employees will follow their own security rules except where they must use organisational systems in the execution of their primary task, and those systems enforce policy at the software or hardware level. Realistically, this is the lowest practical level that employees working in an office environment could function at.

Level 3 – Ad-hoc Knowledge and Application Employees at level 3 follow policy without necessarily a deep knowledge of what it contains. Their security knowledge comes from the 'best practise' or habits associated with their work environment, rather than from being aware of, and understanding, organisational policy.

Level 4 – Policy Compliant Level 4 behaviour demonstrates knowledge and understanding of the policy, and compliance with it, even in situations where the local work environment may include the use of workarounds and frequently made excuses. At Level 4, employees can be considered to be useful role models and guides for security culture within the organisation.

Level 5 – Active Approach to Security At Level 5, employees take an active role in the promotion and advancement of security culture within the organisation. The serve not just the letter of the policy but the intent as well and will challenge breaches at their level appropriately. They see security as a valuable part of the function of the organisation and have internalised this motivation. Level 5 employees are not security zealots, but rather understand the need to balance the security and business processes and champion the cause of security intelligently and effectively.

C. BASIC INTERVIEW QUESTIONS

C.1 Introductory questions

1. What do you do at the company?
2. How long have you been working at the company?
3. What does your usual day involve?

C.2 Security Awareness

1. How does security fit into your day?
2. Do you think your work has any security implications?
3. Do you encounter information that is in any sense confidential or sensitive?

C.3 Clear Desk Policy

1. Is there a policy that says what you should do with your desk when leaving in the evening?
2. Do you have a secure draw or storage area you can use?
3. Do you ever work on paper at all?

C.4 Laptops, Remote working and Removable Media

1. Do you ever use a laptop in the course of your work?
2. How do you share information with colleagues?
3. Do you ever use removable storage devices such as USB sticks?
4. When working from home what systems or technologies do you use?

C.5 Leadership and Management Roles

1. Do you supervise any other people?
2. Does your supervisor ever mention security issues to you?

C.6 Policies, Reporting and Training

1. How much would you say you know about the security policies at your company?
2. Have you ever received any security training?
3. Do you think people generally follow the policy rules?
4. Who would you report a security concern to?

C.7 Optional Topics

1. Compliance and security culture
2. Personal/mobile devices
3. Locking screens
4. Password behaviour
5. Password resets
6. Physical security
7. Customer data
8. Data classification
9. Trust

Intuitions, Analytics, and Killing Ants: Inference Literacy of High School-educated Adults in the US

Jeffrey Warshaw*
University of California, Santa Cruz
1156 High Street
Santa Cruz, CA, USA 95064
jwarshaw@ucsc.edu

Nina Taft
Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
ninataft@google.com

Allison Woodruff
Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
woodruff@acm.org

ABSTRACT

Analytic systems increasingly allow companies to draw inferences about users' characteristics, yet users may not fully understand these systems due to their complex and often unintuitive nature. In this paper, we investigate inference literacy: the beliefs and misconceptions people have about how companies collect and make inferences from their data. We interviewed 21 non-student participants with a high school education, finding that few believed companies can make the type of deeply personal inferences that companies now routinely make through machine learning. Instead, most participant's inference literacy beliefs clustered around one of two main concepts: one cluster believed companies make inferences about a person based largely on a priori stereotyping, using directly gathered demographic data; the other cluster believed that companies make inferences based on computer processing of online behavioral data, but often expected these inferences to be limited to straightforward intuitions. We also find evidence that cultural models related to income and ethnicity influence the assumptions that users make about their own role in the data economy. We share implications for research, design, and policy on tech savviness, digital inequality, and potential inference literacy interventions.

1. INTRODUCTION

The ways that companies gain insights from consumer data have changed drastically in the last few decades, and yet we know little about how the general public's understanding has kept up with those changes. Many decisions that companies historically made through market research and coarse, demographic segmentation are now instead driven by statistical inferencing, through online data mining and machine learning. The ability to algorithmically process behavioral data and look for patterns across millions of users allows companies to infer characteristics that users may believe are difficult to guess or hidden online, such as their hobbies and likes [35], their age and ethnicity [30,51], and their personality and values [17,18]. These inferences are used in a wide range of everyday contexts, for example to offer personalized ads and product recommendations, or to offer differentiated pricing or employment opportunities [14,43].

Because algorithmic inferences can have economic and other far-reaching consequences in people's lives [14,43], it can be valuable for people to have an understanding of what can be inferred about them and how. However, the systems that generate these inferences are often complex and/or opaque. Recent research has emphasized the surprise that many users experience when learning about inferential systems [18,58,61], implying a gap likely exists between what people generally believe companies currently do with their data, and what the state-of-the-art actually is. To date, though, research on *digital literacy* has focused on knowledge of data collection practices [5,44,57] but to our knowledge has not explored beliefs and misconceptions people hold about companies' inferencing methods and capabilities. We argue for the inclusion of these beliefs as a subconstruct of digital literacy, and we introduce the term *inference literacy* to describe it.

In this work, we share results from a qualitative study assessing the inference literacy of 21 US non-student adults with a high school degree but no post-secondary degree, the modal educational attainment in the US, comprising 49% of the adult US population [60]. Inspired by previous work on folk models [62,63], we explored beliefs and misconceptions, and found two distinct clusters within our sample. One cluster believed that online companies rely on now-outdated market research strategies that companies used decades ago [22], such as data collection through surveys rather than through tracking user behavior online. This cluster also interpreted inferential techniques used by companies as constituting *stereotyping*, and expressed worries about hackers and scammers. The other cluster believed that companies mine people's online behavior to infer their preferences, using computer analytics to make intuitive predictions about users. Neither cluster had fully accurate beliefs, and both clusters had misconceptions that have important user experience implications.

Further, we argue for the broadening of cross-cultural studies in usable privacy and security to explicitly include qualitative differences based on social class and ethnicity rather than just on national culture. Building on research that has examined folk models that people have about online phenomena at an individual [62] or national [28] level, we provide evidence within our high school-educated sample that users' interpretation of the privacy ecosystem can vary substantially based on social class and

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22-24, 2016, Denver, Colorado.

* Work performed while at Google Inc.

ethnicity. We link this to cognitive anthropology research on *cultural models*, the sets of assumptions that members of a group form over time based on shared experiences [11,23]. Cultural models about personal agency and choice, both of which might affect a person's online privacy beliefs and behavior, vary between socioeconomic and ethnic groups [54,56,23]. In this work, we saw differences in the framing of privacy decisions as risks or choices as a function of participant income, relating to differences in cultural models of personal agency based on social class [54,56,23]. We also found that ethnic minority participants interpreted companies inferring their preferences based on their ethnicity as stereotyping, which we contextualize in terms of social psychology research on ethnic minority groups' experiences with discrimination in consumer settings [15,34,50].

Our main contributions are:

- We present a novel study of *inference literacy*, describing the beliefs and misconceptions that 21 US adults with a high school education hold about how companies make inferences from their online data.
- We report and describe two clusters of beliefs that together describe 19 out of our 21 participants. These clusters link inference literacy to cultural models of agency and point to apparent digital inequalities based on socioeconomic factors, including income and ethnicity.
- We argue for the redefinition of tech savviness and digital literacy to include inference literacy, as well as for cultural models based on social class and ethnicity to be included in future online privacy research.

2. RELATED WORK

2.1 Inferencing methods

Our work explores users' understanding of current inferencing methods commonly used by companies. Companies currently rely heavily on machine learning to make inferences about users, applying techniques such as supervised or unsupervised learning, reinforcement learning, deep learning, or neural networks to find complex patterns in behavioral data [2,4,41]. These methods are typically applied to datasets containing multiple streams of data aggregated from large groups of users in order to find correlations between variables of interest [6]. These techniques can uncover strong correlations, similar to those humans might intuitively guess when presented with frequently co-occurring behaviors; they can also uncover weaker, more unintuitive correlations that are only detectable by combining data from a large number of users. A key property of the relationships between variables that these techniques uncover is that they are generalizable to new users: learning how to predict variable A from variables B, C, and D based on a group of people who shared all of those variables enables the creation of a system that can infer variable A for people who have not shared it. For example, analyses of large datasets have led to systems that can infer a person's gender based on their movie ratings [66], their religion from their search queries [7], their sexual preference from their social media likes [30], and their personality and values from their social media text [17,18]. A system's confidence in inferring an unknown characteristic typically increases with the number of predictor variables available, but even a small number of data points can be used to make a better-than-random guess, e.g., [30].

Despite the key role that machine learning systems play in the data economy, the workings of inferencing systems are often opaque, lacking transparency to users about what data they use or how they work. There are a few efforts that have studied users' reactions to inferencing systems. [61] presented social media users with personality profiles that an inferential system automatically generated from their social media posts, and users' reactions spanned from surprise at how accurate the profiles were, to creepiness and learned helplessness about whether they could decline to share them in different settings. In [58], the authors studied users' reactions to online behavioral advertising and found that they felt it was both useful and scary at the same time. Kulesza and colleagues found that having the ability to correct an automatic recommender system does not in itself improve users' mental models of the process by which the system makes inferences [31,32]. Instead, they saw participants' confidence in unsound mental models increased over time unless they received a structured educational intervention prior to using the system [31]. In the current study, we assess people's global beliefs about what data companies use in inferencing, what methods companies use to make inferences, and limitations of inferencing systems.

2.2 Folk models of online privacy and security

Several recent online privacy and security studies have explored folk models, sets of beliefs and misconceptions that non-experts have about a particular topic. Rather than assuming non-experts have zero knowledge, folk models acknowledge that people develop their own lay theories to explain ambiguous situations they encounter. The online privacy landscape is often ambiguous, leaving users to come up with their own incomplete explanations for phenomena like hackers and viruses [28,62]. Research on users' understanding about online data collection and specific inferential systems has found non-experts often have consequential misunderstandings about the online landscape [26], including systems they commonly interact with such as autocomplete [48] or behavioral advertising [39,58]. We build on and extend this previous work by exploring folk models related to inferencing.

Importantly, folk models are not independently produced. Social and cultural factors affect them as well. Informal stories and advice about privacy and security are commonly shared [49]. These informal sources of information may include out-of-date information, with some security advice that non-experts endorse being decades behind what experts recommend based on current threats [25]. This *social* aspect of folk models has been discussed in research in online privacy and security, but *cultural* influences are rarely discussed. Folk models of viruses and hackers do appear to differ cross-culturally [28,62], but neither study examined the relationship between culture and the different folk models evidenced in their samples. In the current study, we explore folk models of inference literacy through an explicitly cultural lens, examining the role of cultural models in shaping beliefs and attitudes about the data economy.

2.3 Cultural models and technology

Cultural models are sets of implicit assumptions that develop based on shared experiences and common history, which differ qualitatively between different cultural groups [11,23]. Culture has often been applied in HCI to describe differences based on national culture or language [9,13,21], but other fields describe substantial differences in cultural models based on other features,

including educational attainment, social class, and ethnicity [47,54,56]. Recent research has described cross-cultural differences in definitions of privacy based on national or religious culture, and how those differences relate to the trade-offs users make in online settings [1,59]. In the current work, we link inference literacy beliefs and attitudes to two specific types of cultural models. First, we draw on research showing different cultural models of personal agency for middle-class and working-class Americans [54,56]. Middle-class Americans typically develop a more *independent* sense of agency, expecting to exercise control within their environment. By comparison, working-class Americans tend to develop a more *interdependent* sense of agency, expecting to cope with external factors rather than exercising independent choice themselves. This stems from differences in economic and environmental constraints between these groups [56], and is reinforced by socialization and media consumption that differ by education and income [40,54].

We also draw on research on experiences of marginalized ethnic groups with stereotyping and discrimination. Ethnic minorities in the US often encounter stereotyping and discrimination across various settings, including in education [3,55], while driving [37], and as pedestrians [16]. This pattern extends to consumer experiences as well. Research on "shopping while Black" has shown substantial differences in customers' treatment in US retail settings based on their ethnicity [34,50,52]. Ambiguously discriminatory experiences such as being ignored, followed by retail staff, or not given service can be interpreted as an institutional distrust for or devaluing of them based on their ethnicity [34,50,52]. Inferential systems can themselves be ambiguous to users in how they operate, and the line between personalization and stereotyping may not always be clear. In this study, we include participants from marginalized ethnic groups to obtain their perspectives on this and other inferencing topics.

2.4 Digital inequality

Several studies have looked at digital inequality: ways that offline socioeconomic inequalities related to demographic categories like educational attainment, income, ethnicity, and age are reproduced in online settings. There are differences in internet usage by educational attainment. Those with a high school degree tend to engage in fewer different types of online activities [46,65], fewer capital-enhancing activities online [20,46], and are more likely to be reliant on a smartphone rather than computer to access the internet than the college-educated [53]. Despite the common belief that the "do-it-yourself" opportunities that online access enables are sufficient to decrease social inequalities, socioeconomically disadvantaged internet users benefit from online access at the same or slower rates compared to those with higher incomes or education [12,38,45]. High impact decisions of inferential systems such as credit scoring often contain systematic errors or design decisions that disproportionately disadvantage those with lower SES or non-European-American ethnicity [14,43]. In the current study, we explore a potential digital inequality: whether differences in inference literacy are related to socioeconomic status.

3. METHOD

For our study of inference literacy, we collected data from a sample of US adults with a high school education to learn their beliefs and misconceptions about companies' inferencing methods. Each session contained two main sections that took place consecutively: an interview to elicit the participant's existing beliefs about what, how, and why companies collect and use their data; and a teaching intervention for the participant to learn two basic principles of current inferencing methods. We focus in this paper mainly on data from the interview section, but we include relevant details about the design of the teaching section in Appendix A.

3.1 Study Design

To explore inference literacy in our participants, we adapted Oakleaf's "Information Literacy Instruction Assessment Cycle" (ILIAC) [42]. The ILIAC is an iterative educational research method that aids in creating learning activities that assess a student's knowledge before, during, and after the activity. This method has been applied successfully in educational settings where the goal is to assess and teach digital literacy concepts in a single session [42].

We adapted this method to fit the current study, going through four full cycles, each of which took 1-3 weeks and included 2-12 participants. The overwhelming majority of changes were made to the teaching procedure, with only minor wording changes made to interview questions between cycles.

3.2 Participants

We collected data from 23 participants in total between July and September of 2015, all of whom were recruited by a research recruitment firm with a respondent database containing San Francisco Bay Area residents. We recruited participants who had a high school degree or the equivalent (i.e., GED) but no post-secondary degree, and who were not currently enrolled in post-secondary education. In addition, we aimed to explore socioeconomic and cultural differences in folk models of inference literacy, so we recruited a diverse sample in terms of age, gender, ethnicity, household income, parental status, occupation, and political beliefs. We created a recruitment screener that asked about these demographic categories, as well as several other questions, such as which internet-accessible devices the participant owned, and news sources the participant uses.

Two participants out of the 23 participated in a pilot study. Because the procedure changed significantly based on the pilot, we exclude these pilot participants from the analysis reported here. The 21 participants in the final sample include 10 women and 11 men, ranging in age from 18 to over 65 years of age. Eleven participants identified as White or European-American, 5 as Black or African-American, 3 as Asian-American or Pacific Islander, 3 as Hispanic or Latino, and 3 identified as having mixed or multiple ethnicities. Occupations were varied, including waste management driver, payroll clerk, security guard, HVAC technician, retired, and unemployed. Participants were interviewed in-person in one of two locations, Mountain View, California (n=16), or San Francisco, California (n=5), and were compensated for their time.

3.3 Session Procedure

We first describe the general structure of the session procedure and then detail each component in the order participants experienced it. The same interviewer led each participant through a 90-minute session with two main components: (a) a semi-structured interview meant to elicit existing beliefs and misconceptions about how companies make inferences from their data, and (b) a teaching intervention during which participants engaged with real world examples of inferences that companies can or cannot make from data. Prior to each session, the

interviewer verbally walked the participant through an informed consent form that described the study. With participants' permission, each session was video recorded to facilitate transcription and coding. Another member of the research team observed each session from a separate room either during or after the session, taking notes that included quotes representative of that participant's beliefs, and preliminary themes that arose across multiple participants' sessions. Between sessions, the research team frequently met to discuss observations, develop the analysis plan, and make changes to session procedure for future cycles.

3.3.1 Belief elicitation interview

Inspired by Wash's work on eliciting participants' lay beliefs about home security [62], the first portion of each session consisted of a semi-structured interview that we developed to learn participants' existing beliefs about how companies collect their data and use it to make inferences about them. The interviewer used a paper script containing questions to facilitate the conversation, and began by asking participants their educational background, occupation, and familiarity with machine learning. Only two participants had heard of machine learning, both of whom claimed it referred to some kind of rudimentary artificial intelligence.

To ground the belief elicitation interview in terms of each participant's daily experiences, the first and main prompt for each participant was, "Think about what you'll do online today, and talk me through things that companies will try to figure out about you based on what you do online today". The responses to this prompt were detailed and varied. Participants referred to different settings, with some referring exclusively to smartphones or laptops whereas others described mixed usage of devices. We did not constrain the companies participants talked about, and they described interactions with a wide range of companies for a variety of tasks, including checking email, social media browsing and posting, online banking, retail browsing and purchasing, and watching videos online. Many beliefs about how companies collect and use data came out naturally as participants described their daily online experiences. If they did not arise spontaneously during the interview, the interviewer asked follow-up questions to elicit more detail on each participant's beliefs, including whether, why, and how they believe companies collect data; what kinds of data companies do and do not collect; and whether and how companies make guesses about individuals' characteristics. After probing the contents and sources for each of these potential beliefs, the interviewer concluded the belief elicitation interview and moved to the teaching intervention phase of the session.

3.3.2 Teaching intervention

The goal of the teaching intervention section was to assess participants' explanations about the inferencing processes and capabilities companies deploy, before and after providing participants with brief explanations about modern data collection and inferencing phenomena. After each explanation, the interviewer conducted a card-sorting task with the participant where they rated and discussed the likelihood that companies can make a particular inference from a particular type of user data. Because the focus of the current paper is on participants' pre-existing beliefs, much of the data collected in this section is outside the scope of the coding and results described in this paper. We did use participants' responses to the pre-test assessment, as they were directly relevant to beliefs about inferencing, and the pre-test was given prior to any teaching: "If a social media company wants to learn more about their users, what would they be able to figure out about a user even if that person didn't tell them? How would they figure that out? What would be impossible for a social media company to figure out about someone?" We also used a small number of beliefs that participants shared after the teaching intervention where it was clear that these were pre-existing, e.g., "I always thought it was X" after we taught them Y. We include the remainder of the teaching intervention procedure as Appendix A.

3.4 Coding

In this section, we detail the affinity diagramming and coding of participants videos and transcripts that allowed us to characterize participants' inference literacy beliefs and attitudes about the data economy.

We began analyzing the interview data by creating affinity diagrams [19], taxonomies where participants' perspectives could be grouped across various axes. Some of these diagrams were digital, containing quotes from interviews that we sorted according to thematic differences in how participants described inferencing phenomena. Other diagrams were physical, and used the participants as the unit of analysis. These holistic groupings allowed us to tease apart the key components of qualitatively different folk models about data collection and inferencing, as well as to analyze for cultural and socioeconomic themes such as stereotyping and risk perception. The research team discussed these diagrams as they were created, iterating on them several times during the analysis process.

Additionally, we reviewed the transcripts to identify and define codes similar to [8] to describe the wide range of beliefs participants expressed. The interviewer first coded each transcript, obtaining feedback from the entire research team about ambiguous codes. This coding process was iterative, so that transcripts read early on were reviewed to check for codes that were discovered or refined later in the coding process. To establish intercoder reliability [33], another author coded each of the transcripts for the key beliefs described in the results below. Intercoder agreement was above 75% for the first five transcripts analyzed, and above 80% for the first pass through all 21 transcripts. Disagreements between the first and second coder were resolved by reviewing the transcripts and discussing to come to agreement. In the majority of these disagreements, the two coders agreed about the participant's belief but had different opinions about the level of proof required to confidently assign a code. We took a conservative stance in these cases, requiring supporting statements that were unambiguous or repeated during the interview. After revising the codebook and assessing the remaining disagreements, intercoder agreement was above 90%, indicating that the codes were sufficiently well-defined and reliably assigned during the coding process. The final codebook contained 160 unique codes from the 21 participants.

3.5 Clustering

During data collection and coding, we noticed that some beliefs appeared to frequently co-occur and decided to explore this possibility systematically. As our interest was in describing inference literacy, we focused primarily here on beliefs about data collection and inferencing. After coding the transcripts, we collected the 31 inference-related codes that we had assigned to four or more participants. We then created a vector for each code, each containing the list of participants who had been assigned that code. We manually compared the vectors pairwise, looking for

Table 1. Categorized list of codes contained within each inference literacy cluster, with beliefs that formed the initial core of each cluster in bold. Paper sections discussing each code and related results from affinity diagramming are in parentheses.

	Market Research Cluster (n = 8)	Data Mining Cluster (n = 11)
Data collection beliefs	**Companies collect demographics by surveys. (4.1.1)** **Companies collect personal information from public records. (4.1.1)**	**Companies collect online behavioral data. (4.2.1)** Companies doing retail retargeting taught me that my behavior is collected. (4.2.1)
Inferencing beliefs	**Companies make inferences by having humans make common sense intuitions. (4.1.2)** Companies stereotype users based on their demographics. (4.1.3)	Companies make recommendations using behavioral data. (4.2.2) **Companies use computer analytics to make inferences. (4.2.3)** Companies tailor ads based on what you click. (4.2.3) Inferences are made by analyzing your social network. (4.2.3)
Attitudes	Companies stereotyping is morally wrong. (4.1.3) I am worried about hackers. (4.1.4) I am worried about scammers. (4.1.4)	I feel "watched" or "tracked". (4.2.1)

frequently co-occurring beliefs as well as beliefs that were strongly negatively correlated, such that two vectors had few or no overlapping participants between them.

There were unmistakable links between beliefs about data collection methods and beliefs about inferencing methods that formed the basis for the rest of the clustering process. First, beliefs that online companies collect demographic data by survey or collect personal information by public records were associated with the belief that companies make inferences by relying on common sense intuition rather than computer processing of data. Second, the belief that companies collect online behavioral data overlapped completely with the belief that companies use computer analytics to make inferences. These two sets of beliefs are conceivably complementary in that they each describe one aspect of current inferencing methods, but surprisingly, there was no overlap between these sets of beliefs. Participants who believed that companies use computer analytics did not express that they believed companies collect demographic data by survey, and so on. These two sets of highly distinguishable inference literacy beliefs therefore formed the core for us to explore other connections between our data.

We compared the remaining arrays of belief codes to identify other commonalities, finding two distinct sets of 7 codes that anchored around the core beliefs above. These two clusters of beliefs and attitudes appear in Table 1. Although we began the clustering process seeking to identify sets of beliefs and we did not presuppose that these would be largely mutually exclusive, we found that participants with beliefs in one cluster had few or no beliefs from the other cluster. Because the interviews often surfaced issues related to social class and ethnicity, we holistically analyzed the clusters, drawing on research on cultural models to interpret the codes in light of participant demographics in the results below. Out of 21 participants, 19 were assigned to one of the two clusters. The remaining two participants believed that companies could not or would not collect data about individual users. Although this is a crucial misconception, it was so infrequent that we were unable to explore it systematically in the present study.

The alert reader may wonder whether these clusters constitute "folk models" as described in other literature [28,62]. In that our clusters describe non-expert sets of beliefs held by our participants, it would be reasonable to refer to the clusters as folk models. In this work, we use the word "cluster" for consistency, as it applies equally well to the sets of beliefs themselves and the participants who held them.

3.6 Limitations

We note several limitations of our study methodology that should be considered when interpreting this work. First, due to our focus on describing beliefs of high school-educated adults, we did not include college-educated adults in our sample. This prevents us from comparing inference literacy beliefs across different levels of educational attainment. Second, our sample was not statistically representative of the US adult high school-educated population. The clusters we report should be viewed as a deep exploration of our sample's beliefs and attitudes, but not as generalizing to that population as a whole. Third, we report several misconceptions that people have about inferencing methods, but we do not have data to say that these misconceptions lead to harmful privacy behaviors. Useful behaviors can arise from incomplete or incorrect beliefs [62,63], and that may be the case here as well. Finally, because we touch on socioeconomic status and ethnicity in this work, we include the detail that the research team consisted only of college-educated, European-American researchers. We describe participants' experiences in their own words, but our interpretations may lack context or nuance that may have been more readily available to a more diverse research team.

4. RESULTS

Based on our analysis of the interview data, we identified two main clusters of inferencing beliefs held by participants in our sample. The "market research" (MR) cluster was anchored by a shared belief that companies ask users direct questions about their demographics and personally identifiable information, to sell to them based largely on stereotype. The "data mining" (DM) cluster relied on a shared belief that companies track users' online behavior, to make retail or media recommendations based on their past behavior. We also observed that ethnicity and socioeconomic status were associated with differences in the interpretations participants made about inferencing processes and their own personal agency in the data economy.

Several participants across both clusters had important misconceptions. Participants in both clusters claimed companies rely on human employees to make inferences about users, which we refer to throughout as "humans-in-the-loop". Related to that

misconception, most participants felt inferences are made only based on strong, intuitive connections between two pieces of data, rather than by using multiple pieces of evidence to support an inference. Although data collection practices and inferencing methods differ across companies, our participants rarely made such distinctions.

In the following, we refer to participants with the Market Research cluster of beliefs as MR1, MR2,... MR8, and those with the Data Mining cluster of beliefs as DM1, DM2,... DM11.

4.1 Market Research Cluster

Participants in the MR cluster believed that companies primarily collect data by asking users directly for their personal information (4.1.1), and that companies make shallow, potentially harmful assumptions about them based on their demographics (4.1.2). These participants believed companies make inferences about users based on a priori assumptions about links between two pieces of data, saying things like, "it goes hand-in-hand" (MR1) and "You can make certain summations just by looking at somebody" (MR6). They often described this inferencing process as *stereotyping* (4.1.3), claiming that companies use demographic information like income or ethnicity to make marketing decisions. This was not seen as a benign form of personalization; rather, participants expressed strong moral objections to companies stereotyping in this manner. We also found that despite the interview focusing on companies, participants with these beliefs spontaneously expressed strong concerns about hackers or scammers getting access to their data (4.1.4). This concern about being targeted by criminals often drowned out any apprehension they might otherwise have about what companies would do with their data. In this section, we describe the beliefs belonging to this cluster in greater detail.

4.1.1 Companies collect demographics and personal information from direct sources

The core belief held by participants in the MR cluster was that online companies collect users' demographics and personal information explicitly. There were two main ways they described companies collecting those data: asking a user for it directly in a survey or form, or searching it out themselves from a factual source, like a credit report or public records database.

When asked how companies would figure out characteristics the participant had declined to share with them, participants with this belief felt companies would transparently ask. In response to the interviewer asking how companies would try to learn a user's religion if that user refused to answer a direct question about it:

> "I mean they ask questions and they can somewhat [learn my religion] there. And if they don't, they're gonna ask more questions... If you don't wanna talk about your religion they would probably go…'What type of church do you go to?'...Yeah, ask other questions to try to get around but try to get to the point of whatever it is they're asking about." – MR2

We found that these participants were mostly unaware that companies collect behavioral or other incidental data. Instead, their beliefs hung on largely outdated market research techniques, leaving out automatic or indirect methods of data collection that modern online companies rely on. When we probed whether they believed companies could learn their demographics through a different process, several participants claimed that companies would be unable to learn a characteristic that a user withheld:

> "If you answer that question, it seems like that's what they'll know, that's what they'll have, but if you don't, it seems like they wouldn't know your ethnicity." – MR1

> "I only think that they could figure out my information that I type in." – MR8

Although most focused on companies wanting their demographics, several participants in this cluster also believed that companies are interested in other types of personal information, like addresses, credit card numbers, or social security numbers. They shared stories about personal experiences where their private information was "found out" by companies or individuals searching authoritative sources like public records or credit reports. This method of data collection would be seemingly less visible to the user, but participants who had searched public records for information on themselves or others seemed especially sure that companies would direct their employees to take the same approach. So after obtaining initial information that could seed a search, an employee of the company might look up, for example, a user's age or marital status by seeking out public records.

This is indicative of a common misconception in this cluster about the scale at which companies collect data. It is not feasible for companies to collect data on millions of users by having humans track down public records for each individual, one-by-one. This is, however, the way many in this cluster described companies' data collection processes to us, as humans thumbing through records to find and learn relevant data about an individual:

> "I think they would look at the age. They'd look at the gender. Everything that they have, like where I'm from, where I'm living, what I do, and kind of be like, 'Okay.'" – MR6

This belief that companies make special efforts to directly collect data on each individual was not universal in the MR cluster. Some believed companies simply do not care enough about any single person to hunt down their information, so that companies would ignore and leave alone individuals who decline to share their information. When asked how a company would try to figure out a user's demographics if they withheld it, MR7 said:

> "I think they don't. I think that they just go on. There's so many people. I mean, it's like ants. There's 10,000 of them, and if you kill 9,000, the other 1,000, you're not going to worry about because you got 9,000." – MR7

4.1.2 Companies exploit common sense connections between data to make inferences

The straightforwardness that this cluster ascribed to companies' data collection methods was echoed in their beliefs about how companies analyze data. Market research cluster participants believed companies make inferences by relying not on sophisticated algorithms, but on human employees who make obvious intuitions about the relationship between two pieces of data. The inferences they described companies making were often vague, with their examples tending to revolve around retail recommendations based off of an individual's stated demographics and interests:

> "They ask questions, you answer them, seems like they'll kind of go with whatever you answered. Like if you say you like to ride bikes or something like that, they'll promote bikes, or different things and places you can go to ride bikes. That kind of thing." – MR1

To these participants, companies appear to make an inference based on a single piece of information, and that relationship is intuitive and based on common knowledge. Participants in this cluster did not touch on topics like data aggregation, needing convergent evidence to support an inference, or weak correlations.

A few participants in this cluster did reference retail recommender systems, but their explanations of how these systems work often left out the role of other users' data in guiding recommendations. To some, recommender systems statically present obviously related items as recommendations, e.g., a person buying a phone would be recommended a case for that exact phone. Others believed that companies assume a user's preferences based on their demographics, such as by age or ethnicity.

One misconception about the inferencing process we saw in this cluster was about the directionality of inferences companies make. Although they correctly believed that companies use their characteristics to make inferences about their behavior, many incorrectly believed it was uncommon or impossible for companies to use their behavior to make inferences about their characteristics. When we did prompt them to consider ways that companies might try to infer characteristics from behavior, there was an underlying skepticism that deep insights about a person could come from analyzing online behavior:

> "How could you figure out me by the things I look at?" – MR4

On the contrary, these participants judged companies' inferencing capabilities in terms of their own abilities. We asked participants to explain how companies would infer a characteristic that a user had kept private online, such as their religion or sexual orientation. Participants in this cluster described their own processes as analogues for what they believe companies do, e.g., "While I'm going through somebody's page, I can see a lot about what they're like." (MR8). MR2 put this even more clearly, attaching companies' capabilities to her own:

> "I think they could, 'cause I could." – MR2

As with data collection, these beliefs about inferencing methods appear dated in some respects. Regardless of humans' expertise in making inferences based on intuitive analysis of a person's behavior, companies that serve a large user base have to use inferencing techniques that are scalable in ways that human analysts would not be able to match.

4.1.3 Companies stereotype users based on their demographics, which is morally wrong

The MR cluster included several African-American, Latino, and mixed-ethnicity participants who each expressed concern that inferences companies make appear to be based not on deep knowledge about users but on stereotyping. In their view, companies offer opportunities unequally to people based purely on their ethnicity or income. This was not seen as accidental or benign. Participants who referred to the inferencing process as stereotyping did not mince words. They believed it to be dehumanizing:

> "It begins to be, like, I'm just a statistic for lack of a better word. I'm just a demographic, I'm just a person who spends this amount of money on this in my spare time, and it just becomes - it's so personal but it's impersonal at the same time, you know what I mean? Because it's just information, and they forget that these are people, these are human beings." – MR6

Beyond their moral concerns, they believed stereotyping leads to inaccuracies, particularly due to ignoring intragroup variation:

> "None of us are the same, so we shouldn't be classified as the same...So those companies that put these people in this basket, I think they're sometimes just rounding them up like cattle." – MR4

Those who mentioned this belief were confident that online companies engage in stereotyping, however there was an ambiguity about the exact consequences that result in the examples participants gave of this happening in their own lives:

> "Usually when you do something, they want your background, like your ethnicity or I guess to put you in a certain place, like, maybe they'll know maybe what you want just [based on] your ethnicity. Maybe." – MR1

The ambiguity of the perceived consequences should not obscure the fact that several participants in this cluster believe that this is the process companies engage in. Research on topics like "shopping while Black" [15,34,50] has surfaced how experiences with ambiguous stereotyping are naturally interpreted in light of wider life experiences of racial discrimination, so that online companies' opaque inferencing methods may lead to unflattering interpretations about stereotyping in the absence of clear evidence to the contrary.

4.1.4 High concerns about hackers and scammers can drown out concerns about companies

Although the interviews were only meant to elicit beliefs about companies, several participants in this cluster spontaneously mentioned hackers and scammers as high-stakes threats to their online data. Hackers were described as individuals who would access information either from a device without the owner's knowledge, or via unauthorized access to a company's database. Scammers were described as companies who call, set up phishing websites, or send email in order to obtain information like credit card or social security numbers under false pretenses. The harms participants saw resulting from hackers and scammers were clear: financial loss, identity theft, and damage to their online devices. By comparison, some saw little concrete harm that companies might cause by having their data:

> "You have to worry more about your hackers than you do your companies. Because hackers do bad things with it. They use it, they destroy your credit, they destroy, you know – I don't think a company would want to do that." – MR3

Several participants who shared concerns about companies stereotyping also worried about hackers or scammers misusing their data. These threats appeared to evoke different feelings. Companies stereotyping appeared to create a sense of moral resentment, whereas hackers and scammers came across as adversaries who could be warded off or fought.

4.2 Data Mining Cluster

We now turn to the other main belief cluster. The data mining cluster of beliefs was anchored around a core belief that companies collect data on users' behavior (4.2.1). Participants who had this belief often believed that companies make recommendations based on their prior behavior (4.2.2) (e.g., recommending a song to listen to based on songs the user has previously liked), but they rarely acknowledged that companies can combine demographics with behavioral data to make

inferences. They all believed companies use some computer-based processing of user data to make inferences (4.2.3) such as analysis of social connections to make inferences about them (4.2.3), but they varied widely in the role they believed humans play in making inferences. Some believed algorithms work fully independently, whereas others believed that companies have humans-in-the-loop, employees who oversee individual inferences made by algorithms.

Compared to the market research cluster, the data mining cluster was more familiar with implicit data collection methods. Additionally, these participants were more confident in their beliefs about data collection and inferencing, including in their misconceptions. Participants in this cluster often had mixed feelings about data mining (4.2.4), acknowledging the value it may provide to them personally but often exhibiting signs of resignation in the face of little perceived privacy control.

4.2.1 Companies collect users' online behavior data

The participants in this cluster shared the core belief that companies collect data on users' online *behavior*. The exact data mentioned varied by participant but often included links they click, products they purchase, or videos they watch. Unlike the Market Research cluster's belief that companies ask users to purposefully provide data one survey question at a time, these participants felt companies collect implicit behavioral data automatically. They described companies as "collecting", "tracking", or "watching" all of the things they do online. They were aware that companies depend on their behavioral data to provide online services, drawing from experiences when an inferential system explicitly referenced the data it had collected:

> "I go on Amazon a lot, and say I haven't been on in, like, two months. When I log back on, it remembers. It says, 'Oh, you liked this video game,' maybe, 'People who bought this, buy this.'" – DM1

Although they all believed companies collect some kind of behavioral data, they had varied levels of awareness about *how* and *what* behavioral data companies collect. They most commonly mentioned companies saving their history, e.g., searches, purchases, videos watched. Only a few participants mentioned that companies could collect their location, e.g., through GPS, IP address, or searches made. Those who did mention location tracking believed that companies value location data highly due to the variety of inferences they can make from it:

> "My location, for one, is huge. Pretty much everyone wants to use my location...Probably for marketing purposes so that they can use [it] in some way, like your location to establish where you are a lot...What my hobbies are, what stores I go to and shop [at], and basically what I'm doing with my time. Because it could be used for purposes of marketing, I think." – DM3

> "If you go through my location history for, you know, using public transportation, you're gonna know where I work, how I get there, what I do certain days of the week, things like that. I mean, literally, I'm carrying around a tracking device almost 16 hours a day." – DM11

Unlike the Market Research cluster, participants in the Data Mining cluster were generally aware that companies collect incidental browsing data, such as how long they browsed a website, or what type of device they were using to go online. Still, the examples many participants in this cluster gave about companies collecting activity data contained misconceptions. DM7, for instance, knew that companies can aggregate data from across multiple devices if he is signed in to the same account on each one, but he also mistakenly believed that companies can *only* collect data and make inferences about him if he is signed into an account. This could be a costly misconception, as believing that signed out activities cannot be tracked would provide a false sense of privacy online.

4.2.2 Recommendations are based on a user's past behavior

It seemed apparent to participants in this cluster that recommendations of products and online content such as those on retail or social media sites were based on their own past behavior. This was a conclusion that few in the Market Research cluster had come to. The Data Mining participants, on the other hand, shared several examples of recommendations that companies make to them based on their past behavior:

> "I notice a lot of advertisements on my page, especially to sites that I've been to or things that I've looked at." – DM6

> "YouTube makes guesses on me all the time. When I go to YouTube and it shows me things I watched previously, and they'll show [videos they] recommended, so they're always doing that type of stuff." – DM7

DM participants often spoke about repeated interactions with these systems over time providing them insights about how they function. DM2 described her experience with a streaming music service presenting poor recommendations as a result of songs she "liked", leading her to an insight about how the system worked, and how to change her behavior to prevent inaccurate inferences from being made:

> "'Can't Touch This', right? It's that kind of song that [you think], 'Oh, isn't this a cool song?' And you like it. But when they refer songs [based on] that song, it's like, 'Oh, I shouldn't have liked it.' It's like, "Mm, they're going to do something with this, and they're probably going to refer to me stuff [based on] it.' And so I should be wise about what I like." – DM2

Their awareness about a behavioral basis for recommendations does not mean they had fully accurate beliefs about how recommendations are made. One misconception held by some participants in this cluster was that companies would rely only on behavioral data to make recommendations, to the exclusion of other data commonly used in recommendation systems, such as demographics. DM7, for example, believed that companies ignore his age when recommending products or other content:

> "Not too many websites really have shown me things based on my age group." – DM7

4.2.3 Companies use analytics to infer users' characteristics, with varying levels of humans-in-the-loop

Participants in the Data Mining cluster had a common element in their descriptions of how companies make inferences, in that they all had confidence that companies rely on some form of computer-based processing of data:

> "I'm sure there's some kind of algorithm out there, you know, I fall into a certain box, maybe I'm just a number with a letter at the end." – DM11

In that respect, they showed a more accurate perspective on modern inferencing methods than the Market Research cluster, who believed that companies collect the data they are interested in directly. Some participants in this cluster were aware that companies make inferences about them by analyzing their social connections, such as their friends on social media, in addition to their own behavior. However, the Data Mining cluster's other beliefs about inferencing often contained misconceptions about how companies rely on humans or computers to make inferences.

Despite this cluster's belief that computer processing of data is key to inferences, we observed a surprising diversity of beliefs about the role of human oversight in modern inferencing. Some thought that companies rely on automatic processes that make simple connections quickly. These participants talked in terms of computers establishing patterns in a person's behavior:

> "I'm thinking it is a machine scanning somebody's information and kind of learning and getting what they might be interested in or what their habit might be with something." – DM2

> "It makes inferences...I think it's just the computer doing [that], I don't think it's [people]...like keywords, just looking through that, I guess." – DM4

Others believed humans oversee the inferences made by algorithms, micromanaging the process. To these participants, computers are able to generate speculative inferences, but a human would make the final decision about whether an inference is accurate before using it, e.g., to make a recommendation. DM5 believed that humans closely supervise the implementation and results of inferencing programs, potentially leading to inaccuracies based on human judgment:

> "Even though it's electronically collected, electronically manipulated, it's looked at by a human. A human wrote the program. We're fallible." – DM5

Still others in the Data Mining cluster believed that companies rely on employees using computers as shallow tools to aid their own "reasonable skills of deduction", as DM11 put it. DM9 believed companies only use computers to generate visualizations of raw behavioral data, which human analysts would then review to make each inference about each individual user. He felt that companies use this process to determine a person's vulnerabilities, e.g., a person's values or attitudes that can be used to manipulate them via marketing or political appeals:

> "Certainly they've got to have analysts sitting there, you know, they hire interns to get on there and watch all this stuff, 'OK, now put it all down on a chart and show me where is he vulnerable and where is he not.'" – DM9

4.2.4 Mixed or negative feelings about inferencing are common

Participants in this cluster expressed complex feelings about companies making inferences from their behavioral data. This contrasts with some recent work suggesting the privacy calculus that people engage in is more visceral and gut-driven, rather than a purely rational accounting [29]. DM participants often described their use of online services as a trade-off, perceiving both benefits and drawbacks to using online services that rely on their data. DM8, a waste management driver with a keen awareness of behavioral tracking methods, spoke about his decision making process unambiguously, as "does the good outweigh the bad?". Others more broadly considered the purposes that behavioral inferences can serve, contrasting the use of data to save lives against using it for marketing:

> "If we're talking about harming mass quantities of people, like a 9/11 thing, then I'm all for collection of data. But if we're talking about, you know, you want to sell me a crib. {Laughs} Um, then I'm kind of against that." – DM5

Others felt torn as to whether benefits they accrue from inferencing are worth the costs:

> "They would try to tailor something for you specifically for your interest. So I guess one way to look at it is [as an] invasion of privacy and stuff like that. But the other way, you might say, 'Oh that [is] a little bit helpful.' So it's hard to tell." – DM4

Not all participants in this cluster saw advantages to being a part of the data economy. Several participants expressed resignation over their limited ability to control data collection, given that other people can provide data about them online without their consent. DM11, a store clerk in his 30s, was highly concerned about this. He lamented that despite taking strict action to pare down his data footprint by downgrading his smartphone to a feature phone and conscientiously managing his device's privacy settings, he is unable to prevent companies from collecting data about him through his friends' social media posts:

> "The things that I do in real time with real people, they possibly don't have very much access on my end from that. But I can't stop other people from posting things about me on Facebook, Twitter, et cetera." – DM11

We heard several in this cluster speak broadly about data mining as part of what they saw as a general decrease in personal privacy:

> "I'm uncomfortable with it. I didn't sign the deal with the devil basically, aside from hitting yes to a bunch of apps on my shiny, new tablet. Aside from that, I feel it is a great invasion of privacy." – DM11

> "The way everything seems to be going now, it almost seems like there's just less and less privacy...it's just kind of weird, feeling like people know certain things about you, you have no idea...all this information being gathered about you that you don't really know who they are." – DM6

4.3 Comparative analysis between clusters

In addition to the beliefs that defined each cluster, we found several apparent differences in demographics, attitudes, and sense of personal agency between the two clusters. We also include additional information on two misconceptions that spanned both clusters: that companies rely on humans rather than computers to make inferences, and that inferences are made on the basis of a single piece of information.

4.3.1 MR cluster more ethnically diverse; DM had higher income

Although educational attainment was similar across all of our participants, there were important demographic differences between the clusters even in this small sample, including income and ethnicity. Participants with household incomes over $45,000 were almost all in the DM cluster. Each ethnic group in our sample was represented in both clusters, but the MR cluster had a

greater proportion of ethnic minority participants compared to the DM cluster: 64% of the DM cluster identified as White or European-American, compared to 38% of the MR cluster. The MR cluster also perceived more stereotyping in companies' behavior, which we return to in the discussion below. Both clusters were roughly equally distributed in terms of age. Although the MR cluster believed that companies engage in older inferencing techniques than the DM cluster, we note that younger participants in the MR cluster had similar beliefs.

4.3.2 DM cluster more specific and confident in their beliefs, including their misconceptions

We observed recurring differences in how participants in each cluster described their beliefs. Compared to the MR cluster, participants in the DM cluster tended to speak more confidently about how they thought companies use their data. MR participants often went out of their way to describe their beliefs as speculative (e.g., "I don't know too much about it, but..." – MR4), but DM participants hedged fewer of their beliefs and misconceptions (e.g., "one way or another, you're being tracked...it happens everywhere" – DM8). Although the MR participants were missing an important piece of the inferencing landscape with regards to behavioral data collection, the DM participants' greater confidence in their misconceptions might be a more difficult obstacle to overcome. We saw hints of this in the teaching intervention section, as the participants with the most confident and specific beliefs during the interview section were often openly resistant to changing their beliefs in response to the learning activities.

4.3.3 MR cluster saw risks, DM cluster saw choices

Both clusters shared what they felt were drawbacks to data collection and inferencing, but they differed in the threats they described and the sense of risk or choice they felt they have in the data economy. We interpret these in light of the demographic differences between the clusters, and how they relate to cultural models of personal agency based on income, and cultural models of interacting with institutions based on ethnicity.

The MR cluster worried whether they are targeted by hackers and scammers, and they felt threatened by what they viewed as companies stereotyping. Many described taking protective measures to guard against what they felt were pervasive threats: financial threats from identity theft and ransomware, or threats to their sense of identity from companies treating them as a stereotype. The language they used often evoked a sense of being under attack, even when the danger was unclear, e.g., "I don't know how it works, but I know I just don't want to be a victim of it." (MR4). This was indicative of what appeared to be a lower sense of personal agency in the data economy in the MR cluster. Even though they felt that the methods companies use to make inferences were not far beyond their own capabilities, we often heard a clear protective motive behind the online behavior this cluster described. The MR participants did not describe trading their information to companies to gain a benefit; instead, they talked protectively about how they tried to prevent their information from being used against them.

The DM cluster had very different concerns and interpretations of their role in decisions about their data. In their view, companies largely provide opportunities for them to consciously trade their data (and by extension, their comfort) for material benefits. Companies appeared in some of their narratives as representing a more abstract threat to the concept of personal privacy, but even those participants felt they are choosing to pay the cost they must, to use the products and services they want:

> "I look at both sides of it and say, 'Well, would I rather do this or would I rather do this?' So if it's not hurting anything, and it could help, then I'm fine with it." – DM8

Somewhat paradoxically, although the MR cluster was more convinced that companies collect their data by explicitly asking them to provide it, the DM cluster seemed to feel more agency and control over the decisions they make online with their data. This difference in perspective may relate to cultural models of agency that differ based on social class, as the DM cluster had higher incomes overall than the MR cluster. We discuss further implications for this finding in the discussion section.

4.3.4 Humans-in-the-loop, and weak correlations in modern inferencing

All of the MR cluster and several in the DM cluster believed that companies make inferences about users by having humans-in-the-loop, either relying on human analysts who exploit common sense connections between two pieces of data (e.g., inferring hobbies from purchases), or by employing experts who analyze an individual's behavior like a detective (e.g., manually combing a user's online pictures for evidence of a spouse). These folk explanations for how companies make inferences exaggerate companies' capabilities in some ways while limiting users' ability to imagine current inferencing methods in others.

Believing that companies rely on common sense logic to tie two data points together ignores the multivariate, deep learning methods that companies now deploy to make unintuitive inferences. To our participants, inferences seemed to be snap judgments based on perfect, intuitive correlations between two pieces of data. This may lead to unpleasant surprises when encountering systems that make unintuitive predictions based on data aggregated from multiple sources. At the same time, the belief that companies employ a team of human experts to diligently analyze each user's data may be partially responsible for some of our users believing that there is no limit to what companies can learn about them. The belief that employees with strong detective skills are hunting down their data may lead some participants to misjudge the risks attached to making different types of data available online for companies to see.

5. DISCUSSION

The patterns we observed in our high school-educated sample's beliefs and misconceptions about companies' inferencing methods underscore the need for privacy researchers to consider qualitative, cultural influences on privacy knowledge, attitudes, and behavior. We share two categories of implications that came out of this work: implications of inference literacy in research, design, and education; and implications of cultural models for future research in online privacy and HCI in general.

5.1 Implications of inference literacy

5.1.1 Redefining tech savviness and digital literacy

As technology itself changes, definitions of tech savviness and digital literacy need to change to stay up-to-date. Measuring tech savviness by the ability to perform instrumental tasks on a local device ignores the extent to which daily device activity takes place in a distributed, online context. Although recent attempts to assess online privacy literacy have gone beyond that to include

aspects of how online institutions collect or transmit data [57], digital literacy studies often still rely on self-reported expertise [5,20] or the number of years using the internet [44] as a primary measure of digital literacy. The current study is among the first, to our knowledge, to directly explore this particular aspect of digital literacy: beliefs users have about how companies make inferences from their data.

Our results suggest that inference literacy is worth including as an aspect of digital and online privacy literacy. The overwhelming majority of our participants use multiple devices everyday for various purposes, but they had several misconceptions about current methods of data collection and inferencing that could lead to unpleasant surprises. We advocate for a broadening of the features used to consider what makes a person tech savvy or digitally/privacy literate to include basic inference literacy: (1) that companies can collect and aggregate data from multiple sources including forms, behavioral data, telemetry, and public databases, and (2) that those data are often processed by learning algorithms that can uncover unintuitive or even private connections that can be found due to the massive amount of data available to companies.

5.1.2 The roles of transparency and education in inference literacy

Our participants were active online users who, in the absence of structures to help them build their inference literacy, developed lay theories to explain their online experiences that contained basic misconceptions. We believe this speaks to a need for interventions to support inference literacy, and we discuss potential inference literacy interventions: transparency, as well as informal and formal educational interventions.

First, we consider the issue of transparency. Transparency can inform individuals and surface issues for broader discussion about systemic and policy issues [48]. Users may, for instance, be more comfortable knowing a system does not have humans-in-the-loop when sensitive data are involved, or they may prefer to have humans-in-the-loop if they feel a human could outperform an algorithm. However, transparency is not a silver bullet. It places a heavy burden on the user to learn about the algorithms of each company they engage with, and complex inferential systems are often black boxes even to those who design and deploy them [27]. It may not be feasible to be transparent about inferencing systems that change frequently, and whose true workings require sustained academic research to discover.

Second, there may be a role for informal, "do-it-yourself" interventions that allow users to teach themselves inference literacy concepts [38], such as that aggregating multiple sources of data allows companies to learn unintuitive, weak correlations between data. There are existing resources related to inference literacy that could be used as models for novel interventions. Teachingprivacy.org [69] offers a selection of accessible lessons about online privacy that draw from real world examples, including structured material for deployment by teachers in formal educational settings. Other efforts like R2D3's "A Visual Introduction to Machine Learning" [67] provide more technical knowledge about statistical classification methods. These approaches provide motivated users with resources to correct their misconceptions, but we caution against the idea that these methods will systemically improve inference literacy. Research has shown that relying on "do-it-yourself" approaches to build technology knowledge and skills may widen rather than reduce inequalities in digital literacy [38]. This may be, in part, due to a discoverability issue as a result of jargon used in some informal interventions. Nineteen of our 21 participants had never heard the term "machine learning" prior to the study, which may make it harder for them to find resources like R2D3's.

Finally, we point out that regardless of societal aspirations to increase access to a college education, high school education is still the modal educational attainment in the US. Students graduating with a high school degree should be prepared for more than just college academics; they should also be prepared to live in a world where interactions with inferential systems are common and consequential. Our participants' beliefs were outdated in several crucial respects. The frequent appearance of misconceptions that companies rely on consumers taking surveys to gather demographic data, or on humans-in-the-loop rather than automated analytics to make recommendations, speaks to the danger of assuming that users will osmose the basics of the consumer data ecosystem outside of a formal educational setting. Requiring college or independent study to learn about how personal data may be used to infer a credit score, decide on a loan application, or other important aspects of economic life places that knowledge outside the reach of those who are most likely to be negatively affected by those decisions [14,43]. There is precedent for teaching digital literacy concepts [68] and personal finance [10] at the high school level, and inference literacy is worth considering alongside these topics.

5.2 Implications of cultural models

The current study surfaced several issues related to power and privilege in consumer interactions, which we describe in terms of *cultural models*, sets of assumptions that differ across cultural groups. We share two main insights here: first, that our participants' experiences of risk and choice in online privacy and security relate to cultural models about personal agency that differ by income; second, that our participants from marginalized ethnic groups believed companies' inferencing methods constitute stereotyping, which we link to broader work on ethnic minority experiences in consumer settings. We describe implications of these findings for online privacy research and design, and finish by advocating more broadly for consideration of cultural models as a key lens to critically examine the experiences of marginalized groups in user research.

5.2.1 Income and differences in inference literacy

Educational attainment and income are often treated as equivalent indicators of socioeconomic status, but we saw differences in beliefs and attitudes within our education-controlled sample based on participant income. First, our higher-income participants viewed online privacy decisions as choices they were empowered to make, whereas our lower-income participants framed those same decisions as risks they had to protect against. This echoes prior work showing that middle-class Americans typically develop a sense of agency built around exercising independent choices, whereas working-class Americans often experience greater economic and environmental constraints that preclude such free choice behavior [54,56]. Recently, some inferencing systems have been designed to allow users the ability to modify their workings, either to improve system accuracy or simply to exercise personal choice over their outputs [31,61]. We suspect that users' interactions with these systems may be affected by the larger cultural context in which those choices are made, and we advise system designers to consider how differences in risk

perception and personal agency based on cultural differences may affect users' willingness to engage with different system designs.

Second, prior work has found that inequalities in online skills and knowledge often result from differences in SES [20,24,64], and we found a similar, problematic inference literacy gap related to SES. The two different clusters of beliefs we saw were linked to income differences: the MR cluster had nearly all of our participants with under $45,000 household annual income, and was less aware of current data collection and inferencing practices. Although we cannot say whether this trend in our sample is representative of one in the larger population, given that inferential systems already disproportionately negatively affect working-class people [14,43], we again highlight the need for systemic efforts to prevent and reduce digital inequalities, including those related to inference literacy.

5.2.2 Ethnicity and interpretations of inferencing as stereotyping

Several participants spontaneously brought up beliefs that companies stereotype consumers by ethnicity, all of whom claimed that doing so is immoral. It is undoubtedly true that companies use demographics to profile users, and that this is an inherently imprecise process. Modern inferencing systems may include demographics like ethnicity among many features, but these participants believed that inferences are sometimes made based *only* on assumptions about ethnicity. However accurate or inaccurate this belief about stereotyping is, it remains that these participants' life experiences have resulted in a cultural model about interactions with institutions like companies that assumes companies stereotype.

The complex online ecosystem our study explored is often ambiguous as to how decisions are made: the opacity of algorithms that recommend, advertise, or filter content that users see often means users generally lack context for how online companies' decisions and recommendations are made. This leaves plenty of room for the user to interpret online experiences in light of other experiences they have had, including those of being stereotyped or discriminated against. To the user who has experienced discrimination while shopping [15,34,50], driving [37], or merely walking [16], stereotyping by online companies may appear no different. Designers should therefore take caution in how they include or describe ethnicity as a component of decision-making about users. Lacking clear evidence to the contrary, unflattering interpretations may be made about inferential systems for which the purpose of using demographics like ethnicity is left ambiguous to the user.

5.2.3 Cultural models in user research and design

In this work, we applied the concept of cultural models to describe additional layers of commonalities and differences across our participants' experiences. Although the finding that our two clusters had different views on risk and choice online might stand on its own, incorporating cultural models allowed us to link this finding to different beliefs about personal agency that relate to social class rather than leaving our analysis at the level of the individual participant. This provided us insight into a mechanism that seems to underlie interpretations about online privacy consequences, one that speaks to different economic and environmental constraints between cultural groups. We believe that exploring the ways that cultural models qualitatively affect people's interpretations and attitudes about online phenomena complements other user research approaches by providing a textured, layered perspective on the meaning that users attach to their online privacy experiences.

5.3 Future Directions

We advocate for further research on inference literacy in high school- and college-educated samples to confirm whether the belief clusters we observed exist in the larger population, as well as to further explore whether inference literacy varies by educational attainment or geographic location. We also endorse the adoption of cultural models as a useful lens to apply to other research in online privacy. It would also be valuable to further explore how inference literacy beliefs interact with participants' online behavior and decision-making processes, in order to inform new system designs that can better support inference literacy.

5.4 Conclusion

We began this work by describing the vast difference between companies' past and present inferencing methods. There is little reason to believe that current methods will remain static, but our findings suggest that there is already a substantial gap between what people believe companies are doing with their data, and the current reality of pervasive, automatic algorithms. We point not only to the size of that gap, but also to its heterogeneity: we saw links between inference literacy beliefs and larger cultural models based on income, ethnicity, and potentially educational attainment. Culturally sensitive policy, research, and design may be a route to minimizing digital inequalities that arise as an outcome of group differences in inference literacy.

6. ACKNOWLEDGMENTS

We thank our participants for their contributions to this research. We also thank Sunny Consolvo and Tara Matthews for their valuable guidance. We express our gratitude to Paul Aoki, Erik Ninomiya, Katie O'Leary, Sai Teja Peddinti, Shepherd Pittman, Sonam Samat, Martin Shelton, and Irene Tang for their helpful feedback and assistance.

7. REFERENCES

[1] Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). "Privacy and human behavior in the age of information," *Science, 347*(6221), 509–514.

[2] Alpaydin, E. (2014). *Introduction to Machine Learning,* MIT Press.

[3] Aronson, J., & Inzlicht, M. (2004). "The ups and downs of attributional ambiguity stereotype vulnerability and the academic self-knowledge of African American college students," *Psychological Science, 15*(12), 829–836.

[4] Awad, M., & Khanna, R. (2015). "Machine learning and knowledge discovery," in *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, Berkeley, CA: Apress.

[5] Bartsch, M., & Dienlin, T. (2016). "Control your Facebook: An analysis of online privacy literacy," *Computers in Human Behavior, 56*, 147–154.

[6] Berkovsky, S., & Freyne, J. (2010). "Group-based recipe recommendations: analysis of data aggregation strategies," *In Proceedings of the 4th ACM Conference on Recommender Systems: RecSys '10*, 111–118.

[7] Bi, B., Shokouhi, M., Kosinski, M., & Graepel, T. (2013). "Inferring the demographics of search users: Social data meets search queries," *In Proceedings of the 22nd International Conference on World Wide Web: WWW '13*, 131–140.

[8] Braun, V., & Clarke, V. (2006). "Using thematic analysis in psychology," *Qualitative Research in Psychology, 3*(2), 77–101.

[9] Choi, B., Lee, I., Kim, J., & Jeon, Y. (2005). "A qualitative cross-national study of cultural influences on mobile data service design," *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: CHI '05*, 661–670.

[10] Council for Economic Education. (2016). "Survey of the states: Economic and personal finance education in our nation's schools", Retrieved from Council for Economic Education: http://councilforeconed.org/wp/wp-content/uploads/2016/02/sos-16-final.pdf

[11] D'Andrade, R. (1995). *The Development of Cognitive Anthropology,* New York, NY: Cambridge University Press.

[12] van Deursen, A. J. A. M., van Dijk, J. A. G. M., & ten Klooster, P. M. (2015). "Increasing inequalities in what we do online: A longitudinal cross sectional analysis of Internet activities among the Dutch population (2010 to 2013) over gender, age, education, and income," *Telematics and Informatics, 32*(2), 259–272.

[13] Dinev, T., Masssimo, B., Hart, P., Christian, C., Vincenzo, R., & Ilaria, S. (2005). "Internet users, privacy concerns and attitudes towards government surveillance: An exploratory study of cross-cultural differences between Italy and the United States," *In Proceedings of Bled: BLED '05*.

[14] Fourcade, M., & Healy, K. (2013). "Classification situations: Life-chances in the neoliberal era," *Accounting, Organizations and Society, 38*(8), 559–572.

[15] Gabbidon, S. L. (2003). "Racial profiling by store clerks and security personnel in retail establishments: An exploration of 'shopping while Black,'" *Journal of Contemporary Criminal Justice, 19*(3), 345–364.

[16] Gelman, A., Fagan, J., & Kiss, A. (2007). "An analysis of the New York City Police Department's 'stop-and-frisk' policy in the context of claims of racial bias," *Journal of the American Statistical Association, 102*(479), 813–823.

[17] Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011). "Predicting personality from Twitter," *In Proceedings of Privacy, Security, Risk and Trust and International Conference on Social Computing: PASSAT/SocialCom '11,* 149–156.

[18] Gou, L., Zhou, M. X., & Yang, H. (2014). "KnowMe and ShareMe: understanding automatically discovered personality traits from social media and user sharing preferences," *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: CHI '14*, 955–964.

[19] Harboe, G., & Huang, E. M. (2015). "Real-world affinity diagramming practices: Bridging the paper-digital gap," *In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems: CHI '15,* 95–104.

[20] Hargittai, E., & Hinnant, A. (2008). "Digital inequality: Differences in young adults' use of the internet," *Communication Research, 35*(5), 602–621.

[21] Heimgärtner, R. (2013). "Reflections on a model of culturally influenced human–computer interaction to cover cultural contexts in HCI design," *International Journal of Human-Computer Interaction, 29*(4), 205–219.

[22] Helgeson, J. G., Kluge, E. A., Mager, J., & Taylor, C. (1984). "Trends in consumer behavior literature: A content analysis," *Journal of Consumer Research*, *10*(4), 449–454.

[23] Holland, D., & Quinn, N., eds. (1987). *Cultural Models in Language and Thought,* New York, NY: Cambridge University Press.

[24] Hsieh, J. J. P.-A., Rai, A., & Keil, M. (2008). "Understanding digital inequality: Comparing continued use behavioral models of the socio-economically advantaged and disadvantaged," *MIS Quarterly, 32*(1), 97–126.

[25] Ion, I., Reeder, R., & Consolvo, S. (2015). "'... no one can hack my mind': Comparing Expert and Non-Expert Security Practices," *In Proceedings of Symposium On Usable Privacy and Security: SOUPS '15*, 327–346.

[26] Kang, R., Dabbish, L., Fruchter, N., & Kiesler, S. (2015). "'My data just goes everywhere:' User mental models of the internet and implications for privacy and security," *In Proceedings of Symposium On Usable Privacy and Security: SOUPS '15,* 39–52.

[27] Kapoor, A., Lee, B., Tan, D., & Horvitz, E. (2010). "Interactive optimization for steering machine classification," *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: CHI '10*, 1343–1352.

[28] Kauer, M., Günther, S., Storck, D., & Volkamer, M. (2013). "A comparison of American and German folk models of home computer security," *In Proceedings of Human Aspects of Information Security, Privacy, and Trust: HAS '13*, 100–109.

[29] Kehr, F., Kowatsch, T., Wentzel, D., & Fleisch, E. (2015). "Blissfully ignorant: The effects of general privacy concerns, general institutional trust, and affect in the privacy calculus" *Information Systems Journal, 25*(6), 607–635.

[30] Kosinski, M., Stillwell, D., & Graepel, T. (2013). "Private traits and attributes are predictable from digital records of human behavior," *PNAS, 110*(15), 5802–5805.

[31] Kulesza, T., Stumpf, S., Burnett, M., & Kwan, I. (2012). "Tell me more?: The effects of mental model soundness on personalizing an intelligent agent," *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: CHI '12*, 1–10.

[32] Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., & Wong, W.-K. (2013). "Too much, too little, or just right? Ways explanations impact end users' mental models," *In Proceedings of IEEE Symposium on Visual Languages and Human Centric Computing: VL/HCC'13*, 3–10.

[33] Kurasaki, K. S. (2000). "Intercoder reliability for validating conclusions drawn from open-ended interview data," *Field methods, 12*(3), 179–194.

[34] Lee, J. (2000). "The salience of race in everyday life: Black customers' shopping experiences in Black and White neighborhoods," *Work and Occupations, 27*(3), 353–376.

[35] Lewenberg, Y., Bachrach, Y., & Volkova, S. (2015). "Using emotions to predict user interest areas in online social networks," *In Proceedings of IEE International Conference on Data Science and Advanced Analytics: IEEE DSAA '15*.

[36] Li, Y. (2012). "Theories in online information privacy research: A critical review and an integrated framework," *Decision Support Systems, 54*(1), 471–481.

[37] Lundman, R. J., & Kaufman, R. L. (2003). "Driving while Black: Effects of race, ethnicity, and gender on citizen self-reports of traffic stops and police actions," *Criminology, 41*(1), 195–220.

[38] Matzat, U., & Sadowski, B. (2012). "Does the 'do-it-yourself approach' reduce digital inequality? Evidence of self-learning of digital skills," *The Information Society, 28*(1), 1–12.

[39] McDonald, A. M., & Cranor, L. F. (2010). "Americans' attitudes about internet behavioral advertising practices," *In Proceedings of the 9th Annual ACM Workshop on Privacy in the Electronic Society: WPES '10*, 63–72.

[40] Miller, P. J., Cho, G. E., & Bracey, J. R. (2005). "Working-class children's experience through the prism of personal storytelling," *Human Development, 48*(3), 115–135.

[41] Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective,* Cambridge, MA: MIT Press.

[42] Oakleaf, M. (2009). "The information literacy instruction assessment cycle: A guide for increasing student learning and improving librarian instructional skills," *Journal of Documentation, 65*(4), 539–560.

[43] Pager, D., & Shepherd, H. (2008). "The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets," *Annual Review of Sociology, 34*, 181.

[44] Park, Y. J. (2011). "Digital literacy and privacy behavior online," *Communication Research*, *40*(2), 215-236.

[45] Park, Y. J. (2013). "Offline status, online status: Reproduction of social categories in personal information skill and knowledge," *Social Science Computer Review, 31*(6), 680–702.

[46] Pearce, K. E., & Rice, R. E. (2013). "Digital divides from access to activities: Comparing mobile and personal computer internet users", *Journal of Communication, 63*(4), 721–744.

[47] Pearce, R. R. (2006). "Effects of cultural and social structural factors on the achievement of White and Chinese American students at school transition points," *American Educational Research Journal, 43*(1), 75–101.

[48] Rader, E. (2014). "Awareness of behavioral tracking and information privacy concern in Facebook and Google," *In Proceedings of Symposium On Usable Privacy and Security: SOUPS '14.*

[49] Rader, E., Wash, R., & Brooks, B. (2012). "Stories as informal lessons about security," *In Proceedings of Symposium On Usable Privacy and Security: SOUPS '12*.

[50] Schreer, G. E., Smith, S., & Thomas, K. (2009). "'Shopping while Black': Examining racial discrimination in a retail setting," *Journal of Applied Social Psychology, 39*(6), 1432–1444.

[51] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PLoS ONE, 8*(9).

[52] Sellers, R. M., & Shelton, J. N. (2003). "The role of racial identity in perceived racial discrimination," *Journal of Personality and Social Psychology, 84*(5), 1079–1092.

[53] Smith, A. (2015). "U.S. smartphone use in 2015," Pew Research Center: Internet, Science & Tech. Retrieved from Pew Research Center: http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/

[54] Snibbe, A. C., & Markus, H. R. (2005). "You can't always get what you want: Educational attainment, agency, and choice," *Journal of Personality and Social Psychology, 88*(4), 703–720.

[55] Steele, C. M., & Aronson, J. (1995). "Stereotype threat and the intellectual test performance of African Americans," *Journal of Personality and Social Psychology, 69*(5), 797.

[56] Stephens, N. M., Fryberg, S. A., & Markus, H. R. (2012). "It's your choice: How the middle-class model of independence disadvantages working-class Americans," *Facing Social Class: How Societal Rank Influences Interaction*, New York, NY: Russell Sage Foundation.

[57] Trepte, S., Teutsch, D., Masur, P. K., Eicher, C., Fischer, M., Hennhöfer, A., & Lind, F. (2015). "Do people know about privacy and data protection strategies? Towards the 'Online Privacy Literacy Scale'(OPLIS)," *Reforming European Data Protection Law*, Dordrecht, NL: Springer, 333–365.

[58] Ur, B., Leon, P. G., Cranor, L. F., Shay, R., & Wang, Y. (2012). "Smart, useful, scary, creepy: Perceptions of online behavioral advertising," *In Proceedings of Symposium On Usable Privacy and Security: SOUPS '12*.

[59] Ur, B., & Wang, Y. (2013). "A cross-cultural framework for protecting user privacy in online social media," *In Proceedings of the 22nd International Conference on World Wide Web: WWW '13*, 755–762.

[60] U.S. Census Bureau. (2015). Educational Attainment in the United States: 2014 - Detailed Tables. Retrieved from U.S. Census Bureau: http://www.census.gov/hhes/socdemo/education/data/cps/2014/tables.html

[61] Warshaw, J., Matthews, T., Whittaker, S., Kau, C., Bengualid, M., & Smith, B. A. (2015). "Can an algorithm know the 'real you'?: Understanding people's reactions to hyper-personal analytics systems," *In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems: CHI '15*.

[62] Wash, R. (2010). "Folk models of home computer security," *In Proceedings of the Symposium on Usable Privacy and Security: SOUPS '10*.

[63] Wash, R., & Rader, E. (2015). "Too much knowledge? security beliefs and protective behaviors among United States internet users," *In Proceedings of the Symposium on Usable Privacy and Security: SOUPS '15*.

[64] Wei, L. (2012). "Number matters: The multimodality of internet use as an indicator of the digital inequalities," *Journal of Computer-Mediated Communication, 17*(3), 303–318.

[65] Wei, L., & Hindman, D. B. (2011). "Does the digital divide matter more? Comparing the effects of new media and old media use on the education-based knowledge gap," *Mass Communication and Society, 14*(2), 216–235.

[66] Weinsberg, U., Bhagat, S., Ioannidis, S., & Taft, N. (2012). "BlurMe: Inferring and obfuscating user gender based on ratings," *In Proceedings of the 6th ACM Conference on Recommender Systems: RecSys '12*, 195–202.

[67] Yee, S., & Chu, T. (2015). "A Visual Introduction to Machine Learning". Retrieved from R2D3: http://www.r2d3.us/visual-intro-to-machine-learning-part-1/.

[68] "DigitalLiteracy.gov: Your destination for digital literacy resources and collaboration," Retrieved from National Telecommunications and Information Administration: http://www.digitalliteracy.gov/

[69] "Teaching Privacy". Retrieved from International Computer Science Institute: http://teachingprivacy.org/

Appendix

Appendix A. Teaching intervention procedure

The teaching intervention included a pre-test assessment; two interventions to teach inference literacy concepts, each followed by a card sorting task to assess users' developing explanations of inferencing phenomena; and a post-test assessment to gauge changes in inferencing beliefs after the interventions.

The **pre-test assessment** consisted of one main prompt and follow-up questions about the inferencing capabilities of a social media company.

Next, we provided the **first teaching intervention**, in which the interviewer explained that companies may collect behavioral data while people use a device. Following this first intervention, the interviewer gave the participant the **first card sorting task**, in which participants ranked the likelihood that a given inference could be drawn from a particular piece of data, e.g., "Data: List of apps on phone, Inference: Whether they have kids". These inferences were chosen because they could be made intuitively by people or by an algorithm, allowing us to learn which explanation participants gravitated towards. Afterwards, the interviewer provided feedback on which inferences are or are not likely to be possible for companies to make using current technology.

Next, we provided the **second teaching intervention**, sharing a simplified explanation of classification through machine learning. After the second teaching intervention, we provided the **second card sorting task**, with inferences chosen to explore capabilities related to classification, e.g., "Data: Text from social media posts, Inference: Their personality" [14]. We hoped that after the explanation of classification, participants' explanations would include details of the machine learning process. Again, the interviewer provided feedback on the feasibility of each inference.

Finally, the interviewer administered a **post-test assessment**, asking about the inferencing capabilities of a cell phone service provider. We finished the session by soliciting feedback on the teaching intervention, which we used to refine materials between cycles.

Forgetting in Social Media: Understanding and Controlling Longitudinal Exposure of Socially Shared Data

Mainack Mondal
MPI-SWS
mainack@mpi-sws.org

Johnnatan Messias
MPI-SWS
johnme@mpi-sws.org

Saptarshi Ghosh
IIEST Shibpur
sghosh@cs.iiests.ac.in

Krishna P. Gummadi
MPI-SWS
gummadi@mpi-sws.org

Aniket Kate
Purdue University
aniket@purdue.edu

ABSTRACT

On most online social media sites today, user-generated data remains accessible to allowed viewers unless and until the data owner changes her privacy preferences. In this paper, we present a large-scale measurement study focussed on understanding how users control the longitudinal exposure of their publicly shared data on social media sites. Our study, using data from Twitter, finds that a significant fraction of users withdraw a surprisingly large percentage of old publicly shared data—more than 28% of six-year old public posts (tweets) on Twitter are not accessible today. The inaccessible tweets are either selectively deleted by users or withdrawn by users when they delete or make their accounts private. We also found a significant problem with the current exposure control mechanisms – even when a user deletes her tweets or her account, the current mechanisms leave traces of residual activity, i.e., tweets from *other* users sent as replies to those deleted tweets or accounts still remain accessible. We show that using this residual information one can recover significant information about the deleted tweets or even characteristics of the deleted accounts. To the best of our knowledge, we are the first to study th information leakage resulting from residual activities of deleted tweets and accounts. Finally, we propose an exposure control mechanism that eliminates information leakage via residual activities, while still allowing meaningful social interactions with user posts. We discuss its merits and drawbacks compared to existing mechanisms.

1. INTRODUCTION

"every young person one day will be entitled automatically to change his or her name on reaching adulthood in order to disown youthful hijinks stored on their friends' social media sites". – Eric Schmidt [14]

The unprecedented sharing of personal, user-generated content on online social media sites like Twitter and Facebook has spawned numerous privacy concerns for the users of the sites [5, 6, 10, 13, 16, 24]. In this paper, we focus on a dimension of user privacy that becomes more challenging to manage with the passage of time, namely, *longitudinal privacy*. Users' privacy preferences for sharing content are known to evolve over time [5, 6]. There can be many reasons for such temporal changes in privacy preferences – e.g., the sensitivity or relevance of shared content changes with time; the biographical status of users and their friend relationships change over time. The challenge of managing longitudinal privacy for a user refers to the difficulty in *controlling the exposure of the user's socially shared data over time*. This challenge becomes more complex over time as the set of contents shared in the past grows larger and new technologies like archival (timeline-based) searches make it easier to access historical content shared under outdated privacy preferences.

Against this background, this paper asks and investigates the following *two* foundational questions related to *understanding* and *controlling* longitudinal exposure of user data in social media sites, respectively:

1. In practice, is there evidence for users changing their privacy preferences for content shared on social media sites 5 to 10 years in the past? If so, what is the extent of the change in longitudinal exposure of user data?

2. In practice, how effective are the mechanisms provided by social media sites to enable users to control the exposure of their shared data over time? Could we improve the effectiveness of longitudinal exposure control mechanisms?

To address these questions, we have gathered extensive longitudinal data (over 6 years) from the Twitter social media site. Compared to the Facebook social networking site, the privacy preferences of users for messages (tweets) posted (tweeted) in Twitter are relatively simple – each tweet is either publicly visible to everyone, or privately visible only to the user's followers, or deleted from the site by the user. However, the simplicity of privacy choices in Twitter allows us to measure the temporal evolution of their users' privacy preferences by simply tracking the public visibility of users' tweets over time.

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22–24, 2016, Denver, Colorado.

Our analysis of Twitter messages[1] reveals striking evidence of a significant fraction (∼35%) of all Twitter users changing their privacy preferences over time. Only a minority (∼8%) of all Twitter users selectively withdraw (i.e., delete or make them private) a small (∼10%) fraction of all their public posts. On the other hand, a sizeable fraction (∼27%) of all Twitter users withdraw all of their public posts older than a few (4-6) years. While a few recent studies have attempted to understand how user's privacy preferences might change with time through user surveys [5, 6], to our knowledge, our work presents the first large-scale measurement study of how users actually change their privacy preferences in practice. Since our exploration is data driven (as opposed to user surveys), we could not investigate the user intentions behind the changes in privacy preferences. A limitation of our work lies in the assumption that these changes are driven by users' privacy concerns.

Our investigation of the effectiveness with which Twitter users control the public exposure of their tweets reveals a fundamental problem. Even after a user withdraws her public posts, the past interactions of her friends and other users with those posts (by the way of comments and replies) leave a trail of residual posts that remain on the site (as the residual posts are not authored by the same user, they cannot be withdrawn by her). We show that these residual activities are in many cases sufficient to recover significant amounts of information about the withdrawn posts. Our analysis of residual activities highlights this inherent flaw with the longitudinal exposure controls currently being provided to Twitter users. To make users more aware of the flaws in the existing exposure control mechanisms, we also design a Twitter app, deployed at `http://twitter-app.mpi-sws.org/footprint/`, where any one can login with their Twitter account and check the residual activities around their posts.

Having identified the limitations of existing longitudinal exposure controls, we discuss *why* devising a perfect solution to control longitudinal exposure is extremely difficult. Then we present an investigation into merits and drawbacks of a few advanced longitudinal exposure control mechanisms. Specifically, we focus on the recent trend towards ephemeral posts in new social media sites like Snapchat, where every post is timed to be deleted once it reaches a pre-set age (expiry time). The challenge with such ephemeral posts, however, lies in determining the "correct" pre-set deadlines for post deletion. We show that a different approach, where a post is deleted based on a pre-set duration of inactivity, offers users comparatively better control over their longitudinal privacy.

2. RELATED WORK

In this section we explore the related work in this space along three axes.

Are users concerned about privacy of their old data? Understanding and improving privacy control in online social media sites garnered quite a bit of attention in recent times [5–8, 10, 11, 13, 16, 19, 20, 22, 24, 28]. The focus of these studies range from identifying regrettable / deletable content, to understanding the usage of privacy management mechanisms for sharing data, to designing better privacy management tools. However, there has been relatively little research on exploring the longitudinal privacy management mechanisms. Two recent studies [5,6] surveyed tens to hundreds of users to explore how online social media users want to manage their longitudinal privacy for old content uploaded in the recent past (last week, month, year). The study in [5] performed a user survey and found that a user's willingness to share content drops as the content becomes old. Moreover, willingness of share further decreases with a life-change, e.g., graduating from college or moving to a new town. The other study [6] performed two surveys and discovered that users want some old posts to become more private over time and their desired exposure set for the content remained relatively constant over the years. Both of these studies indicate that users are, in general, concerned about the privacy of their old content, possibly because these content do not reflect who they are at present (possibly after a change in life). Hence, these studies provide a strong motivation for us to study at large scale how users in the real-world behave to address their privacy concerns.

How do users control longitudinal exposure of their old data? One natural way for a user to protect her longitudinal privacy is to delete her old content. Some recent studies have focused on content deletion by users. For instance a PEW survey [18] on 802 teenagers found that 59% of respondents edited or deleted their content in OSNs. Almuhimedi *et al.* [4] reported the largest study so far on deleted tweets using real world data, however they only collected data which are deleted at most one week after posting. Specifically, they collected 67 million tweets from 292K users posted during a week, and found that 2.4% of those tweets are deleted within that week. Out of their set of deleted tweets, 89.1% were deleted on the same day on which they were posted. Moreover 17% of those deleted tweets were removed by the user due to typos or to rephrase the same tweet. However, note that, they primarily focused on content posted in the near past (no more than one week old) which were selectively deleted by the user. We will report later in this study how the exposure controls are quite different for the content posted in the near and far past, and show that the study [4] missed a large part of deleted tweets posted in far past (e.g., 6 years back).

A few other studies [12, 17] explored the changing behavior of Twitter users over time. Out of them, Liu *et al.* [17] analyzed the collective tweeting behavior over time including deletion of content. They observed that social media users are either selectively deleting their tweets or deleting their entire account. However, they did not check if there are limitations of these mechanisms to control exposure. Neither did they explore the relative merits and drawbacks of different exposure control mechanisms. We explore these unanswered questions in detail.

What are some proposed mechanisms to help users control longitudinal exposure? Some recent studies mentioned possible mechanisms to improve the usability of longitudinal privacy mechanisms in OSNs. Bauer *et al.* [6] observed that users are possibly becoming more privacy-aware about their longitudinal data. This change in users' privacy concerns is further reflected by the advent and pop-

[1]This study was conducted respecting the guidelines set by our institute's ethics board and with their explicit knowledge and permission.

ularity of systems like Snapchat [2] which deletes all users' posts after a predefined expiry time. Aylan and Toch [5] proposed longitudinal privacy management mechanisms like allowing users to set expiration dates on content or having an archive feature for old content. We build upon these studies and propose a smart policy for content withdrawal, which dynamically tries to decide which content to delete or archive based on its longitudinal exposure.

3. UNDERSTANDING LONGITUDINAL EXPOSURE

In this section, we aim to understand how users are presently withdrawing their socially shared content to control longitudinal exposure. We start by answering the simple question – *what are the longitudinal exposure control mechanisms available today in Twitter, for withdrawing shared content?*

3.1 Exposure controls in Twitter

We found three distinct mechanisms of withdrawing socially shared content (tweets) in Twitter today:

1. Withdrawing tweets via selective deletion: The reasons for such deletion ranges from regrettable content in the tweets to simply correcting typographical errors or rephrasing [4].

2. Withdrawing tweets via deleting account: All tweets posted by a user can be withdrawn by deleting her whole account.

3. Withdrawing tweets via making account private: In Twitter, user-accounts are either 'public' or 'private'. Tweets posted by a public account are visible to anyone online, but tweets posted by a private account are visible to only the followers of that account, who must be approved by the private account owner before they can be a follower. Unlike Facebook, Twitter does *not* have sophisticated access control mechanisms whereby a tweet can be made visible to only a subset of one's followers. In Twitter, a tweet is either public to all users, or at least to all followers of the user who posted the tweet. Thus, if a user makes her account 'private', all tweets posted from this account are no longer available publicly.

Note that there is another factor that will result in tweets becoming inaccessible – if Twitter suspends a user's account for violating their terms of service, all tweets posted by that account will became inaccessible. However, we do not consider this factor as a mechanism for exposure control, since suspension is not carried out by the user herself.

To perform this study at scale, we needed to identify a large set of tweets that have been withdrawn by Twitter users. Additionally, we also needed to ascertain *why* a tweet has become inaccessible, so that we can ignore tweets that have become inaccessible due to Twitter suspending the users, and focus only on tweets that have been withdrawn by the users themselves. The rest of this section describes how we identified such tweets.

Methodology for identifying tweets withdrawn by users: Our methodology consisted of taking a large set of tweets posted and archived in the past, and checking which ones have become inaccessible at the time of this study (October 2015). We observed that if we query the Twitter API with a tweet-id (a Twitter-generated unique identifier for a tweet) that was archived in the past when the tweet was public, if the tweet is inaccessible at present, the Twitter API sends back an error code and an error message as explanation. These error codes are customized by Twitter and are different from the normal HTTP error codes 404 (resource not found) and 403 (access forbidden) that are also obtained during this querying process. During our experiments consisting of querying for millions of tweet-ids (details given later), we noticed four distinct error codes that are shown in Table 1, along with the corresponding HTTP error codes, the corresponding error messages, and the practical interpretation of the error codes. These practical interpretations are based on the Twitter error messages and experiments performed using one of the author's Twitter account (as described below).

Twitter error codes	**Corresponding HTTP error codes**	**Twitter error message**	**Practical interpretation of Twitter error codes**
179	403	Sorry, you are not authorized to see this status	User account made private
63	403	User has been suspended	User account suspended by Twitter
34	404	Sorry, that page does not exist.	Tweet (or user account) withdrawn
144	404	No status found with that ID	Tweet (or user account) withdrawn

Table 1: Error codes and error messages returned by the Twitter API when we try to access a tweet that has become inaccessible. The last column presents a practical interpretation of each error code.

As shown in Table 1, the error messages accompanying codes 179 and 63 respectively identify the cases where the tweet has become inaccessible because the user made her account private, and where Twitter suspended the account. In this study, we will henceforth ignore the tweets that returned error code 63, since these tweets became inaccessible *not* due to user controlling their exposure, but rather due to Twitter suspending the users.

However, neither the Twitter official documentation[2] nor the error messages help to practically interpret the difference between the error codes 34 and 144. We experimented using the Twitter account of one of the authors of this paper, and observed that, both these error codes practically correspond to the case where the tweet has been withdrawn. However, these two error codes do *not* distinguish between the cases where the user selectively deleted a tweet and where the user deleted her account as a whole. To distinguish between these two scenarios, we further queried the Twitter API to check the *status of the user account* that had posted the tweet. The interpretation of codes is much simpler for user accounts (as compared to those for tweets) – the Twitter API returns HTTP code 200 OK for existing accounts, and error code 404 for deleted accounts.

Thus, by querying the Twitter API with archived tweet IDs (and the userids of users who posted the tweets), and ob-

[2] `https://dev.twitter.com/overview/api/response-codes`

serving the error codes returned, we can determine whether a previously public tweet has been withdrawn.

Limitations of our methodology: We do not know exactly *when* a tweet became inaccessible, i.e., how long after posting was it withdrawn. However, this limitation does not have much effect on the analyses we intend to conduct in the later sections. As we mentioned in the introduction, we also do not capture the user intention behind the withdrawal, i.e., we do not know exactly *why* a user withdrew her tweet or account. That said, we do view historical tweet withdrawal as being implicitly motivated by the desire for controlling longitudinal exposure of prior posts.

3.2 Longitudinal exposure of user data

To measure the longitudinal exposure of user data over the last *six years* from the time of the experiment (October 2015), we used two sets of archived data – (i) a near-complete crawl of Twitter done in September 2009 [9], consisting of 1.7 billion tweets posted by 54.9 million users, and (ii) a 10% random sample provided by Twitter (Gardenhose sample) collected from 2011 till the time of this study. Note that all of these archived tweets were publicly shared when the data was originally collected.[3]

We fixed twenty-two time periods over the last six years, ranging from 1 day ago (from the date of our experiment in October 2015) to 6 years ago (see the x-axis in Figure 1). Then we randomly sampled 5,000 tweets from each of those time periods from our archived data.[4] We used the method described in the previous section on these tweet samples to check how many of the tweets from each time period have been withdrawn today due to exposure control of user data. We repeated the experiment over multiple consecutive days to make sure that the particular day examined was not an outlier (e.g., a holiday, the day a privacy news story broke, etc.). Specifically, for each of the time periods earlier than 2 months ago, we sampled 5,000 random tweets *per day* for a week around that time period and repeated our experiment.

1. How much of the archived data has been withdrawn? Figure 1 shows the variation in the percentage of tweets that have been withdrawn for each time-period. We show box and whiskers for time periods that are greater than or equal to 2 months, representing results from multiple days around those timestamps. We observe that there is little variation among results from the repeated experiments over multiple consecutive days. Unless otherwise stated, we will report the median from the values obtained through the repeated experiments.

We discover that a substantial amount of past data has been withdrawn today. As shown by the solid red curve in Figure 1, the percentage of withdrawn tweets increases from 4.3% of the tweets archived 1 day ago to 28.3% of the tweets archived in 2009. Our observation suggests that users control the exposure for a significant amount of their past data. Hence the natural next question is: how do the different exposure control mechanisms account for this inaccessibility?

2. What is the relative usage of different control mechanisms for longitudinal exposure? Figure 1 further shows the variation of the percentage of tweets withdrawn via the three longitudinal exposure controls – (i) users selectively deleting tweets (green dashed curve), (ii) users deleting their account (blue curve), and (iii) users making their account private (pink curve). Surprisingly, we find that tweets posted from the near to far past have been withdrawn via very different exposure controls. Tweets posted in the near past (e.g., 1 month ago) have mostly been withdrawn via users selectively deleting some of their tweets. However the percentage of tweets withdrawn via selective deletion quickly stabilizes over time. On the other hand, the percentage of tweets withdrawn due to users deleting their accounts or making their accounts private, ramp up as we go further back in the past. In fact, these tweets account for the bulk of the older withdrawn tweets (e.g., 6 years back).

Specifically, out of 8.9% withdrawn tweets from September 2015 (1 month back), 5.9% consists of tweets selectively deleted by users and only 3% is contributed by users who deleted their account or made it private. Whereas, out of 28.3% withdrawn tweets posted in 2009, as much as 16.2% is contributed by users who deleted their account and only 3.2% by users who selectively deleted tweets.

It is important to note that prior studies on deleted tweets, e.g., by Almuhimedi *et al.* [4] *exclusively* focused on data from the near past (e.g., 1 week in the past), most of which are deleted shortly (within a few days) after they are posted. Hence, they ended up analyzing only the selectively deleted tweets, and missed the significant fraction of tweets posted in the far past that have been withdrawn due to users deleting their accounts or making the accounts private.

Summary: We analyzed the longitudinal exposure of socially shared data by measuring the percentage of tweets posted at different time periods in the past, that have been withdrawn as of today. We discovered that a surprisingly large fraction of old tweets has been withdrawn. Moreover, the exposure controls responsible for this withdrawal are very different for the near and far past. This global view motivates us to better understand privacy related behaviors at a user-level, i.e., *how are individual users controlling their longitudinal exposure?* We address this question next.

3.3 Understanding user behaviors

In this section, we assess individual users' behavior for controlling longitudinal exposure in the long-term. From the near-complete snapshot of Twitter data collected in September 2009 [9], we randomly selected 100,000 users who posted at least 100 tweets. For each selected user, we randomly sampled 100 tweets out of all the tweets posted by her (as obtained from the dataset). To simplify further analysis, we selected only the tweets that are in English, i.e., tweets in which at least 50% of the words appear in an English dictionary. Further, we ignored users who were later suspended, and the tweets posted by these users. We were left with 8,950,942 tweets (more than 89% of all tweets), posted by 97,998 users (97.9% of the users).

[3] We observed that Twitter provides a tweet in their random sample nearly instantaneously (within seconds) after a user posts the tweet. Consequently, there is at most a minimal chance that a user deleted a tweet even before it could appear in our random sample.

[4] We only considered original tweets (and not retweets) during sampling since our goal is to understand how much of the tweets originally posted by users are withdrawn today.

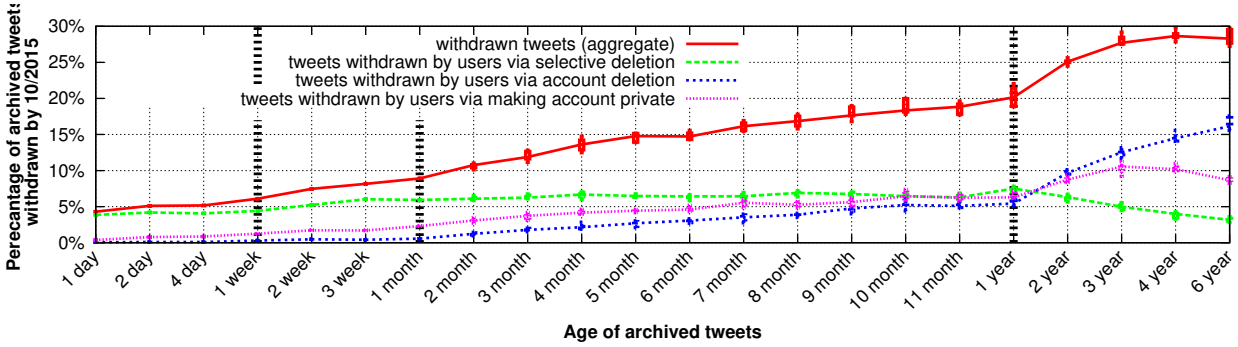


Figure 1: Percentage of tweets in our sample of archived tweets that have been withdrawn as of October 2015. The age of a tweet is the difference between the time when the tweet was posted and the time of querying the Twitter API with the tweet-ids (October 2015). The amount of withdrawn tweets is increasing considerably over time – more than 28% of tweets posted 6 years back have been withdrawn today. The dotted vertical lines in the figure demarcate the points on the x-axis where the scale changes (days vs. months vs. years).

Using the methodology described earlier, we found that 29.1% of all the tweets that we checked have been withdrawn in the last six years, and these tweets were posted by 34.6% of our selected users.

3.3.1 *Longitudinal privacy preferences of users*

We start with categorizing our users into 3 distinct categories based on their usage of longitudinal exposure controls for withdrawing their tweets.

1. Non-withdrawers: users who did not withdraw any of their tweets. 65.4% of the users in our random sample fall in this class.

2. Partial withdrawers: users who only selectively withdrew some of their tweets. 8.3% of users in our sample are in this class. They have contributed 9.7% of the tweets that have been withdrawn.

3. Complete withdrawers: These are the users who have withdrawn all of their old tweets by either deleting their account or making their account private. As many as 26.3% of our selected users (25,751 in total) are in this class. Out of these users, 60.4% users have controlled exposure of their data by deleting their account, while 39.6% have made their account private. Out of all the withdrawn tweets in our sample, these users have contributed the bulk – 90.3% of all withdrawn tweets.

Table 2 shows the relative presence of each category of users in our dataset. We also show the breakdown of these users across different countries where only the top few countries (according to number of users) are shown.[5] The percentage of users with the different privacy preferences remains relatively constant across locations. This observation gives us some confidence that these privacy preferences are not location-specific, rather they are more universal.

One concern with our methodology is that, since we randomly sampled 100 tweets per user, we might potentially undercount the fraction of partial withdrawers. To check how serious this concern is, we repeated our experiments using *all* tweets posted by a set of users. However, due to the presence of some very active users, our sampled users posted more that 60 million tweets in total, and given the rate limitations imposed by the Twitter API, it is very difficult to obtain the present status of all these tweets. Hence, we analyzed a slightly less active set of $\sim$ 97k random Twitter users from 2009, who posted between 10 to 100 tweets each. We repeated the same analysis as above considering *all* of their 2,622,808 English tweets. We found out that 13.6% of the users in this new random sample are partial withdrawers, which is only slightly higher than the fraction of partial withdrawers in our original sample of active Twitter users (8.3%).

We also found that, for a large majority of the users who posted between 10 to 100 tweets, the amount of information available is *not* sufficient for most of the analyses that we performed further (as described in the subsequent sections) due to lesser activity of these users. Hence, in the rest of our study, we will report results for our original set of 97,998 active users who posted 100 or more tweets each.

[5]We obtained the country of our users by leveraging location data of Twitter users gathered by Kulshrestha *et al.* [15]. They used the location and timezone field of the Twitter profile for inferring location of users.

Country	Total users	Non withdrawer	Partial withdrawer	Complete withdrawer
All	**97,998**	**65.4%**	**8.3%**	**26.3%**
US	43,412	65.4%	8.6%	26.0%
UK	4,870	69.7%	8.7%	21.6%
Brazil	4,576	60.8%	8.5%	30.7%
Canada	2,818	67.9%	10.7%	21.4%
Japan	1,740	73.2%	3.6%	23.2%
Australia	1,602	67.6%	7.9%	24.5%
Germany	1,439	67.7%	8.6%	23.7%

Table 2: A breakdown of all users by their privacy preferences as well as by their countries. Note that the breakdown of users by privacy preferences remains relatively consistent across countries.

3.3.2 *Correlating privacy preferences with demographics*

Having identified users with different privacy preferences, we now check who these users are, by correlating the lon-

category	Total #users	# users with inferred gender	% female users
Random population	97,998	65,438	50.3
Non-withdrawers	64,073	41,054	44.5
Partial withdrawers	8,174	5,667	55.7
Complete withdrawers	25,751	18,717	61.5

Table 3: Percentage of female users among different categories of Twitter users whose gender is inferred. The percentage of female users is higher among the partial and complete withdrawers than in a random Twitter population.

gitudinal privacy preferences of the users with their demographics. Twitter maintains only minimal demographic information for users, which includes only a profile bio and location. In spite of the absence of user-reported fine grained demographics information, there has been lot of prior work to infer different demographics characteristics for Twitter users [15, 21, 23]. We leverage this prior work to infer one important demographic for users from the available profile information – gender of these users. We focus on the gender since Tufekci *et al.* [26] noted a correlation between gender and privacy preferences of users in online social media.

We infer the gender from the self-reported first names specified in the user profiles using the methodology developed in [21]. Table 3 shows the percentage of female users among the users with different longitudinal privacy preferences. Interestingly, a majority of the partial and complete withdrawers are female, whereas the exact opposite is true for non-withdrawers. As a baseline, we checked that in a random sample of Twitter users, the percentage of males and females is similar. These results suggest that female users are controlling exposure of their old data more than male users. This finding is also supported by an earlier study on Facebook [26] which reported that women are more likely than men to delete social media content.

Summary: We identify three distinct categories of users based on their individual use of longitudinal exposure control mechanisms. These privacy preferences of individual users do not vary significantly across countries. We also find that a majority of the content withdrawers are female.

After understanding the privacy preferences of different users, and observing the significant use of longitudinal exposure controls among them, we investigate our next question – are there any limitations of the current exposure controls?

4. LIMITATIONS OF EXISTING LONGITUDINAL EXPOSURE CONTROLS

Across online social media sites, the existing longitudinal exposure control mechanisms have an inherent limitation in the form of retained *residual activities* associated with a withdrawn post (e.g., a deleted tweet) or a withdrawn (deleted or private) account.

In these sites users frequently engage in conversations with other users, spurring interactions linked to their posts or to their accounts themselves (e.g., by mentioning a user in a tweet or by tagging a user in a Facebook post). Such interactions also include someone publicly replying to a specific post. When a user selectively deletes her post or withdraws her whole account, those old interactions (from others) associated with her withdrawn post or account become *residual activities* which still points to the withdrawn tweet or account. We show later in this section that, anyone today can collect a number of residual activities (e.g., *residual tweets* on Twitter) around both withdrawn tweets and accounts posted as far as six years back from the time of this study.

We acknowledge that such residual activities might exist even when a user deletes her recent post or withdraws her account created in recent past. However, intuitively, the amount of residual activities grows over time as an account stays longer in an online social media site, and consequently the associated privacy concerns become higher. Thus, we focus our analysis on the residual tweets around withdrawn tweets and accounts posted long back in the past (in 2009).

The presence of residual activities raises an immediate privacy concern – do the residual activities actually breach the longitudinal exposure control mechanisms? In other words, in the context of Twitter, can one recover information about selectively deleted tweets and deleted/protected accounts by simply collecting and analyzing the residual tweets associated with them?

4.1 Recovering information about selectively withdrawn tweets

We first focus on the *selectively* withdrawn tweets, which are deleted by their account holder while *retaining* some other tweets posted from their accounts. Specifically, we ask: what is the amount of the retained residual activities associated with these withdrawn tweets today, and what can we learn from them about withdrawn tweets?

4.1.1 Residual activities around withdrawn tweets

Data collection: We analyzed all the users who selectively withdrew one or more of their tweets from our random sample of 97,998 active users from 2009 (the same dataset as employed in Section 3.3). We then used Twitter search to collect conversations that mention any of those user accounts. Among these conversations, replies to a tweet still contain the tweet id of the tweet. Thus, we also identified the reply posts i.e., residual tweets involving those selectively withdrawn tweets from our dataset.

Limitation of our data: Modified residual tweets like *RT@XTZ:<copiedPartialTweetText>* are easy to (programmatically) assign to withdrawn accounts (@XYZ) but not to particular withdrawn tweets. Therefore we included such residual tweets in the analysis of withdrawn accounts in Section 4.2, but not for the analysis of withdrawn tweets in this section. Thus, the data used in this section is effectively a lower bound on the residual activity around tweets. However, even so, we will show that one can still infer significant information about withdrawn tweets using this data.

How many residual tweets remain around the selectively withdrawn tweets?: In our dataset, a total of 8,174 users selectively withdrew their 253,853 tweets. We were able to collect 12,415 residual tweets posted in response to 9,738 of the withdrawn tweets. Although only 3.8% of all selectively withdrawn tweets have at least one residual tweet, these withdrawn tweets with residual activ-

ities were selectively withdrawn by a significant fraction of the users – 29.2% of 8,174 users who controlled longitudinal exposure by selective withdrawal. We further analyze the number of residual activities per withdrawn tweet. Figure 2 shows that, although a majority (89.2%) of these 9,738 selectively withdrawn tweets (with residual activities around them) have only one residual tweet, 3.8% of those tweets have more than two residual tweets. There is a maximum of 59 residual tweets around a single selectively withdrawn tweet in our data.

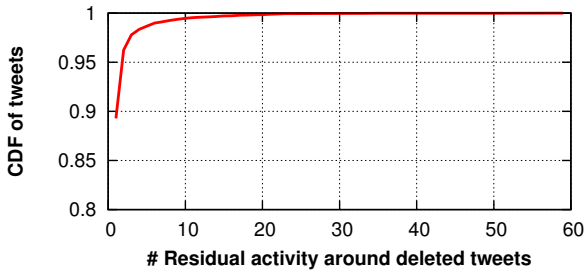


Figure 2: Cumulative distribution function (CDF) for number of residual activities per selectively withdrawn tweet. Each of the withdrawn tweets have non-zero residual activity around it.

4.1.2 *Recovering keywords from withdrawn tweets*

We start by asking – can we recover meaningful words from the original withdrawn tweets just from the residual replies? To answer this question, we first removed all stopwords [6] (no hashtags were removed in the process) from selectively withdrawn tweets and their associated residual activities, then stemmed the remaining words. We call the resulting set of words for a tweet *keywords*. We then checked what fraction of keywords from a withdrawn tweet also appears in the keywords from the set of residual tweets around it.

How many keywords can we recover from the withdrawn tweets?: Figure 3 shows the fraction of keywords shared by the withdrawn tweets and the residual tweets, as the number of residual tweets increases. We report the median values (unless otherwise stated) in this section, and the boxes in Figure 3 indicate the 25th and 75th percentiles. Note that we could recover 16.7% of the keywords when the withdrawn tweets received two or more replies. Moreover, as expected, more residual tweets allow recovery of more information – the fraction of common keywords increases as the number of residual tweets increases.

Keywords revealed from the residual tweets: Table 4 shows some sample withdrawn tweets along with their residual tweets and the keywords gathered from the residual tweets. The keywords that also appear in the withdrawn tweets are highlighted using a bold font. Note that even if all the keywords from residual tweets do not match the ones in the withdrawn tweet, they offer significant contextual information regarding the withdrawn tweet. This becomes more evident as the number of residual tweets increases. This observation motivated us to consider another ambitious idea: *to what extent is it possible for a human observer to guess*

[6] We use a list of English stopwords and a list of Twitter-specific stopwords from [27].

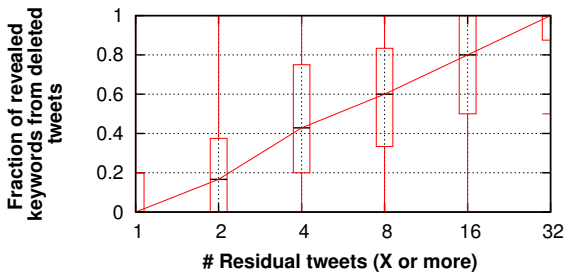


Figure 3: Fraction of keywords that could be extracted for each of the withdrawn tweets (with at least one residual tweet) with varying number of residual tweets. The boxes indicate the 25th and 75th percentiles in the fraction, and the whiskers indicate the minimum and maximum values. The recovered keywords from withdrawn tweets increase with the number of residual tweets.

the meaning of a withdrawn tweet from the residual tweets? Specifically, we asked human observers to guess a withdrawn tweet from its residual tweets, and then informally checked whether the meaning of the guessed tweets is qualitatively similar to the meaning of the original withdrawn tweet.

4.1.3 *Recovering meaning of withdrawn tweets*

Since guessing the meaning of a tweet automatically is a hard problem, we instead took help of human annotators from Amazon Mechanical Turk (AMT) for a preliminary demonstration. We used three AMT master workers from the USA for this survey. Each worker was first shown 5 example tweets and their replies. We first binned all of our selectively withdrawn tweets into five bins by the number of their residual tweets (i.e., tweets with $1, 2, \ldots, 5$ or more residual tweets) and selected ten withdrawn tweets from each bin. For our randomly sampled 50 withdrawn tweets, all the AMT workers were then shown the residual tweets of each withdrawn tweet and were simply asked to "Guess the original tweet". Finally we read through the guessed tweets and informally checked the (qualitative) resemblance between the meaning of the original withdrawn tweet and that of the guessed tweets.

Table 5 shows a part of the result from our AMT experiment.[7] As expected, when the number of residual tweets is small, the AMT workers were sometimes unsure about the meaning of the withdrawn tweet. Nevertheless, as the number of residual tweets increased, all the human observers guessed the meaning of the withdrawn tweets reasonably well (as reflected in their guessed tweets). This observation indicates that residual tweets often give out sufficient information for a human observer to guess the meaning of selectively withdrawn tweets.

Summary: We demonstrate that it is possible to recover both keywords and meaning from the withdrawn tweets by collecting and analyzing the available residual tweets associated with them. This is definitely a bad news for the users

[7] For an interested reader to check the resemblance in meaning between the guessed and original tweets, we put our complete AMT evaluation result at `http://twitter-app.mpi-sws.org/soups2016/amt_guess.html`.

Original withdrawn tweet	#Residual tweets	Example keywords from residual tweets	Example residual tweets
Saw The Cove last night. Made me think about how much ALL animals need our respect – dolphins, cats, pigs, dogs, cows, chickens...	1	**cove**, **respect**, **animals**, extend, yeah, sea, recommending, veganfail, eat	"@[username] Yeah, but too bad "The Cove" doesn't extend that respect by recommending to not eat any animal from the sea"
[url] - Is it bad for you to eat unbaked cookie ? Hope not	3	**cookie**, **eat**, dough, batter, yummy, eveyone	"@[username] Cookie dough is awesome! Eat it up.", "@[username] i don't think so. isn't it like eating cookie dough? i do it with cake batter all the time. it's yummy"
What happened with Palin?	7	**palin**, resigining, alaska, safe, dearly, white, house, fantastic, definitely	"@[username] she's resigning. awww...", "@[username] she's going to act now....Nat'l Lampoon: Palin goes to Hollywood."

Table 4: Examples of withdrawn tweets, example keywords from the residual tweets, and actual examples of residual tweets. The keywords common in withdrawn tweets is shown in the bold font. As the number of residual tweets increases, their keywords give out more context about the withdrawn tweet.

Original withdrawn tweet	#Residual tweets	Guessed tweet from AMT workers		
		Guess 1	Guess 2	Guess 3
Saw The Cove last night. Made me think about how much ALL animals need our respect – dolphins, cats, pigs, dogs, cows, chickens...	1	The Cove has vowed to not eat any animals, good start!	Loved The Cove!	I think it's cool that the cove doesn't eat animal meat.
[url] - Is it bad for you to eat unbaked cookies? Hope not	3	Cook cookies? no thanks, I'll just eat them raw.	Are you sure I can eat this stuff? It's got raw food in it	I made cookie dough, but I can't seem to actually bake the cookies because I can't stop eating the dough!
What happened with Palin?	7	Sarah Palin finally stepping down, good day!	Read Sarah Palin's governorship resignation speech here: <link>	I wonder why Palin is resigning??

Table 5: Examples of selectively withdrawn tweets and the corresponding tweets guessed by AMT workers who were shown only the residual tweets for a withdrawn tweets. As the number of residual tweets increases, the AMT workers guessed the meaning of the original withdrawn tweet more closely.

who wish to control exposure of their old post through selective withdrawal.

4.2 Recovering information about withdrawn accounts

Twitter users widely employ two mechanisms towards controlling longitudinal exposure of their accounts – some prefer to delete their accounts, while others prefer to make accounts private making their content inaccessible to a public observer. We collectively call these deleted or protected accounts *withdrawn accounts*. Here, we study two questions: what amount of residual activity around a withdrawn account is available, and what information does this residual activity reveal about the withdrawn accounts?

4.2.1 Residual activities around withdrawn accounts

We collected residual tweets around withdrawn accounts using a similar methodology as described in Section 4.1.1. We considered the withdrawn accounts from our random sample of 97,998 users from 2009 (same dataset from section 3.3), and then used Twitter search to collect posts that mentions any of those user accounts. We limited our search to the period when the withdrawn accounts were active in our dataset, i.e., from the account creation date to the date of the last tweet appearing in our data.

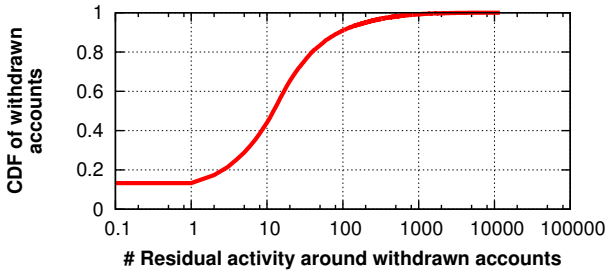


Figure 4: CDF of number of residual activities per withdrawn account. More than 55% of withdrawn accounts have more than 10 residual tweets.

How many residual activities remain around withdrawn accounts?: We collected a total of 1,403,716 residual tweets that mentioned 23,526 withdrawn accounts. In other words, a substantial fraction (91.4%) of the 25,751 withdrawn accounts have some residual tweets around them. We analyzed the number of residual activities around each account. Figure 4 shows that a significant amount of residual

activities remain even at an individual account level – 55.9% of all withdrawn accounts have 10 or more residual tweets. Next, we ask what information can we recover about these withdrawn accounts, using both the residual tweets and the existing accounts that posted those residual tweets?

4.2.2 Recovering social connections

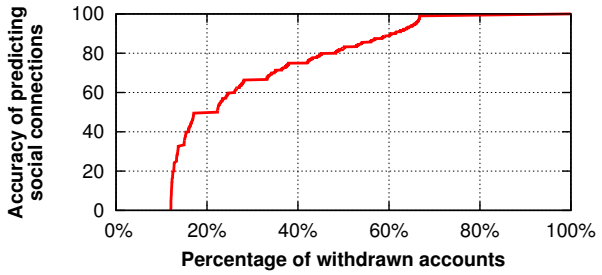


Figure 5: The accuracy of our social connection inference with the percentage of withdrawn accounts for which we get this accuracy. For more than 30% of withdrawn accounts, all of their residual tweets came from their social connections.

We expect that two users converse mostly when they are socially connected. Thus, as a first test, we check if the users who mentioned a withdrawn account were connected to the withdrawn account by the follower-following relation. Cha *et al.* [9] had collected all the followers and followings of all Twitter users in 2009 and our withdrawn accounts are part of their dataset. Leveraging their collected data, we took all the social connections (both followers and followings) for each withdrawn account as our ground truth. Then we did a simple prediction: we predicted that each of the accounts mentioning a withdrawn account are either followers or followings of the withdrawn account. The accuracy of our inference for each user was: for what percentage of cases was our prediction correct?

Figure 5 shows the accuracy of our inference and for what percent of users we have a specific accuracy. Significantly, for 33.3% of the withdrawn accounts, the accuracy is 100%, i.e., all residual activities around these withdrawn accounts were posted by their social connections. For 48.3% of the withdrawn accounts, accuracy is more than 80%. Therefore, simply by checking who posted the residual tweets associated with a withdrawn account, we can recover some social connections for a significant number of withdrawn accounts.

A large number of existing studies pointed out that connected users in online platforms show homophily, i.e., have similar characteristics [3, 25]. So we next check if we can recover some of the demographic attributes, like location, for the withdrawn accounts by leveraging the demographics of the accounts who contributed to the residual posts.

4.2.3 Recovering demographics

We here focus on whether we can infer the location of an withdrawn account from the location of the accounts who contribute to the residual activity around the withdrawn account. As stated earlier, we obtained the ground truth country-level location for user-accounts from the study [15]. We then picked the most frequent location among the accounts which posted the residual tweets, as our predicted location for the corresponding withdrawn account. Our accuracy was decided by the number of withdrawn accounts for which our prediction was correct. As a baseline for comparison, we take the accuracy of a trivial predictor that selects USA as location every time (the most popular country in Twitter population).

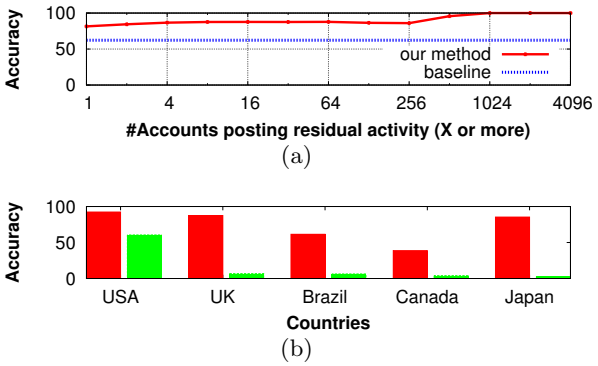


Figure 6: 6(a) Accuracy of our location inference leveraging residual activities. We can infer location with high accuracy and the inference is consistently better than baseline. 6(b) the accuracy for withdrawn accounts from different countries. First bar for each country is accuracy of our method and second bar is percentage chance that a random user will belong to that country.

Demographics prediction accuracy: Figure 6(a) shows the accuracy of our prediction with increasing number of user accounts associated with residual tweets. Significantly, when a withdrawn account has three or more accounts posting residual tweets around it, just by leveraging the residual activities we can infer the withdrawn account's location in 85.8% cases. This is consistently better than the baseline.

We also analyzed accuracy of our location inference for top five countries for the withdrawn accounts with some residual activities. The baseline accuracy for each country in this analysis was the accuracy of a predictor that outputs location based on the chance that a random Twitter user will belong to that country (computed using the full random sample of ∼98K users from Section 3.3). Figure 6(b) shows the comparison of accuracy for top five countries. We note that even for countries like Japan, where the chance of a random user coming from the country is as low as 2.25%, our inference is accurate for more than 87% withdrawn accounts.

4.2.4 Recovering topics of interest

To recover potential topics the withdrawn accounts could have been interested in, we leveraged a special type of keyword – *hashtags*. Hashtags are words in tweets that starts with a '#' symbol and are included to provide the tweet a specific context. Practically hashtags are used to group together multiple tweets on the same topic. For example, there were multiple tweets posted with "#iranelection" in 2009 to identify the topic of the tweet related to Iran election 2009.

Using data from [9], we determined that 3,855 accounts in our set of withdrawn accounts posted at least one tweet with a hashtag. Out of those, for 58.7% accounts (2,263 in total),

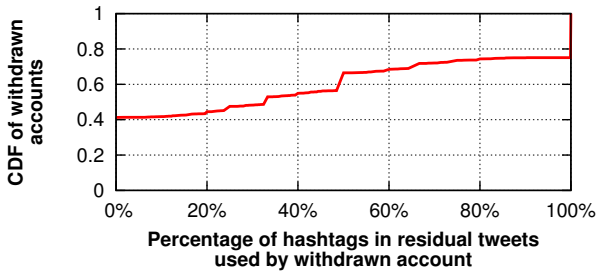


Figure 7: The percentage of hashtags revealed by residual tweets that were originally also used by a withdrawn account. 25% of the withdrawn accounts, who ever used any hashtag in their tweets, used all of the hashtags revealed from their residual activities.

the residual tweets revealed at least one of their hashtags, and in total 3,625 unique hashtags were revealed for these withdrawn accounts. This correlation encouraged us to further check what percentage of the hashtags revealed by the residual tweets were also used by the withdrawn accounts. Figure 7 shows our results: interestingly, in 25% of the cases, *all* the hashtags revealed by the residual tweets were also used by the withdrawn account.

User serial	Topics	Hashtags used by withdrawn accounts, that are revealed by residual tweets
1	Politics, Sports, Technology	#iranelection, #prisoners, #strike, #frenchopen, #tech
2	Politics	#conservativebabesarehot, #teaparty, #tcot, #obamacare
3	Sports, LGBTQ issues	#daviscup, #samesexsunday, #india, #lgbt, #followfriday
4	Sexuality, Entertainment	#furgasm, #nsfw, #gay, #shazam, #music
5	LGBTQ issues	#housing, #dcmetro, #protest, #gaymarriage
6	Politics	#immigrationreform, #iranelection, #peace #lgbt
7	Religion	#jesus, #truth, #idol
8	Sports	#grandrapids, #nascar
9	Sexuality	#hugeboner, #carchat
10	Sports, Entertainment	#collegefootball, #seinfeld

Table 6: Hashtags revealed by residual tweets for 10 withdrawn accounts. These users themselves used each of these hashtags. Also shown are some manually annotated topical categories these hashtags fall into. These hashtags give us an idea of what might be the topics of interest of the withdrawn accounts.

We further analyzed the hashtags revealed from residual tweets for some individual withdrawn accounts, and manually annotated the hashtag topics. Table 6 presents some example hashtags from the residual tweets of 10 users, who had used all of these hashtags in their now withdrawn tweets. As shown by our manual topical annotation of these hashtags, these hashtags shed light on the user's interests partially if not fully. Interestingly, some of these hashtags like "#iranelection", "#nsfw" might even be considered sensitive, while other hashtags such as "#daviscup", "#tech" or "#nascar" give away specific interests of the withdrawn accounts. This observation provides evidence that the residual tweets still reveal information about what a withdrawn account was interested in, even when the account become inaccessible.

Twitter app to raise awareness about residual activities: To increase user awareness about their residual activities, we designed a Twitter app, using which any Twitter user can check what information about her account and individual tweets can be inferred by simply analyzing her residual activities on Twitter. We invite readers to use the app by visiting `http://twitter-app.mpi-sws.org/footprint/`.

Summary: We found significant evidence that the residual tweets and their associated user-accounts can be leveraged to at least partially recover the social connections, demographics (location) and even topical interests of the withdrawn accounts. Hence, the goal of the withdrawn tweet / account owners to control exposure of their (past) data cannot be achieved by the existing exposure control mechanisms. In the next section, we discuss the relative merits and demerits of a few exposure control mechanisms, and how such mechanisms can be improved.

5. TOWARDS BETTER LONGITUDINAL EXPOSURE CONTROL MECHANISMS

Our analyses in the earlier sections show that a large number of users withdraw their past social content, but often a significant amount of residual information is left behind, which might lead to significant information leakage about withdrawn social content (and consequent privacy violation). This calls for an improvement of longitudinal exposure control mechanisms, which will directly increase the usability of such systems from a privacy perspective.

However, it must be understood that improving longitudinal exposure control mechanisms is a complex problem, as this has to take into consideration multiple (and sometimes contradictory) factors, such as the desire to retain some old content while allowing other content to be completely removed without a trace [6]. In fact, analyzing the effectiveness of such a mechanism might require a far richer understanding of many dimensions like incorrectly (not) limiting exposure of (non-)desirable content, potential privacy impact of such false flags, ownership of residual activities, ease of use and even user sentiment. Hence, it is very unlikely that there is a silver bullet to solve all the problems with longitudinal exposure control. The longitudinal exposure control mechanisms that are being deployed in different online social sites today, aim towards improving different dimensions of the problem, some of which we discuss below. We also propose a novel mechanism for longitudinal exposure control, which addresses some of the limitations of the existing mechanisms.

5.1 Existing Mechanisms

1. Putting users in charge of controlling their longitudinal exposure: This mechanism is used in most of the popular online social media sites, including Twitter and Facebook, where the users are expected to control their own longitudinal exposure by withdrawing individual posts / accounts. On the positive side, this mechanism perfectly captures the user intent of retainment or withdrawal of specific content. However, as the previous section demonstrated, even when users withdraw their posts or accounts, the residual activity surrounding the withdrawn posts (authored by

other users) could leak significant information about the withdrawn content.

It can be argued that withdrawing the residual activities along with the withdrawn posts and accounts is a natural solution to this issue or residual information. However, any such tampering of the content authored by other users (other than the one who specifically wishes to delete her content) raises several difficult questions associated with ownership and control of the content.[8]

2. Age based withdrawal: *Ephemeral social media sites* such as Snapchat [2] and Cyber Dust [1] offer a potential way out of the residual activity problem. On such sites, every message is associated with an expiry time after which the post is automatically withdrawn and becomes inaccessible to the users. Ayalon *et al.* [5] also suggested that the system operators of non-ephemeral social media sites can offer their users similar timed expiry option such that the posts will become inaccessible to the public after the expiry time.

Though this mechanism solves the problem of residual activities (since even the residual activities will be inaccessible over time), it has two limitations. First, the default expiry time used in such mechanisms is generally too small (e.g., few seconds or few minutes), which prevents any meaningful discussion around any post. Since the most interesting posts also get deleted after the expiry time, such mechanisms might not be preferred in sites like Twitter which promote social discussions. Second, as noted in [6], users are generally poor at anticipating when a post should be deleted, which reduces the practical use of this mechanism even if users are given the option of setting the expiry time.

5.2 Our proposal: Inactivity-based withdrawal

Our proposal is based on a simple intuition – when a post becomes inactive, i.e., it does not generate any more interaction or receive any more exposure, the post can be safely withdrawn (deleted/archived/hidden) from the public domain. Note that 'interaction' is a general term that can involve several tasks based on the social media site; e.g., it can mean sharing the post (e.g., retweeting in Twitter), replying to the post or even viewing the post by the original posting account or other users. Large social media operators today collect all of these interactions.[9] Hence, they can easily check if a post is inactive for more than T days (for any given definition of inactivity), and then the post can be withdrawn from the public domain. Also note that a user can be given various options for withdrawing her posts which become inactive; for instance, instead of fully deleting the posts, she may instead decide to limit access to the post to only select friends or may even anonymize the posts by removing any identifiable information. Here we generally consider withdrawal of posts from the public domain, and leave the details of the exact access control decisions to the social media operators.

Compared to age based withdrawal, this mechanism has the following advantages. First, the users need not be burdened with deciding expiry times of their posts. Second, this mechanism allows meaningful discussions around interesting posts, since the posts are withdrawn only after the discussion around them has died down.

However, we do acknowledge that just like earlier mentioned mechanisms, our proposal is not a silver bullet. For instance, this mechanism does *not* capture a user's intent to retain some old content even after it becomes inactive (e.g., because it had acquired large popularity, or because of some user-sentiment around a particular post). Another limitation of this mechanism is that, if a post is continuing to get interactions because it is controversial in nature, this mechanism would lead to the post remaining in the public domain. To address such issues, this mechanism should be coupled with other exposure control mechanisms such as a user being able to specifically withdraw some posts, or indicating her desire to retain a post even after it becomes inactive.

Even if a user wishes to adopt our proposed mechanism, a technical question needs to be addressed – how to select a value for T, the number of days after which a post will be withdrawn? With a very small value of T (say, 1 day), we may end up losing some valuable interactions; on the other hand, if T is too high (e.g., six years) users run a significant risk of someone digging up information about their past lives. Next, we demonstrate how the system operators can leverage the past interaction history to select an appropriate value of T.

Deciding an inactivity threshold: We ask a simple question in this direction: if we set a threshold of T days of inactivity before withdrawing a post, how much of the interaction generated by a post is likely to be lost? To that end we perform the following experiment. We randomly sample 700,000 tweets posted in the first week of November 2011, i.e., more than four years back. Note that all of these tweets are accessible today. In our experiment we take "retweets" as a proxy for generated interactions by a tweet. For a given tweet, we can obtain this interaction information directly from the Twitter API (unlike interactions like residual activities). In our dataset, 30,014 tweets received at least one retweet and they received 74,705 retweets in total. We collect information about when each tweet received their retweets using the Twitter API, and simulate setting our inactivity threshold at T days, i.e. each of these tweets will become inaccessible after T days of not getting any retweets. We analyze the number of future retweets we would lose for different values of T.

Figure 8 shows that if we set our threshold to be too low, say 1 day, we will lose a significant 5.5% of all the retweets. However, if we set our threshold at only 180 days (i.e., decide that after six months of inactivity a tweet might be withdrawn from the public eye) then only 0.4% of the future retweets will be lost. Note that the parameter T need not to be global, and every user may choose her own value. In fact, the system operator can show a range of values of the threshold and point out the associated percent of stopped activities based on a user's past history, and allow the user can make an informed decision.

A comparison between the inactivity-based withdrawal and the age-based withdrawal: To demonstrate advantages of inactivity-based withdrawal over the

[8] For example, Twitter today automatically deletes re-tweets of a deleted tweet, but *not* replies or mentions generated by other users.

[9] `https://support.twitter.com/articles/20171990#` `https://www.facebook.com/help/437430672945092`

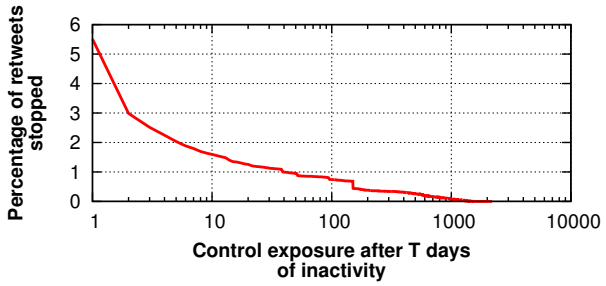


Figure 8: Percentage of lost retweets if tweets were withdrawn after T days of inactivity, for different values of T. When T is set to 180 days only 0.4% of the future retweets will be lost.

Threshold in days	Inactivity based withdrawal		Age based withdrawal	
	#Retweets stopped	#Tweets these retweets came from	#Retweets stopped	#Tweets these retweets came from
1	4,117	1,584	7,798	1,681
7	1,342	556	2,678	587
30	842	317	947	339
90	609	235	744	243
180	300	181	579	193

Table 7: Comparison of age and inactivity-based threshold when both have the same threshold. Retweets of more active tweets are stopped by age-based threshold.

age-based withdrawal, we also simulated age-based withdrawal policy with different thresholds over the same dataset of 700,000 random tweets and their retweets. Our age-based withdrawal policy is simple: after T days the tweet will be withdrawn and all future retweeting will be stopped. We closely investigate how many retweets will be affected by both these policies if we set same threshold. Table 7 shows the absolute number of retweets stopped and the number of tweets these retweets come from. It demonstrates that for the same threshold T, inactivity-based withdrawal stops comparatively fewer retweets than age-based withdrawal.

From our experiments, we make a more interesting observation: age-based withdrawal also affects tweets which generates lot of interaction (i.e., retweets) over a longer period of time, e.g, a tweet from the president of the United States. Let us take an example: Table 7 shows that when the threshold is set to 180 days, inactivity-based withdrawal stops 300 retweets from our dataset as it makes 181 tweets inaccessible. For the same threshold, age-based withdrawal makes 12 more tweets inaccessible (total 193), but stops 279 retweets from those additional 12 tweets, (i.e., on average 23 retweets per tweet). Notice that, by generating a lot of activity, popular tweets increase the usefulness of social content sharing systems. Thus, since age-based withdrawal might affect popular tweets, even with a high threshold it might not be suitable in the real-world adaptation. To demonstrate the effect of this issue, we measure actual time when a tweet will be withdrawn when we set an inactivity-based threshold of T days for different values of T. In Figure 9, we plot the withdrawal age of the (inactivity-based) withdrawn tweets, and rank them in a sorted order based on their age. From the slope of these plots for different values of T, it is clear that the actual age of most tweets is significantly higher than their inactive age (or period).

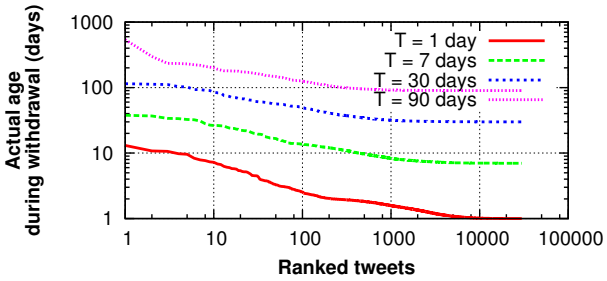


Figure 9: Actual time when a tweet will be deleted when we set an inactivity-based threshold of T days.

Summary: We consider our inactivity-based withdrawal method to be an improvement over the age-based withdrawal, as it removes the need for a user to guess when her content should be withdrawn. Instead, the social site operator can present suggestions to users when a post becomes inactive, and facilitate the withdrawal. Our proposed mechanism does not solve all the problems with longitudinal exposure control, but we do believe it is a step toward more usable longitudinal exposure control mechanisms.

6. CONCLUSION

In this paper, we explored a dimension of user privacy that becomes more challenging to manage with passing time, namely, longitudinal privacy. Specifically, using extensive data from the Twitter social media site, we studied whether online users employ longitudinal exposure control mechanisms in real world to limit exposure of their old data. We find that a surprisingly large fraction (28%) of tweets posted in the far past are withdrawn by users today. After exploring the usage of existing privacy mechanisms by individual users, we find a significant problem with mechanisms to control data exposure today – social media sites retain residual activities around withdrawn content, which can be used to recover various important information ranging from social connections to user interests and even parts of the withdrawn content. We also proposed an exposure control mechanism called *inactivity based withdrawal* – an embodiment of the simple idea that old content can be safely withdrawn when it does not generate any more activity – and showing its benefits for controlling longitudinal exposure over existing age-based exposure controls. However, our study also calls for further research in this field, as much remains to be done in this space of understanding and improving longitudinal exposure controls of socially shared data.

7. REFERENCES

[1] Cyber Dust. `https://www.cyberdust.com/`, 2016.

[2] Snap chat. `https://www.snapchat.com/`, 2016.

[3] L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, M. Benjamin, and F. Menczer. Friendship prediction and homophily in social media. *ACM Transactions on the Web*, 6(2):1559–1131, 2012.

[4] H. Almuhimedi, S. Wilson, B. Liu, N. Sadeh, and A. Acquisti. Tweets are forever: A large-scale quantitative analysis of deleted tweets. In *Proceedings of the 16th Conference on Computer Supported Cooperative Work (CSCW'13)*, February 2013.

[5] O. Ayalon and E. Toch. Retrospective privacy: Managing longitudinal privacy in online social networks. In *Proceedings of the 9th Symposium on Usable Privacy and Security (SOUPS '13)*, July 2013.

[6] L. Bauer, L. F. Cranor, S. Komanduri, M. L. Mazurek, M. K. Reiter, M. Sleeper, and B. Ur. The post anachronism: The temporal dimension of facebook privacy. In *Proceedings of the 12th ACM Workshop on Privacy in the Electronic Society (WPES'13)*, November 2013.

[7] M. S. Bernstein, E. Bakshy, M. Burke, and B. Karrer. Quantifying the invisible audience in social networks. In *Proceedings of the 31st SIGCHI Conference on Human Factors in Computing Systems (CHI'13)*, April 2013.

[8] A. Besmer and H. R. Lipford. Moving beyond untagging: Photo privacy in a tagged world. In *Proceedings of the 28th SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*, April 2010.

[9] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in twitter: the million follower fallacy. In *Proceedings of the 4th AAAI Conference on Weblogs and Social Media (ICWSM'10)*, May 2010.

[10] R. Dey, Z. Jelveh, and K. W. Ross. Facebook users have become much more private: A large-scale study. In *Proceedings of the 10th Annual IEEE International Conference on Pervasive Computing and Communications (perCom'12)*, March 2012.

[11] C. M. Hoadley, H. Xu, J. J. Lee, and M. B. Rosson. Privacy as information access and illusory control: The case of the facebook news feed privacy outcry. *Electronic Commerce Research and Applications*, 9(1):50–60, 2010.

[12] P. Jain and P. Kumaraguru. On the dynamics of username changing behavior on twitter. In *Proceedings of the 3rd IKDD Conference on Data Science (CODS'16)*, March 2016.

[13] M. Johnson, S. Egelman, and S. M. Bellovin. Facebook and privacy: It's complicated. In *Proceedings of the 8th Symposium on Usable Privacy and Security (SOUPS'12)*, July 2012.

[14] H. W. J. Jr. Google and the search for the future. `http://www.wsj.com/articles/SB10001424052748704901104575423294099527212`, August 2010.

[15] J. Kulshrestha, F. Kooti, A. Nikravesh, and K. P. Gummadi. Geographic dissection of the twitter network. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM'12)*, June 2012.

[16] Y. Liu, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Analyzing facebook privacy settings: User expectations vs. reality. In *Proceedings of the 11th ACM/USENIX Internet Measurement Conference (IMC'11)*, November 2011.

[17] Y. Liu, C. Kliman-Silver, and A. Mislove. The tweets they are a-changin': Evolution of twitter users and behavior. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM'14)*, June 2014.

[18] M. Madden, A. Lenhart, S. Cortesi, U. Gasser, M. Duggan, A. Smith, and M. Beaton. Teens, social media, and privacy. `http://www.pewinternet.org/2013/05/21/teens-social-media-and-privacy/`.

[19] M. Madejski, M. Johnson, and S. M. Bellovin. The Failure of Online Social Network Privacy Settings. Technical Report CUCS-010-11, Department of Computer Science, Columbia University, 2011.

[20] A. Mazzia, K. LeFevre, and E. Adar. The pviz comprehension tool for social network privacy settings. In *Proceedings of the 8th Symposium on Usable Privacy and Security (SOUPS'12)*, July 2012.

[21] L. Mullen. Predicting gender using historical data. `https://cran.r-project.org/web/packages/gender/vignettes/predicting-gender.html`, 2015.

[22] S. Petrovič, M. Osborne, and V. Lavrenko. I wish I didn't say that! analyzing and predicting deleted messages in twitter. *CoRR*, abs/1305.3107, 2013.

[23] L. Sloan, J. Morgan, P. Burnap, and M. Williams. Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *PLoS ONE*, 10(3):e0115545, 2015.

[24] F. Stutzman, R. Gross, and A. Acquisti. Silent listeners: The evolution of privacy and disclosure on facebook. *Journal of Privacy and Confidentiality*, 4(2), 2012.

[25] M. Thelwall. Homophily in myspace. *Journal of the American Society for Information Science and Technology*, 60(2):219–231, 2009.

[26] Z. Tufekci. Facebook, youth and privacy in networked publics. In *Proceedings of the 6th International Conference on Weblogs and Social Media (ICWSM'12)*, June 2012.

[27] M. B. Zafar, P. Bhattacharya, N. Ganguly, K. P. Gummadi, and S. Ghosh. Sampling content from online social networks: Comparing random vs. expert sampling of the twitter stream. *ACM Transactions on the Web*, 9(3):12:1–12:33, 2015.

[28] L. Zhou, W. Wang, and K. Chen. Tweet properly: Analyzing deleted tweets to understand and identify regrettable ones. In *Proceedings of the 25th International Conference on World Wide Web (WWW'16)*, April 2016.

Sharing Health Information on Facebook: Practices, Preferences, and Risk Perceptions of North American Users

Sadegh Torabi
University of British Columbia
Department of Electrical and Computer Engineering
sadeq@ece.ubc.ca

Konstantin Beznozov
University of British Columbia
Department of Electrical and Computer Engineering
beznosov@ece.ubc.ca

ABSTRACT

Motivated by the benefits, people have used a variety of web-based services to share health information (HI) online. Among these services, Facebook, which enjoys the largest population of active subscribers, has become a common place for sharing various types of HI. At the same time, Facebook was shown to be vulnerable to various attacks, resulting in unintended information disclosure, privacy invasion, and information misuse. As such, Facebook users face the dilemma of benefiting from HI sharing and risking their privacy.

In this work, we investigate HI sharing practices, preferences, and risk perceptions among Facebook users. We interviewed 21 participants with chronic health conditions to identify the key factors that influence users' motivation to share HI on Facebook. We then conducted an online survey with 492 Facebook users in order to validate, refine, and extend our findings.

While some factors related to sharing HI were found in literature, we provide a deeper understanding of the main factors that influenced users' motivation to share HI on Facebook. The results suggest that the gained benefits from prior HI sharing experiences, and users' overall attitudes toward privacy, correlate with their motivation to disclose HI. Furthermore, we identify other factors, specifically users' perceived health and the audience of the shared HI, that appear to be linked with users' motivation to share HI. Finally, we suggest design improvements—such as anonymous identity as well as search and recommendation features—for facilitating HI sharing on Facebook and similar sites.

1. INTRODUCTION

Individuals with health condition(s) can benefit from sharing their health information (HI)[1] in different ways: seeking or providing social support, learning from the shared experiences, and self-management education [34, 39, 47]. Furthermore, sharing HI was shown to be beneficial not only to the individuals themselves, but also to their social peers. Motivated by the two-way benefits, people have used different online services to exchange their HI and experiences (e.g., blogs, emails). Among these services, social networking sites (SNSs), which have attracted the largest number of active online users, have emerged as a common place for sharing different types of personal information, including HI [15, 38].

Recent studies suggest that various user groups with health conditions (e.g., breast cancer patients) may benefit from sharing HI on SNSs [34, 39]. On the other hand, revealing HI to other people has been always associated with privacy concerns. Not only have researchers identified an implicit consensus among people that their HI should be considered as "personal" and "private" [33, 39], but also Facebook and SNSs were shown to be vulnerable to various attacks that could result in unintended information disclosure, privacy invasion, and medical data misuse [18, 46].

One can argue that the users' attitude toward disclosing HI might be influenced by their perception of privacy and the expectation of benefits (privacy calculus) [31]. Although a number of studies brought to the attention of the research community the benefits and the privacy concerns related to HI sharing on Facebook, we need to do more work to understand the interplay among various factors (including privacy concerns) and the users' motivation to engage in HI sharing [30, 34, 35]. In order to increase the benefit of sharing HI by SNS users, it is important to investigate socio-technical features that motivate and enable users to share their HI effectively and safely. This, however, demands building a better understanding of users' HI sharing practices and risk perceptions.

To this end, we studied why, how, and with whom users share their HI on Facebook. Following a grounded theory approach [7], we interviewed 21 individuals who had chronic health conditions. We focused on exploring participants' practices, preferences, and risk perceptions when sharing HI on Facebook. The interviews enabled us to develop a better understanding of the key factors linked to users' motivation to share HI on Facebook. We then conducted an online survey with 492 active Facebook users, in order to confirm and extend upon our findings.

All studies were reviewed and approved by our university's ethics committee. We minimized risks to participants by excluding any personally identifiable information from the collected data, generated results, and published reports. Participation in all studies were completely voluntary, and participants were able to withdraw from the study at any time.

[1] Any type of information related to the health of an individual including personal health information (PHI), electronic health records (EHRs), and personal health records (PHRs)

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22–24, 2016, Denver, USA.

The results of our investigation suggest that participants who previously shared their HI on Facebook, especially those who gained some benefits, were more willing to share their HI on SNS in the future. Yet, despite the perceived benefits, participants who had strong privacy concerns were always unlikely to share their HI, as compared to participants with medium or low privacy concerns, who showed more flexibility in the presence of different motivating factors (e.g., perceived benefits). Furthermore, we found that participants' perceived health status correlates with their motivation to share HI with different Facebook users, even strangers. It also matters for all types of users who the intended recipients are. Based on the findings, we suggest a number of features (e.g., anonymous identity, specialized search and recommendations, trusted SNS provider) that could motivate users toward engaging in effective HI sharing on Facebook.

In summary, this work makes the following contributions:

- We provide a better understanding of Facebook users' HI sharing practices, preferences, and risk perceptions.
- We identify factors linked to users' perceived privacy and motivation to share HI on Facebook.
- We suggest design features that could facilitate effective HI sharing among Facebook users.

In what follows, we present background and related work (Section 2). In Sections 3 and 4, we present details of the exploratory and confirmatory studies. In Section 5, we discuss the main findings along with study limitations and implications for design. We conclude by presenting conclusions in Section 6.

2. BACKGROUND AND RELATED WORK

Several studies have indicated that HI sharing is becoming a common behavior among a considerable number of SNSs users [15, 32, 40]. The results of the Pew Internet survey suggest that a considerable number of internet users in the U.S. went online to follow their friends' personal health experience, with a noticeable increase when compared to previous years [14]. Meanwhile, 16% of the surveyed participants reported going online to find others who had similar health concerns [15]. Moreover, people with health concerns have been shown to visit their SNSs (e.g., Facebook) to seek support from other online peers [38].

The benefits of using SNSs for HI sharing have been investigated by a number of studies [33, 34, 47]. Lederman et al. [29] discussed the benefits of addressing socio-technical needs by utilizing SNSs and developing engaging therapeutic solutions for mentally ill patients. Following a user-centered design approach, Skeels [39] captured breast cancer patients' HI sharing requirements and designed an online interactive technology to facilitate HI sharing and management. Kamal [22] also used a similar approach to design a SNS prototype for promoting healthy behavior changes.

Despite the reported benefits for people with chronic health conditions, only a small number of studies explored the effects of using SNSs on health management. For instance, Newman et al. [37] interviewed 14 participants who joined health-focused online communities in order to investigate the way people think about sharing HI as they pursue social goals related to their personal health. The methodological limitations (data collection/analysis) and the focus on the niche demographics in their study render the findings non-generalizable to the user (or even patient) population at large. In addition, Newman et al. explore the mixture of online and offline user experiences, unlike our research of users' HI sharing behaviors on Facebook. In another relevant investigation, Merolli et al. [34] reviewed the literature and found that among all the examined studies (N=19), only five focused on SNSs (referred to as web 2.0 sites). Moorhead et al. [35] surveyed primary research and identified the lack of information about the uses, benefits, and limitations of social media for health communication among the general public, patients, and health professionals. Similar conclusions were drawn by Lefebvre and Bornkessel [30], where they suggest further investigations, in order to better understand how SNSs can be effectively and efficiently used to improve health across the population.

People's motivation to engage in protective health behaviors was shown to be influenced by the severity and the likelihood of their health conditions [43]. In the context of HI sharing on SNSs, a number of studies have shown that people who suffered from chronic health conditions were likely to visit SNSs to seek or share their HI with social peers [14, 38]. For instance, Lederman et al. [29] highlighted the motivation of mentally ill patients toward engaging in online therapeutic procedures on their proposed SNSs. Skeels [39] on the other hand studied breast cancer patients' engagement in HI sharing on an online SNS that was built to help them manage their health issues. Both studies were conducted with participants who suffered from chronic health conditions (mental illness and breast cancer). Therefore, while the likelihood of having a health condition for their participants was at its maximum value (100%), the severity of their health conditions was assumed to play a major role in motivating them toward discussing their HI on SNSs.

There are different ways to assess one's overall health status and the severity of his health conditions. A number of studies used the self-reported perceived health status as a reliable measurement of individuals' overall health status [20, 45]. Also, they found a correlation between the perceived health status and the number of health conditions, with those who had "poor" health to have more health conditions. On the other hand, the self-reported assessment of health conditions might not always accurately describe the overall health status. For instance, one might suffer from a number of severe health conditions and yet consider his health to be stable or good, while another person might have a minor health issue and feels completely devastated by his health issue.

Discussing overly personal information on SNSs have been associated with privacy concerns [10, 36]. The nature of SNSs can lead to the diffusion of personal information beyond its intended targets, while resulting in the lack of subsequent control over its exposure [5, 19, 35]. In general, information revelation on SNSs was shown to be influenced by the raised privacy concerns due to both the personal experiences and the negative reports in the media [46]. In the context of SNSs, privacy concerns have been always associated with sharing HI among users [10, 42]. A survey of 1060 U.S. adults found that 63% raised concerns related to publicly sharing their HI on SNSs, while 57% were concerned that their HI might be hacked or leaked from the SNSs [1]. Morris et al. [36] surveyed different types of questions that SNSs users asked their social peers about and found that "health" was a type of topic that people tend to consider too personal.

It has been shown that internet users' privacy concerns and their attitudes toward privacy could highly influence their motivation to disclose personal information to online sites [6, 10, 40]. The Westin privacy index was introduced as a way to meaningfully classify internet users based on their overall attitudes to-

ward privacy and motivations to disclose personal information on the internet [6, 9, 25, 26]. Although being commonly used in literature, the Westin based categorization was criticized for its flaws [25]. Researchers have also raised concerns with respect to the predictive value of the Westin privacy index categorization and its correlation to online information disclosure in specific contexts [12, 44]. They showed that in specific scenarios, users' behavioral intention might not be accurately represented using the Westin categories, while suggesting more fine-grained classifications considering other factors (e.g., consequences). In general, despite the flaws with the Westin privacy based categorization, we believe that the literature provides reasonable evidence to reflect on the "overall" correlations between people's privacy attitudes, as classified by Westin, and their motivation to disclose personal information online.

3. EXPLORATORY STUDY: INTERVIEWS

In an effort to develop a better understanding of users' motivation to share HI on SNSs, 21 chronically ill patients were interviewed about their HI sharing experiences. Following a grounded theory approach [7], we explored participants' HI sharing practices, perceptions, and preferences. We identified the main factors that influenced participants' perceived privacy and motivation to share HI on SNSs. We aimed at answering the general research questions: *Why, how, and with whom patients share their HI on SNSs?*

3.1 Sampling and Participants Recruitment

Following a theoretical sampling approach [7], 21 individuals with chronic health condition(s) were recruited through media advertisements (e.g., craigslist). Potential participants were invited to visit the study webpage, where they viewed details of the study, along with the consent form. To be eligible for the study, participants must be: 19 years of age or older, living in Metro Vancouver, Canada, maintaining at least one active account on an SNS that they visited regularly, and having at least one chronic health condition. Participants were compensated with $25 (CAD) for taking part in the study.

A total of 21 participants were interviewed throughout the study. The purposive sampling of participants who had chronic health condition(s) assured their involvement in HI sharing practices. The sample included 7 women and 14 men, between 21 and 68 years old. Participants came from diverse ethnic backgrounds but all were speaking English fluently. A summary of participant demographics is given in Table 1. Participants had different health conditions, including physical, mental, or a combination of both. Details about participants' health conditions are presented in Appendix A.2.

3.2 Data Collection

Data collection was done by means of audio recorded interviews during the months of February-May, 2014. The semi-structured interviews lasted approximately one hour each. Interviews were conducted in different locations to meet participants' needs and requirements (e.g., at participant's home due to his disability and limited mobility). An interview guide was developed to help in managing the interview flow and assuring purposeful data collection (Appendix A.1). Participants were always invited to tell their stories according to their style and conventions. Data collection was directed by a theoretical sampling approach, where new data was collected and analyzed to elaborate and refine the identified themes respectively [7]. After analyzing 16 interviews, the total number of identified unique codes reached a plateau where analyzing further interviews did not result in new findings (Figure 1). Data collection was stopped after conducting 21 interviews, when "theoretical saturation" was achieved in the analysis.

Table 1: Participants demographics.

Demographic	Category	Count (N=21)
Gender	Male	**14**
	Female	7
Age range (21-68)	19-30	2
	31-40	**9**
	41-50	5
	50+	5
Completed Education	High School	3
	Some college/university	6
	Post secondary diploma	**7**
	University (BSc., MSc.)	5
Health Conditions	Physical (e.g., heart disease)	**14**
	Mental (e.g., post-traumatic stress disorder)	3
	Physical and Mental	4

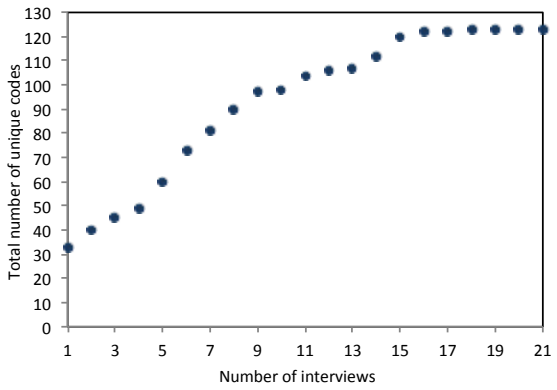


Figure 1: Data analysis and theoretical saturation (21 interviews and 123 unique codes).

3.3 Analysis

The analysis process started immediately after transcribing the first interview and continued concurrently with the data collection process. The recorded interviews were transcribed verbatim by professional transcriptionists. Interview transcripts were anonymized by removing identifiable information (e.g., names). NVivo (Version 10.1) was used as the main qualitative data analysis tool for coding and analyzing the transcripts.

Constant comparison of coding and analyzing data through several iterative stages of *open*, *axial*, and *selective* coding were applied [7]. Open coding was initially used to identify, describe, and categorize interesting phenomena that were found in the data. The first set of transcripts were read line by line and coded accordingly, resulting in 90 unique codes after analyzing 8 interviews. At that point, we also started to look at interrelated codes that formed meaningful categories (axial coding). The identified categories were as following: *perceived privacy, perceived benefits, the recipients of the shared HI, used technologies, prior HI sharing experiences, HI sharing motivation, trusted entities, anonymous communication, HI sharing preferences, and health status*. Then, we identified participants' *Motivation to*

share HI on SNSs as the core category (selective coding). We also identified the following sub-categories: *perceived privacy, perceived health status, the recipients of the shared HI, prior HI sharing experiences, and health status*. The transcripts were further analyzed by selectively coding new data that was related to the core category until theoretical saturation was reached. Finally, memoing was used frequently to describe coded events, as well as explain observed concepts and their relations.

The analysis resulted in a total of 2,521 coded excerpts, with an average of 120 coded excerpts per interview. The quality and consistency of the analysis was checked by a second researcher, who reviewed and coded a total of 100 randomly selected excerpts using our generated codes. The two coders reached about 90% agreement.

3.4 Results

3.4.1 HI Sharing Practices on SNSs

While participants used a variety of SNSs, all participants were active users on Facebook. Considering the fact that more than 71% of North American internet users are on Facebook,[2] it is not an anomaly to have all participants to be Facebook users. Participants indicated going on their SNSs on regular basis. Moreover, participants indicated using a variety of sites to share or seek HI online (e.g., SNSs, blogs). Despite the fact that the identified sites were not designed to support HI sharing among social peers, the majority of participants recalled sharing HI instances on them in the past:

"I've got a lot of pictures on Facebook of when I was in hospital. I had pictures of myself, my scar, and everything else. All of those are on my Facebook."–P2 (M, 59, fractured back/defective knee)

3.4.2 Perceived Benefits

From simply sharing how a person feels at a specific moment, to sharing detailed information about treatments, participants experienced sharing HI with select individuals or groups in the past. Participants shared their HI with others for the sake of getting benefits. The benefits of sharing HI include but not limited to: learning from the shared experience, initiating conversations with online peers, justifying specific behaviors, reaching out to others who had similar health conditions, and engaging in social support. Moreover, participants showed interest in helping others by providing social support, empathy, and experience-related feedback. It was also interesting to see that regardless of the expected reactions and responses, some participants felt relieved simply by talking about their problems with others:

"I feel better letting them know. Whether they understand or not, I feel relieved telling them."–P15 (M, 37, bipolar depression/anxiety)

3.4.3 The Recipients of the Shared HI

A number of participants (5/21) shared detailed HI with select family members and/or close friends via online services (e.g., email, SNSs). For instance, P21 (F, 35, herniated disks at L4-L5) used Facebook occasionally to communicate her health issues with her friend, who happened to be an experienced therapists, and tried to ask for her opinion and advice. Generally speaking, while participants preferred to have in-person discussions of their health issues with other friends and family members, the online services have provided them with a convenient way of communication, especially when physically distanced from friends and/or family members:

"I do [talk about health on Facebook], and especially with my wife [who lives in a different region]. Because my wife is a nurse so, rather than going to a doctor, she would be somebody that I would talk to first."–P12 (M, 59, degenerative disc disease/brain injury)

In addition to close friends and family members, participants shared their HI with others who had been through similar health experiences. In fact, they believed that the mutual health experiences had helped them in understanding each others and communicate with less effort:

"I talk about all kinds of things I'd never talk to my able-bodied friends about, because these people know what our lives are like. Our lives are all different but they have a commonality that doesn't exist with able-bodied people."–P4 (F, 68, C4-C5 quadriplegic)

3.4.4 Perceived Health Status

Participants developed an overall perception of their health status based on their knowledge of their health conditions and their perceived control over its outcomes. For instance, P19, who suffered from HIV, considered his health condition as yet another manageable disease that required only few tweaks to his life style:

"Totally manageable. You got to watch your cholesterol, watch your liver, take two pills in the morning, one at night, and that's it."–P19 (M, 50, HIV)

This was mainly because he was completely aware of his condition, its complications, and the necessary ways to control it. Interestingly, participants who perceived their health status to be "manageable" were found to be less motivated to engage in sharing their HI on SNSs. Moreover, participants who suffered from chronic pain due to physical injuries and/or arthritis (9/21), considered their health status to be stable and "manageable." As such, they showed less interest in using SNSs for sharing their HI with other people.

On the other hand, P9 (F, 42), who suffered from a rare disease called Neuromyelitis optica (NMO), was heavily engaged in sharing her HI on blogs and SNSs (e.g., MS society, Twitter, Facebook). She described a number of reasons for her enthusiasm toward sharing her HI online: helping newly diagnosed patients, finding new information about the disease, and participating in research. Moreover, the insufficient scientific knowledge about the health condition, and the relatively small population of diagnosed patients with similar health condition, were also motivating her to actively engage in online HI sharing activities.

3.4.5 Perceived Privacy

Despite the perceived benefits, users' attitude toward disclosing HI on SNSs is also influenced by their perception of privacy (privacy calculus) [24, 31]. By exploring users' HI sharing practices and preferences, we tried to develop a better understanding of the factors that shaped chronically ill patients' perception of privacy when sharing HI on SNSs. In what follows, we highlight some factors that contribute to users' perception of privacy.

The Shared HI.

To minimize the privacy concerns related to sharing HI on SNSs, the majority of participants tried to keep their shared information very general, with the least details about their personal health. Moreover, participants altered their HI sharing behaviors with respect to the audience in different SNSs. For instance, while P11 (M, 40, L3-L4 fusion) shared information

[2] http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/

about his back injury with a group of people who suffered from similar injuries on LinkedIn, he tried to maintain his professional image by not posting irrelevant and unprofessional details on LinkedIn (e.g., feelings and emotions, certain drug usage). Some participants on the other hand, avoided talking about their health issues on their SNSs because their social peers had not experienced similar health issues. They felt that their social peers might misread their situation and make judgments:

"I'm not the kind of guy that is just going to throw it out there [on Facebook] and get responses from anyone on a medical issue. I think it's just common sense like, people judge. People rumour. You don't want to throw out a bunch of stuff that's going to be misconstrued."–P1 (M, 38, chronic sciatica).

It was interesting to see that some participants considered the shared HI instances to be mostly of abstract nature. For instance, P2 (M, 59, chronic back/knee pain) shared pictures of himself and his scar on Facebook while staying at the hospital. Some participants believed that the shared HI contained no important details about them, and therefore, they did not mind sharing it with their social peers:

"I posted pictures of the brace that I had on Facebook. There's no story behind it, it's just like a picture, like "oh, this is gross", you know?"–P10 (M, 37, osteoarthritis)

Health Conditions and Associated Stigma.

Participants shared general aspects of their health on different SNSs. Moreover, they were likely to share information related to their physical health conditions (e.g., injuries, chronic pain, arthritis), as compared to other types of HI. Participants showed more discomfort when sharing information related to their mental health. In fact, the stigma associated with such health issues stopped them from sharing their HI with specific audiences, especially with people whom they knew. Furthermore, we noticed that some male participants were less motivated to talk about their health issues with friends and family members on SNSs. They believed that there is a stigma of men talking about their health, especially mental health issues:

"It's just a stigma of men not talking about stuff like that. With you, you're a complete stranger and I'll never see you again probably. So, it's not that hard for me to be candid and open about. But with friends, I'm always worried about bumping into people I've known."–P20 (M, 50, depression/chronic neck pain)

Few female participants also raised concerns about sharing information related to their mental health issues on their SNSs. For instance, P5 used an online website (reddit)[3] to anonymously ask questions related to her depression. She also tried to maintain her privacy by hiding her reddit name from her friends. Another way of avoiding the stigma associated with sharing information regarding mental health was to engage in online discussions and express feelings and health issues in an indirect manner. For instance, P15, who suffered from chronic depression, talked about his mental health issues by posting philosophical questions on his blog and Facebook page. He used these questions as a way to indicate his willingness to talk about his feelings and mental health issues with others.

Anonymity and Online Identity.

Despite the existing concerns of sharing HI with known people, P20 (M, 50, depression/chronic neck pain), for instance, shared his health issues with a couple of friends on Facebook. Those friends were not living in the same city where he lived in and therefore, there was a very little chance of running into them on a typical day. Interestingly, few participants indicated that the physical distance had provided them with some level of privacy, and therefore, they felt more comfortable to communicate their health issues with physically distanced people through Facebook:

"Even if I'm using my real name, it still feels kind of anonymous because they're not right there beside me. I'm not looking at them while I'm talking to them. They could be in Sydney, Australia."–P16 (M, 48, post-traumatic stress disorder)

As described by P16, his anonymity was maintained by keeping himself physically unreachable. The importance of the physical proximity in defining privacy in the online space was clearly present in participants' responses during the interviews. Participants also raised serious concerns with regards to revealing their physical location in online environments. Regardless of their healths status, participants wanted their current and/or future location to be kept strictly confidential. Moreover, even though participants did not mind being in the same virtual space with several other people (e.g., friends, acquaintance, possible strangers), they were concerned with the unexpected presence of their social peers in their physical proximity. As a result, some participants tried to hide their location information from different recipients while posting information on the SNSs:

"You just don't know who's reading it [online posts]. So, I don't want to say: "Oh, I'm going to Location today," and I get there and then there's somebody there. It would just be creepy. So, for things like that, I will post later: "Hey, went to Location today." So, it's done and I'm back home now."–P9 (F, 42, NMO).

SNSs Vulnerabilities and HI Misuse.

A number of participants perceived existing SNSs to be vulnerable to privacy and security exploits and therefore, risking the confidentiality of their information and increasing the chances of undesirable information disclosure. On top of that, some participants were also concerned about the probability of having their HI being misused by insurance companies and some governmental agencies. Participants recited several stories about themselves and other individuals in their social networks that became victims of shared information misuse. For instance, P7 (M, 54, quadriplegic) was overwhelmed by the attempts made by insurance companies towards cutting disability benefits by misusing patients' shared information on their SNSs. Moreover, both P1 (M, 38, chronic sciatica) and P5 (F, 30, chronic depression) raised concerns with respect to sharing specific "risky" information regarding their health on their SNSs, especially if their behavior was classified as illegal in a different jurisdiction (e.g., licensed drug consumptions). Furthermore, P5 was worried about being denied access to the U.S. if she shared information about being hospitalized for depression or bi-polar disorders. She mentioned knowing over a dozen of stories about people who were turned away from the U.S. borders just because they shared similar HI on Facebook.

3.4.6 HI Sharing Preferences

Preferred User Groups.

Participants identified three main user groups, with whom they were willing to share their HI on SNSs:

1. *Doctors and Health Professionals.* Almost all participants preferred to have their doctors involved in their SNSs. They are the source of information, advice and medical care. Also, they

[3] http://www.reddit.com/

have the knowledge, experience, and the authority to initiate health management decisions [4]. As a result, having them in any SNS that will be used for sharing HI might be preferable.

2. *Select Friends and Family Members.* Participants preferred to keep their family members and friends updated about their overall health status. More importantly, participants indicated their interest in sharing further details of their HI with select friends and family members. However, the nature and the level of details of the shared HI was dependent on the mutual health experiences and the closeness of their relationships. Participants were also open to discuss details of their health issues with those friends and family members who had expertise in the medical field.

3. *Others with Similar Health Condition(s) and Experience(s).* All participants identified the importance of having access to a pool of people who had gone through similar health issues. Due to the mutual experiences, the perceived benefits were higher when communicating HI with others who had gone through similar health experiences. It was also important for participants to consider other mutual factors (e.g., age, ethnicity, treatments) when deciding to share their HI with other people.

The SNS Environment.

For the majority of participants, it was important to know who owns/operates the SNS. Most of the participants (20/21) considered the government and/or their doctors' offices to be the most reliable and trusted entities with their HI. Moreover, participants did not necessarily trust private companies with their health records, unless recommended by their doctors. The ability to maintain an online version of their health records in the SNS was essential to all participants. Nevertheless, participants required to have their health records fully contained in the SNS environment. Participants preferred to keep their health records private and not shared with other users. Participants also required adequate security measures for protecting their stored data (e.g., using proper encryption).

Communication and HI Presentation.

To maintain their boundaries while communicating with strangers, participants required having anonymous communication capabilities in the SNS. Anonymity does not necessarily mean hiding all personal information. In fact, the majority of participants did not mind revealing their first name and their city of residence. However, the anonymity was necessary to maintain privacy by managing the identity and hiding some HI from other social peers. In general, participants preferred to perform one-to-one communications whenever they wanted to discuss details about their health with other social peers. Participants also indicated their need to maintain the way their HI was viewed by others. For instance, while participants did not want others to view every detail of their HI, they did not mind sharing an aggregate view of their HI with others who had similar health conditions (e.g., viewing progress updates during a course of treatment).

3.4.7 Results Summary

We interviewed 21 SNS users who had chronic health conditions about their HI sharing practices and risk perceptions. We explored their prior experiences with sharing HI on SNSs while inquiring about their preferences for the ideal HI sharing environment. We highlighted the main factors that related to users' motivation to share HI on SNSs (perceived benefits, perceived privacy risks, and perceived health status). We also showed that the recipients of the shared HI can influence users' perceived benefits and perceived privacy risks. Furthermore, we characterized the preferred recipients of the shared HI (people with medical expertise, mutual health experiences, and/or strong social ties). Finally, we discussed requirements for creating a trusted SNS environment that facilitate HI sharing among social peers (e.g., anonymity, trusted owner/operator, HI communication/presentation).

4. CONFIRMATORY STUDY: ONLINE SURVEY

We conducted an online survey to confirm our findings from the exploratory study. The online survey consisted of a mixture of close- and open-ended questions. The survey gave us the opportunity to reach a larger sample of SNSs users, which in return helped in achieving more generalizable findings.

4.1 Why Facebook?

Results of our previous interview study indicated that the majority of participants were Facebook users (Section 3.4). Facebook is one of the few SNSs that have been extensively studied by social and computer scientists. This has resulted in a good understanding of how it is generally used and for what purposes [13,27,28]. Facebook is also the most popular SNS today, consisting of more than a billion users, with a large user population that goes on Facebook on daily basis.[4] As of August 2015, Facebook remains by far the most popular SNS in the U.S., with 72% of online adults to use Facebook (62% of all adults in the U.S.) [11].

4.2 Participants Recruitment

Participants were recruited via Amazon Mechanical Turk (MTurk),[5] which is a crowdsourcing website that provides a reliable source of high-quality data for research involving human-subjects [41]. A respondent was expected to finish the survey in less than 30 minutes. To ensure quality data collection and analysis, we used MTurk's features to recruit participants who had successfully completed 100 tasks or more on MTurk while having a minimum approval rate of 95%. Participants were limited to a single submission only. Participants were compensated with $1 (U.S.) through MTurk for successfully completing the survey. To ensure successful compensation on MTurk, participants were required to submit a unique code, which was assigned to them after completing the survey.

4.3 Data Collection

A total of 537 participants responded to the online survey between October 16–23, 2015. As shown in Appendix B.1, the online survey consisted of the following items: (1) demographics and background; (2) health conditions and perceived health status; (3) previous HI sharing experiences; (4) motivation to share HI on Facebook; (5) preferred recipients of the shared HI; (6) anonymous online identity; (7) trusted SNS providers; and (8) attitudes toward privacy. The average completion time was approximately 10 minutes, with an overall survey completion rate of 96.5%. Responses were closely examined based on completion time. Submissions that lasted less than 4 minutes were fully examined to ensure quality of the provided responses. Finally, to insure consistency of the sample, and avoid the effects of cultural differences, submissions made from people residing

[4] http://newsroom.fb.com/company-info/
[5] www.mturk.com

outside of the U.S were excluded. The remaining 492 submissions were included in further analysis through the study.

4.4 Data Analysis

The survey was employed using our university's online survey tool. We used MS Excel and SPSS (Version 23.0) to perform statistical analysis on the data. We also used NVivo (Version 10.1) for coding and analyzing qualitative text responses. Descriptive statistics were used to explain the underlying properties of the collected data (e.g., mean, SD), while a number of inferential statistic analysis tests were used to highlight correlations and significant differences among groups (e.g., person's correlation). A series of between-subjects tests were used to explore participants' motivations and perceptions. Non-parametric statistics were used when the normality of the data was not assumed, especially with ordinal data (e.g., Likert-Scale). We used Kruskal-Wallis test for comparing $k-$independent samples, with post-hoc pair-wise comparisons using Mann-Whitney U tests (if necessary). We also employed Friedman's test and/or Wilcoxon signed-rank tests to check for statistically significant differences in participants' responses when repeated measurements were collected from the same participants (within-subjects).

4.4.1 Privacy Attitudes

Westin explored people's attitudes and concerns toward a number of privacy-related topics by conducting several surveys since 1978 (e.g., confidence in organizations that handle personal information). In order to summarize results and highlight trends in privacy, Westin created "privacy indices" for most of his surveys (e.g., General Privacy Concern Index, Computer Fear Index). Despite its flaws [12, 25, 44], the Westin privacy index has been used as an indicator of internet users' general attitudes toward privacy and their motivation to disclose personal information online [6, 9, 26]. According to Westin, people could be categorized based on their overall privacy attitudes, as follows: (1) privacy *Fundamentalists*, who highly value privacy and feel very strongly about it; (2) privacy *Pragmatists*, who have strong feelings about privacy but can also see the benefits from surrendering some privacy in situations where they believe they can prevent the misuse of their information; and (3) privacy *Unconcerned*, who have no real concerns about privacy or about how other people and organizations use their information [25].

In this study, we modified the statements typically associated with the Westin privacy index in order to fit them into the context of HI sharing on SNSs. We replaced the words "consumers" and "companies" with "internet users" and "social networking sites" respectively (as shown in Appendix B.1.8). Inspired by the Westin categorization procedure, we used participants' responses to the modified statements to group them into people with high, medium, or low privacy concerns, as corresponding to privacy *Fundamentalists*, *Pragmatists*, and *Unconcerned* categories. About 54% of participants were categorized to have high privacy concerns, while approximately 34% and 12% of participants were categorized to have medium and low privacy concerns respectively.

Contextualizing the Westin privacy index (e.g., by using brand names) can have a significant effect on the categorization outcomes [44]. Therefore, although we used a categorization procedure similar to Westin, we do not know how the modifications to the original Westin privacy index have impacted our analysis, as compared to using the original Westin statements. Nevertheless, we believe that our categorization could be of interest to the community. In fact, our categorization proportions were very close to those presented in Woodruff et al. [44], where they implemented the Westin privacy index to categorize MTurk workers (49% *Fundamentalists*, 40% *Pragmatists*, and 10% *Unconcerned*). In general, our sample included a larger number of participants with high or medium privacy concerns, as compared to the general population [23, 44].

To corroborate our categorization outcomes, we asked participants to indicate the privacy-preserving actions that they had performed on Facebook (e.g., changing profile visibility). Participants selected all that applies from a list of 10 common privacy-preserving actions (Q34 in Appendix B.2). On average, participants performed 7.16 privacy-preserving actions in the past ($\sigma = 2.63$). About 26% of participants performed all 10 privacy-preserving actions on Facebook. The correlation analysis using Spearman's test showed a negative correlation between the number of performed privacy-preserving actions and participants' attitudes toward privacy ($r(490) = -0.176, p < 0.001$). This supports our analysis of the Westin inspired categorizations, which relates participants with higher privacy concerns to performing more privacy-preserving actions, as compared to those with lower privacy concerns.

4.5 Results

4.5.1 Demographics

We analyzed responses from 492 participants residing in the U.S. with ages ranging between 19 and 74 years ($mean = 34.7$ and $\sigma = 10.8$). A summary of participant demographics is presented in Table 2. The sample consisted of almost equal number of male and female participants, with a wide range of employment categories including Students (32/492) and Unemployed (72/492). While about 75% of participants were younger than 40 years old, almost half of all participants were between 19 and 30 years of age (46.1%). About 60% of participants completed a post-secondary degree (e.g., Diploma, Bachelor's, Master's, or PhD). Approximately 20% of participants indicated having a degree and/or work experience in fields related to Computer/IT. Furthermore, the vast majority of participants (91.1%) spent more than two hours on the Internet on daily basis ($mean = 6.5$ and $\sigma = 3.4$). These demographics reflect the nature of MTurk workers, who were shown to be highly active internet users with higher education levels and younger ages than the general population [41].

We also asked participants about their Facebook usage. About 97% of participants have been on Facebook for at least 4 years ($mean = 7.7$ and $\sigma = 2.3$). On average, participants had approximately 289 Facebook friends ($min = 0, max = 3165$). The majority of participants (98.8%) were checking their Facebook account at least once a week, while 84.6% of all participants checked their Facebook on daily basis. Participants were asked to describe their Facebook friends by selecting all that applies from a list of categories. Family members and relatives, offline friends, colleagues/co-workers, and friends' friends represented the top four friends' categories. A comparison of participants' Facebook usage frequency and friends' demographics with Pew research centre's recent report shows that our sample is in fact representative of U.S. Facebook users with slightly more active participants, which is typical for MTurk workers [11].

4.5.2 Perceived Health Status

We asked participants about their overall health status and existing health conditions. Only 73 participants (14.8%) did not have any chronic health conditions while the remaining 419 par-

Table 2: Participants demographics ($N = 492$).

Demographic	Category	Count	(%)
Gender	Male	**246**	**50.0**
	Female	245	49.8
	Unspecified	1	0.20
Age range (19–74)	19–30	**227**	**46.1**
	31–40	145	29.5
	41–50	62	12.6
	51+	58	11.8
Completed Education	Undergraduate University (Bachelor's)	**208**	**42.3**
	Some college/university courses	136	27.6
	Graduate University (Masters's/PhD)	58	11.8
	High School	51	10.4
	Diploma (post-secondary courses)	33	6.70
	Less than High School	3	0.60
	Other	3	0.60
Employment Categories (Top 5)	Business, management, or financial	**65**	**13.2**
	Services (e.g., retail)	62	12.6
	Computer engineer, IT professional	41	8.30
	Administrative support	34	6.90
	Education (e.g., teacher)	33	6.70

ticipants (85.2%) reported 55 different health conditions. Allergies, anxiety, depression, stress, arthritis/chronic pain, asthma, obesity, diabetes, heart disease, and cancer represent the most frequent health conditions reported by participants (Figure 6 in Appendix B). About one third of all participants (33.9%) suffered from one chronic health condition, while slightly over half of all participants (51.2%) reported two or more chronic health conditions. Among participants who had chronic health conditions ($n = 419$), the majority (96.9%) reported having the chronic health condition(s) for at least two years.

Participants were asked to identify their perceived health status on a 4-point Likert scale ("poor", "fair", "good", and "excellent"). A number of studies showed that the self-reported health status could be considered as a reasonable indicators of one's overall health [20, 45]. Despite that, in Section 3.4.1 of the exploratory study, we discussed that patients' perceived health status could be influenced by their perceived control over their health conditions. In line with our previous findings, we noticed that 253 of the online survey participants (about 51%), had one or more health conditions and yet perceived their health status to be "good" or "excellent." Furthermore, 13 participants reported "fair" health status without having any health conditions. Therefore, we used a combination of the self-reported health status and the number of health conditions in order to group participants into three meaningful categories: (1) *Healthy* (14.8%), individuals who had no chronic health conditions; (2) *Manageable* (51.4%), individuals who had at least one chronic health condition and perceived "good/excellent" health status; and (3) *Unhealthy* (33.7%), individuals who had at least one chronic health condition and perceived "fair/poor" health status. We believe that these categories provide a better representation of participants' overall health, and therefore, we used them for further comparison of participants' behaviors according to their health status.

We also explored the relationship between participants' privacy attitudes and their perceived health status. While the correlation analysis was marginally significant (p=0.035), the resulted correlation coefficient was very small (r=-0.095). Therefore, we did not include this relationship in further analysis.

4.5.3 HI Sharing Experiences

We asked participants to indicate if they ever shared details of their health information with different people on Facebook. About half of participants (48.6%) never shared their HI on Facebook. Among the remaining participants, 71.1% indicated sharing their HI with "some close friends or family members," while 37.9% shared their HI details with "select friends who had medical expertise and/or mutual health experiences." Furthermore, we asked participants to evaluate their prior HI sharing experiences on Facebook (*Positive*, *Negative*, *Both* positive and negative, or *Neither* positive nor negative). Three participants were not able to provide an evaluation for their prior HI sharing experiences on Facebook. Among the remaining 250 participants, more than half of them (57.7%) evaluated their prior HI sharing experience to be *Positive*, while about 18.2% had *Both* positive and negative experiences. It is interesting to see that only 8 participants (3.2%) indicated having only *Negative* experiences, while the remaining participants (19.8%) indicated *Neither* positive nor negative experiences. We also asked participants to explain in their own words why they thought that their experiences were *Positive* or *Negative*. In general, *Positive* experiences were related to gaining benefits (e.g., positive social support), while *Negative* experiences resulted mainly from the lack of benefits (e.g., impractical advice) or privacy concerns (e.g., over-sharing one's HI, judgments). Detailed analysis of participants' responses is presented in Appendix B.3.1.

4.5.4 Motivation to Share HI

Participants were asked to indicate the reasons that might motivate them to share their HI on Facebook by selecting all that applies from a list of common reasons. About 41.7% of participants considered Facebook as a place for seeking social support from friends and family whenever necessary. About one third of participants (33.5%) were motivated to share their HI on Facebook in an exchange for other people's expertise and experiences. Furthermore, 32.3% of participants were motivated by their previous positive experiences. It is also interesting to see that 28.9% of participants were passionate to help others by sharing their own health-related experiences on Facebook. This highlights the two-way nature of information sharing on SNSs, where some people tend to generate and disseminate content for the rest of the population. Finally, it seems that the lack of knowledge about the health issues, and the fact that Facebook could help in connecting to other people with similar health issues, were also motivating about 20% of participants to share their HI on Facebook.

Prior HI Sharing Experiences.

A series of Mann-Whitney U tests resulted a statistically significant difference among participants' willingness to share HI on Facebook when compared based on their prior HI sharing experiences, with mean ranks of 328.3 and 160.0 for the two groups respectively ($p < 0.001$ and large effect size $r = 0.61$). This indicates that those who had previously shared their HI on Facebook are more willing to share their HI on Facebook in the future. To investigate further, we used participants' evaluation of their prior HI sharing experiences to group them into the following categories: (1) *Positive*, those with only positive experiences; (2) *Negative*, those with only negative experiences; (3) *Both*, those with both positive and negative experiences; and (4) *Neither*, those with neither positive nor negative. A Kruskal-Wallis test followed by a series of pair-wise comparisons using Mann-Whitney U tests showed statistically significant differences for all pair-wise comparisons except when comparing *Both* and *Neither*

groups. The results showed that having only positive experiences in the past can highly influence the motivation to share HI in the future. Moreover, participants who had only negative experiences were also shown to be less motivated to share their HI details on Facebook, as compared to other groups.

Privacy Attitudes and Motivation to Share HI.
Participants were grouped based on their privacy attitudes (high, medium, or low privacy concerns). A Kruskal-Wallis test showed a statistically significant difference in willingness to share HI on Facebook ($\chi^2(2) = 33.42, p < 0.001$), with mean ranks of 218.4, 263.5, and 325.1 for participants who had high, medium, and low privacy concerns respectively. The pair-wise comparisons using Mann-Whitney U tests showed significant differences between all three groups, with $p \leq 0.001$ for all pair-wise comparisons ($r_{1-2} = 0.168$, $r_{1-3} = 0.287$, and $r_{2-3} = 0.226$). This confirms that people with higher privacy concerns are less willing to share their HI on Facebook, as compared to those with lower privacy concerns.

Health Status and Motivation to Share HI.
To investigate the effect of health status (*Healthy*, *Manageable*, and *Unhealthy*) on the motivation to share HI on Facebook, we conducted a of Kruskal-Wallis test. The test showed a statistically significant difference in the motivation to share HI details on Facebook among the three groups ($\chi^2(2) = 8.11, p < 0.017$), with mean ranks of 241.4, 242, and 267.4 respectively. Furthermore, the pair-wise comparisons using Mann-Whitney U tests showed a significant difference in the motivation to share HI on Facebook between *Healthy* and *Unhealthy* groups only ($p = 0.007$ and $r_{1-3} = 0.176$). This conforms with prior findings that associated online HI seeking/sharing activities with the overall health status and the number of health conditions [11, 16, 45]. A closer look at the participants shows that about 91% of those who were motivated to share their HI on Facebook were categorized as *Unhealthy* or *Manageable*. This might also be a good indication on the influence of health status on users' overall motivation to share HI on Facebook.

4.5.5 Preferred Recipients of the Shared HI

We asked participants to indicate their willingness to share their HI with different recipients on Facebook. As shown in Figure 2, about 67% of participants considered sharing their HI with "some close friends and/or family members," while about 65% considered sharing their HI with "friends and/or family members who had medical expertise and/or mutual health experiences." On the other hand, about 73% of all participants did not consider sharing their HI publicly with "all their Facebook friends." Furthermore, about 53% of participants did not consider sharing their HI with strangers through Facebook, even if they had "expertise in the medical field or mutual health experiences." Within-subjects comparison of the repeated measures showed that participants were significantly more willing to share their HI with "close friends and/or family members" and "friends/family who had medical expertise and/or mutual health experiences," as compared to other recipients. Moreover, while the "closeness" of the relationships among friends and family members was shown to influence their motivation to share HI with each other, the "medical expertise and/or mutual health experiences" were also considered as important motivating factors that encouraged people to share their HI.

To extend our investigation, we compared participants' willingness to share HI with different recipients on Facebook.[6] We found statistically significant differences in participants' willingness to share HI with all recipient groups when compared based on their prior HI sharing experiences and privacy attitudes. However, when comparing participants based on their health status, we only found a statistically significant difference in their willingness to disclose HI with "non-friends Facebook users who had medical expertise and/or mutual health experiences" ($\chi^2(2) = 7.43, p = 0.024$), with mean ranks of 208.6, 247.9, and 261 for *Healthy*, *Manageable*, and *Unhealthy* groups respectively. We found that *Unhealthy* participants were significantly more willing to share their HI with non-friends Facebook users as compared to *Healthy* participants ($p = 0.007$ and $r = 0.18$). The results suggest that while participants' health status was not a determining factor when sharing HI with friends and family members, it might have influenced participants' motivation to share HI with non-friends Facebook users.

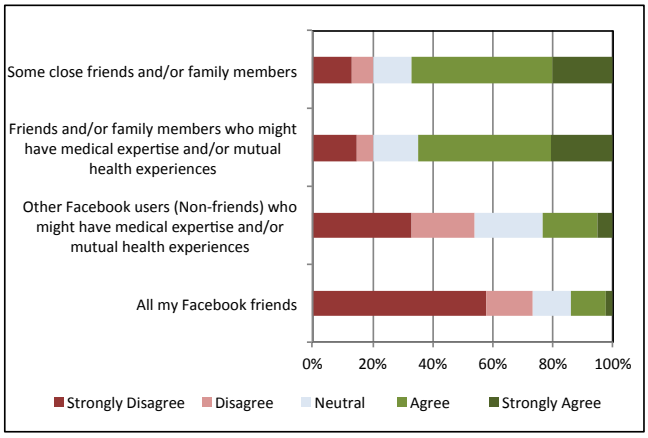


Figure 2: Willingness to disclose HI to different recipients on Facebook.

4.5.6 Willingness to Search for Specific Users

In a hypothetical situation, participants were asked to identify their willingness to use customized search features that could help in finding other Facebook users who had "mutual health experiences" or "expertise in the medical field." Between 32-29% of all participants were "(Very) Likely" to use the search features to find other Facebook users who had "mutual health experiences" or "medical expertise" respectively. On the other hand, about half of all participants were "(Very) Unlikely" to do the same. Within-subjects comparison of participants' willingness to use the search features for finding different users showed that participants were significantly more willing to search for other Facebook users who had mutual health experiences, as compared to users with expertise in the medical field ($p < 0.001$ and $r = 0.2$).

We compared participants' willingness to use the search features to find other users on Facebook by performing a series of between-subjects tests. We found that participants' who had *Positive* experiences to be more likely to use the search feature as compared to those who had neither positive nor negative experiences. When comparing participants' willingness to use the

[6]The "All Facebook friends" group was excluded from the pairwise comparisons since it was not representing specific recipients.

search features based on their privacy attitudes, we found statistically significant differences among all groups, with participants who had high privacy concerns to be significantly less likely to use the search features as compared to those who had medium or low privacy concerns.

4.5.7 Anonymous Identity

We asked participants to indicate their willingness to use an anonymous online identity for sharing their HI on Facebook. About 47% of participants were "(Very) Unlikely" to use an anonymous identity when sharing their HI. On the other hand, about 36% of participants were "(Very) Likely" to do so. A between-subjects comparison of participants' willingness to use anonymous identities for sharing HI on Facebook showed that participants who had IT/Computer knowledge were significantly more willing to use anonymous identities on Facebook, as compared to those who had no IT/Computer knowledge ($p = 0.036$ and small effect size $r = 0.1$). Furthermore, our comparisons showed that participants with medium privacy concerns were significantly more willing to use anonymous identities for sharing their HI on Facebook than people who had high or low privacy concerns ($p = 0.016$ and $p = 0.015$). This however might be due to the pragmatic nature of people with medium privacy concerns, who might be more willing to mitigate risks in exchange for the expected benefits.

Moreover, we were unable to find statistically significant difference in the willingness to use anonymous identities for sharing HI on Facebook, when comparing participants based on their health status. This means that regardless of participants' health status, their motivation to use an anonymous online identity for sharing HI on Facebook is mainly influenced by their privacy attitudes and their IT/Computer knowledge and experience.

Participants were asked to indicate their willingness to "hide" different personal information when creating their anonymous identity that would be used for sharing HI with strangers. As shown in Figure 3, about 95% of participants were "(Very) Likely" to hide their residential address and phone number, while approximately 90% preferred to hide their current/future location information, identifiable profile picture, email address, and last name. On the other hand, slightly over 60% of participants were "(Very) Unlikely" to hide their gender. Also, it is interesting to see that while 29% of participants were "(Very) Likely" to hide their health condition(s), about 50% of all participants were "(Very) Unlikely" to do so.

We performed Principle Component Analysis (PCA) in order to reduce the correlated personal information items presented in Figure 3 into fewer meaningful components.[7] The analysis showed that about 66% of the cumulative variance was described by selecting three components, as shown in Table 3. We considered an information item to be a part of a component if it had a factor loading of at least 0.6 for the particular component and a factor loading under 0.4 for the other components. Moreover, KMO and Bartlett's tests showed adequate sampling and statistically significant correlations that were appropriate for using PCA ($KMO = 0.87$, $p < 0.001$, $df = 91$).[8]

As shown in Table 3, twelve information items were grouped into three components, while the remaining two items did not conform to any particular component (occupation and employment, and city of residence). We named the identified components as following: (1) *Contact and location* information, which consisted of information that could be used to directly reach an individual (e.g., phone number, residential address); (2) *Demographic* information, which consisted of information that were not considered to be identifiable by themselves but could be used to describe properties of an individual in real life (e.g., age, gender, hobbies); and (3) *Identity* information, which represented information that could lead to revealing one's real identity (e.g., picture, first/last name). We created an index variable for each component by averaging participants' rating for each information item within the component.

Considerably more participants were "(Very) Likely" to hide information related to *Contact and location* information and *Identity* information, with average scores of 91% and 81% respectively. *Demographic information* on the other hand had the least score among all components (average score of 38%), with sta-

[7] PCA with Varimax rotation method was used.

[8] A KMO test result >0.8 is considered as "meritorious".

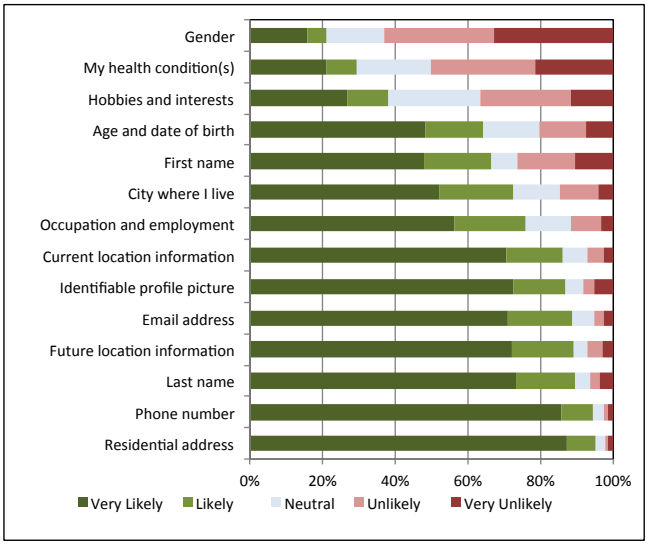


Figure 3: Willingness to hide different information items when creating an "anonymous" online identity.

Table 3: PCA results for different personal information items. The last column represents the percentage of participants who were likely to hide each information item.

Component	Factor loading	Agreement (%)
Contact and location information	—	**91**
Phone number	0.87	94
Residential address	0.80	95
Email address	0.74	89
Future location information	0.73	89
Current location information	0.72	86
Demographic information	—	**38**
Gender	0.83	21
My health condition(s)	0.80	30
Hobbies and interests	0.69	38
Age and date of birth	0.62	64
Identity information	—	**81**
Identifiable profile picture	0.79	87
Last name	0.75	89
First name	0.72	66
Information items that did not conform to any component	—	—
Occupation and employment	NA	76
City where I live	NA	72

tistically significantly fewer participants who were likely to hide their demographic information on Facebook, as compared to identity, contact, or location information. It was also interesting to see that "health conditions" were categorized as *Demographic* information, with about 30% of participants who were likely to hide their health conditions.

By comparing participants' motivation to hide different information based on their privacy attitudes, we found statistically significant differences among all groups with the following two exceptions: (1) motivation to hide *Contact and location* information, when comparing participants who were classified as having medium privacy concerns to those with low concerns. This confirms the relatively higher level of concerns raised by most participants toward revealing their *Contact and location* information, and (2) motivation to hide *Demographic* information, when comparing participants who were classified as having high privacy concerns to those with medium concerns. This might indicate the overall lower concerns with regards to revealing demographic information.

Finally, comparing participants' motivation to hide different information based on their prior HI sharing experiences resulted in a statistically significant difference for hiding *Demographic* information ($p = 0.003$), with mean ranks of 113.8, 182.8, 127.2, and 149, for those who had different experiences (*Positive*, *Negative*, *Both* positive and negative, and *Neither* positive nor negative). Furthermore, the pair-wise comparisons showed that participants who had *Positive* experiences in the past were less likely to hide their *Demographic* information, as compared to other participants.

4.5.8 Willingness to Trust SNSs with HI

We asked participants to indicate their willingness to trust an SNS with their HI, based on its provider. About 27% of all participants trusted an SNS with their HI if it were provided by a governmental health authority, while slightly less than 20% of participants trusted a governmental agency (non-health related) and a recognized private company. On the other hand, about 58% of participants did not trust an SNS if it were provided by a non-health related governmental agency, which was relatively more than the percentage of participants who did not trust a recognized private company and a government health authority (about 52% and 48% respectively). Within-subjects comparisons showed that significantly more participants were willing to trust an SNS with their HI if it were provided by a governmental health authority, as compared to other providers.

We also asked participants to indicate their willingness to trust an SNS with their HI if it were recommended by different people (doctors, friends with mutual health experiences, friends with medical expertise, and close friends/family members). The results of a Friedman's test and the post-hoc comparison using Wilcoxon signed-rank test showed statistically significant differences in participants' willingness to trust an SNS with their HI if it were recommended to them by their doctors, as compared to other people. We imagine that the higher level of trust might also influence participants' willingness to use an SNS for sharing/seeking HI if it were recommended to them by their doctors.

5. DISCUSSION

The results of the online survey showed that participants' overall willingness to share HI on Facebook was linked to the following factors: (1) prior HI sharing experience; (2) privacy attitude; (3) perceived health status; and (4) the intended recipient(s) of the shared HI.

We found that participants' prior HI sharing experiences had a significant correlation with their willingness to share HI on Facebook, with participants who previously shared their HI on Facebook to be more willing to do the same in the future. Furthermore, participants who described their prior HI sharing experience to be *Positive*, were significantly more likely to disclose their HI on Facebook, as compared to participants who had *Negative* experience. By analyzing participants qualitative responses, we found that *Positive* HI sharing experiences were described as online communications with other social peers that benefited the participants (e.g., positive support). It appears that sharing HI on SNS might be a way for some people to initiate conversations and discussions with other social peers, while creating the opportunity toward finding other people that might had similar health experiences. *Negative* experience was mainly due to the lack of gained benefits. Participants were also intimidated by the loss of control over their shared HI in the semi-public SNS environments, and by the fear of oversharing their HI, which might lead to unforeseen consequences such as gossips, rumours, and judgments.

Inspired by the Westin privacy index [25], participants' attitudes toward privacy were used to group them into people with high, medium, or low privacy concerns. In general, we found that higher privacy concern was associated with performing more privacy-preserving actions on Facebook. This indicates that participants who were classified as having higher privacy concerns were willing to put more effort into protecting their online privacy in the context of SNSs. Furthermore, when sharing their HI on Facebook, participants with high privacy concerns were significantly less likely to disclose their HI than the other groups (medium or low concerns). This is inline with findings from previous studies of the influence of users' privacy attitudes on their overall willingness to disclose sensitive personal information on websites [6, 31]. Also, it indicates that HI might be treated as sensitive/personal information by users and therefore, should be handled with extra care in the context of SNSs.

We used participants' health conditions along with their self-reported health status to categorize participants into *Healthy*, *Manageable*, and *Unhealthy* groups. We discovered that *Unhealthy* participants (who had one or more health conditions and perceived their health to be *poor* or *fair*) were significantly more likely to disclose their HI on Facebook than *Healthy* ones (with no health conditions). This is in line with previous findings, which showed that those who perceived their health poor, were more willing to share and/or seek HI online, as compared to people in good health [16, 20, 45]. Furthermore, our participants with *Manageable* health status (i.e., had at least one chronic health condition yet perceived their health to be good or excellent) were somewhere in between *Healthy* and *Unhealthy* in terms of their motivation to share HI on SNSs. We conclude that patients' motivation to share HI on SNSs is linked to their confidence in the level of control over their health conditions, with those who had higher control to be less motivated to discuss their HI issues with other online users.

We explored participants' willingness to disclose their HI with different audiences on Facebook. The results suggest that regardless of participants' health status, they were more willing to disclose their HI to friends and family members than other Facebook users (e.g., non-friends). Moreover, while the "closeness" of the relationship among friends and family members was likely to increase their willingness to share HI with each other, "medical expertise" or "mutual health experiences" appear to be contributing factors that encourage friends and fam-

ily members toward exchanging their HI with each other. On the other hand, participants were less likely to share their HI with non-friends Facebook users, even if those users had expertise in the medical field or had mutual health experiences. At the same time, *Unhealthy* participants were significantly more willing to share their HI with those non-friends who had medical expertise or mutual health experiences, as compared to *Healthy* participants. This indicates that those users who have poor health might be more willing to discuss their health issues with strangers on Facebook, especially if those strangers have expertise in the medical field or mutual health experience.

5.1 Limitations

Individual interviews have few limitations: First, the interview results are limited by participants' experiences with existing HI sharing services. Therefore, we restricted the participation to patients who were also active SNSs users, with at least one SNS account that they used regularly. Second, it is possible that participants indicate some behavioral preferences during the interviews that they are not necessarily practicing in their real lives [3]. To address that, we tried to infer privacy preferences from participants' previous HI sharing practices rather than directly asking them. Third, to address generalizability of our findings, we conducted a followup online survey in order to test our findings with a more representative sample. Finally, to minimize interviewer's biases on both the data collection and analysis processes [21], we asked open-ended questions and tried to probe the participants to tell their story from their own perspectives. Furthermore, we tried to validate our coding scheme by comparing our results to the results of a second researcher who analyzed 100 randomly selected excerpts from the interview transcripts. Ideally, we believe that involving more than two researchers throughout the data collection and analysis will always help in minimizing existing biases.

The main limitation of the online survey was in the self-reported nature of the data, which was difficult to verify in practice. For instance, participants reported a number of health conditions that were difficult to confirm without violating participants' privacy. Furthermore, we used a contextualized version of the Westin privacy index in order to categorize participants according to their privacy attitudes. While our findings might be of interest to the community, a formal validation of our Westin inspired categorization would be necessary before comparing our categories to the Westin based categories.

5.2 Implications for Design

By exploring participants' motivation to use a hypothetical search feature for finding different Facebook users, we found that participants were more willing to search for Facebook users who had mutual health experiences, as compared to users who had expertise in the medical field. Furthermore, while our results showed that *Unhealthy* participants were more willing to share their HI with different user groups, we did not find statistically significant difference in their willingness to use the search features, when compared to participants with *Healthy* or *Manageable* health status. Aside from the reasons behind participants' motivation to use the search features, we believe that SNSs can utilize users' shared HI in order to provide automatic recommendations that could facilitate finding the preferred user groups on behalf of users. For instance, while we showed that patients were less sensitive toward revealing their health conditions when creating their online anonymous identity, we believe that current recommendation systems on Facebook can utilize this information to automatically search and suggest other users who might have mutual health experiences or medical expertise.

Using an anonymous online identity to share HI with strangers was considered to be a preferable option for overcoming the privacy concerns [2, 34]. Similarly in our exploratory study (Section 3), participants considered using anonymous identities to protect their privacy when discussing their health issues with online users, especially strangers. We believe that providing the ability to anonymously share HI on SNS can encourage users, especially people with medium privacy concerns (i.e. pragmatists), to engage in more active HI sharing by regaining some of the privacy surrendered when users disclosed their HI online. In order to maintain anonymity, it is important for users to have the ability to hide *contact and location information* and *identity information* from other users. Furthermore, we imagine that SNSs can also benefit from users' low sensitivity towards revealing their health conditions, in order to facilitate HI sharing among users and increase their interactions by offering them an option to use anonymous online identities whenever needed.

Internet users' trust in web-based services was shown to influence their motivation to provide personal information to these services [8, 17, 45]. In the context of sharing HI on SNSs, we identified a number of trusted SNS providers, among which a "governmental health authority" appeared to be the most trusted SNS provider by the participants. Furthermore, we found that regardless of the SNS provider, participants were more likely to trust an SNS with their HI if it were recommended by their doctor(s), as compared to others (e.g., friends with mutual health conditions). We believe that SNS providers, especially those specialized in HI sharing and management, can benefit from patients' trust towards their doctors and utilize them as intermediate channels for attracting new users. This however will require incentivizing, educating, and motivating doctors, which might be a challenging process by itself.

6. CONCLUSION

We employed qualitative and quantitative instruments to investigate users' motivation to share HI on Facebook. Our results indicate that users' prior HI sharing experiences, attitudes toward privacy, and perceived health status, are linked to their motivation to share HI. In addition, we identified the key characteristics of the recipients that users preferred to share their HI with. Armed with such an understanding, we discussed the opportunities of utilizing existing features in order to optimize the gained benefits, while improving users' privacy when sharing HI. Also, our results indicate that users' health conditions could be used to facilitate HI sharing on Facebook without compromising their online privacy. Finally, by hiding *Contact and location* information, Facebook users' can maintain some level of anonymity and privacy when sharing HI with strangers.

Through this study, we (1) provide a better understanding of Facebook users' HI sharing practices, preferences, and risk perceptions, (2) identify factors linked to users' perceived privacy and motivation to share HI on Facebook, and (3) suggest design features that could facilitate effective HI sharing among Facebook users.

7. ACKNOWLEDGMENTS

We would like to thank the members of the Laboratory for Education and Research in Secure Systems Engineering (LERSSE) at UBC for their constructive feedback.

8. REFERENCES

[1] Social media "likes" healthcare: From marketing to social business. PwC HRI Social Media Consumer Survey, 2012.

[2] S. A. Adams. Revisiting the online health information reliability debate in the wake of "web 2.0": an inter-disciplinary literature and website review. *International journal of medical informatics*, 79(6):391–400, 2010.

[3] S. B. Barnes. A privacy paradox: Social networking in the united states. *First Monday*, 11(9), 2006.

[4] T. Bodenheimer, K. Lorig, H. Holman, and K. Grumbach. Patient self-management of chronic disease in primary care. *Jama*, 288(19):2469–2475, 2002.

[5] J. A. Calandrino, A. Kilzer, A. Narayanan, E. W. Felten, and V. Shmatikov. 'you might also like:' privacy risks of collaborative filtering. In *Proceedings of the 2011 IEEE Symposium on Security and Privacy*, SP '11, pages 231–246, Washington, DC, USA, 2011. IEEE Computer Society.

[6] F. Chanchary and S. Chiasson. User Perceptions of Sharing, Advertising, and Tracking. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*. USENIX Association, 2015.

[7] K. Charmaz. *Constructing Grounded Theory*. SAGE publications, 2006.

[8] V. Choudhury and E. Karahanna. The relative advantage of electronic channels: a multidimensional view. *Mis Quarterly*, pages 179–200, 2008.

[9] S. Consolvo, I. E. Smith, T. Matthews, A. LaMarca, J. Tabert, and P. Powledge. Location disclosure to social relations: why, when, & what people want to share. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 81–90. ACM, 2005.

[10] M. De Choudhury, M. R. Morris, and R. W. White. Seeking and sharing health information online: Comparing search engines and social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 1365–1376, New York, NY, USA, 2014. ACM.

[11] M. Duggan, N. B. Ellison, C. Lampe, A. Lenhart, M. Madden, L. Rainie, and S. Aaron. Social Media Update 2014. Pew Research Center, January 2015.

[12] S. Egelman and E. Peer. The myth of the average user: Improving privacy and security systems through individualization. In *Proceedings of the 2015 New Security Paradigms Workshop*, pages 16–28. ACM, 2015.

[13] N. B. Ellison, C. Steinfield, and C. Lampe. The benefits of Facebook "friends:" social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168, July 2007.

[14] S. Fox. The social life of health information. Pew Research Center Report, May 2011.

[15] S. Fox and M. Duggan. Health online 2013. Pew Research Center's Internet and American Life Project, January 2013.

[16] S. Fox and M. Duggan. The Diagnosis Difference. Pew Research Center, November 2013.

[17] D. Gefen, E. Karahanna, and D. W. Straub. Trust and TAM in Online Shopping: An Integrated Model. *MIS Q.*, 27(1):51–90, Mar. 2003.

[18] R. Gross and A. Acquisti. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society*, WPES '05, pages 71–80, New York, NY, USA, 2005. ACM.

[19] C. Hawn. Take two aspirin and tweet me in the morning: how twitter, facebook, and other social media are reshaping health care. *Health affairs*, 28(2):361–368, 2009.

[20] T. K. Houston and J. J. Allison. Users of internet health information: differences by health status. *Journal of Medical Internet Research*, 4(2), 2002.

[21] G. Iachello and J. Hong. End-user privacy in human-computer interaction. *Found. Trends Hum.-Comput. Interact.*, 1(1):1–137, 2007.

[22] N. Kamal. *Designing online social networks to motivate health behaviour change*. PhD thesis, University of British Columbia, 2013.

[23] R. Kang, S. Brown, L. Dabbish, and S. Kiesler. Privacy attitudes of mechanical turk workers and the us public. In *Symposium on Usable Privacy and Security (SOUPS)*, 2014.

[24] H. Krasnova, S. Spiekermann, K. Koroleva, and T. Hildebrand. Online social networks: why we disclose. *Journal of Information Technology*, 25(2):109–125, 2010.

[25] P. Kumaraguru and L. F. Cranor. Privacy indexes: a survey of Westin's studies. Technical report, Carnegie Mellon University, 2005.

[26] M. Kwasny, K. Caine, W. A. Rogers, and A. D. Fisk. Privacy and Technology: Folk Definitions and Perspectives. In *CHI'08 Extended Abstracts on Human Factors in Computing Systems*, pages 3291–3296. ACM, 2008.

[27] C. Lampe, N. Ellison, and C. Steinfield. A face(book) in the crowd: social searching vs. social browsing. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, CSCW '06, pages 167–170, New York, NY, USA, 2006. ACM.

[28] C. Lampe, N. B. Ellison, and C. Steinfield. Changes in use and perception of facebook. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, CSCW '08, pages 721–730, New York, NY, USA, 2008. ACM.

[29] R. Lederman, G. Wadley, J. Gleeson, S. Bendall, and M. Álvarez Jiménez. Moderated online social therapy: Designing and evaluating technology for mental health. *ACM Trans. Comput.-Hum. Interact.*, 21(1):5:1–5:26, Feb. 2014.

[30] R. C. Lefebvre and A. S. Bornkessel. Digital social networks and health. *Circulation*, 127(17):1829–1836, 2013.

[31] Y. Li. Theories in online information privacy research: A critical review and an integrated framework. *Decision Support Systems*, 54(1):471–481, 2012.

[32] Liquid Grids. *Social Media Survey Report*, May 2014.

[33] F. Lupiáñez-villanueva, W. Lusoli, M. Bacigalupo, I. Maghiros, N. Andrade, and C. Codagnone. Health-related information as personal data in Europe: Results from a representative survey in Eu27. Medicine 2.0: Ethical and legal issues, confidentiality and privacy, 2011.

[34] M. Merolli, K. Gray, and F. Martin-Sanchez. Methodological review: Health outcomes and related effects of using social media in chronic disease management: A literature review and analysis of affordances. *Journal of Biomedical Informatics*,

46(6):957–969, 2013.

[35] A. S. Moorhead, E. D. Hazlett, L. Harrison, K. J. Carroll, A. Irwin, and C. Hoving. A new dimension of health care: Systematic review of the uses, benefits, and limitations of social media for health communication. *J Med Internet Res*, 15(4):e85, Apr 2013.

[36] M. R. Morris, J. Teevan, and K. Panovich. What do people ask their social networks, and why?: a survey study of status message q&a behavior. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1739–1748. ACM, 2010.

[37] M. W. Newman, D. Lauterbach, S. A. Munson, P. Resnick, and M. E. Morris. It's not that i don't have problems, i'm just not putting them on facebook: challenges and opportunities in using online social networks for health. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 341–350. ACM, 2011.

[38] H. J. Oh, C. Lauckner, J. Boehmer, R. Fewins-Bliss, and K. Li. Facebooking for health: an examination into the solicitation and effects of health-related social support on social networking sites. *Computers in Human Behavior*, 29(5):2072–2080, 2013.

[39] M. M. Skeels. *Sharing By Design: Understanding and supporting personal health information sharing and collaboration within social networks*. PhD thesis, University of Washington, 2010.

[40] S. Torabi and K. Beznosov. Privacy Aspects of Health Related Information Sharing in Online Social Networks. In *USENIX Workshop on Health Information Technologies (HealthTech '13)*. USENIX Association, August 2013.

[41] B. Ur, P. G. Kelley, S. Komanduri, J. Lee, M. Maass, M. L. Mazurek, T. Passaro, R. Shay, T. Vidas, L. Bauer, et al. How does your password measure up? the effect of strength meters on password creation. In *Proc. USENIX Security*, 2012.

[42] M. Velden and K. El Emam. "not all my friends need to know": a qualitative study of teenage patients, privacy, and social media. *Journal of the American Medical Informatics Association*, 20(1):16–24, July 2012.

[43] N. D. Weinstein. Perceived probability, perceived severity, and health-protective behavior. *Health Psychology*, 19(1):65, 2000.

[44] A. Woodruff, V. Pihur, S. Consolvo, L. Schmidt, L. Brandimarte, and A. Acquisti. Would a privacy fundamentalist sell their DNA for $1000... if nothing bad happened as a result? the westin categories, behavioral intentions, and consequences. In *Symposium on Usable Privacy and Security (SOUPS)*, 2014.

[45] N. Xiao, R. Sharman, H. Rao, and S. Upadhyaya. Factors influencing online health information search: An empirical analysis of a national cancer-related survey. *Decision Support Systems*, 57:417–427, 2014.

[46] A. L. Young and A. Quan-Haase. Information revelation and internet privacy concerns on social network sites: a case study of facebook. In *Proceedings of the fourth international conference on Communities and technologies*, pages 265–274. ACM, 2009.

[47] D. M. Zulman, K. M. Nazi, C. L. Turvey, T. H. Wagner, S. S. Woods, and L. C. An. Patient interest in sharing personal health record information: A web-based survey. *Annals of Internal Medicine*, 155(12):805–810, 2011.

APPENDIX

A. SUPPLEMENTARY MATERIALS FOR THE EXPLORATORY STUDY (INTERVIEWS)

A.1 Interview Guide and Questions

Will collect qualitative data by means of semi-structured interviews. The interview lasted between 60-90 minutes, and were audio-recorded and transcribed. The interviews started by reviewing the consent form and the collection of demographic information (age, gender, education, etc.). Then, a number of research-related questions were asked, as shown in the following subsections. A follow-up telephone call or email communication was made when necessary to clarify issues arising from the discussion. The interview questions are presented as following:

A.1.1 Health Condition Background

- What is the health condition you have?
- How/When did the health condition appeared or started the first time?
- How does the health condition affect your daily life?
- What are the challenges that you face due to the health condition you have?
- How does the health condition you have affect your social life?
- Is there anything specific about your health condition that is of your concern?

A.1.2 Health Management

- How do you manage your health condition?
- How others (if any) are involved in your health management process?
- What is your relationship with doctors, physicians, and nurses?
- Do you have any concerns regarding your health management?

A.1.3 SNS Usage and Background

- How many SNS accounts do you maintain?
- How often do you log into your SNS accounts and what do you usually do there?
- Who do you connect to using the SNSs? Who are your online friends?
- What do the SNS environments mean to you (e.g., Facebook)?

A.1.4 HI Sharing on SNSs

- Have you ever shared HI in your SNSs? Why?
- Whom do you usually share your HI with?
- How do you think sharing HI could be helpful/beneficial to you or others in your social network?
- When it comes to sharing HI, do you have specific preferences about the type of the SNS where you share your HI in? Why?
- How do you think about existing privacy settings in SNSs?

A.1.5 A Specialized SNSs for Managing Health Conditions

- Have you ever considered using an SNS to manage, share, and seek HI?
- What do you expect from a specialized SNS that is used to help you manage your health conditions and get connected to others?
- How do you define your privacy?

A.1.6 Study Related Feedback

Do you have any comments, suggestions or concerns related to this study? We appreciate your constructive feedback?

A.2 Supplementary Results

Participants came with different health issues. Nine participants suffered from chronic pain and arthritis in different parts of their body. We interviewed two quadriplegic participants with limited physical mobility, among whom one had also suffered from chronic lung and heart diseases. We also interviewed an HIV positive patient, who was infected as a results of an accidental needle poke while doing his job as a paramedic. Finally, one participant had Neuromyelitis Optica (NMO), which is a rare disease that attacks the central nerve system and causes blindness, paralysis, and other health issues. The remaining participants suffered from a combination of mental and/or physical illnesses (e.g., eating disorder and depression, arthritis and lung disease). More details about participants' health conditions are presented in Table 4.

Table 4: Participants demographics and health conditions. The first column represents participants' ID.

ID	Gender	Age	Health condition(s)
P1	M	38	chronic sciatica due to an accident
P2	M	59	back fracture and defective left knee
P3	M	31	severe arthritis in right hand due to a car accident
P4	F	68	C4-C5 incomplete quadriplegic due to damaged neck in a sport accident
P5	F	30	chronic depression
P6	F	21	curved spine and chronic back pain
P7	M	54	C5-C6 quadriplegic due to a motor accident, and chronic heart/lung disease
P8	M	38	chronic back pain
P9	F	42	Neuromyelitis optica (NMO), episodes of blindness, headaches, and fatigue
P10	M	37	osteoarthritis (deformed leg) and defective knee
P11	M	40	L3-L4 fusion due to a work-related accident and COPD (lung problem)
P12	M	59	degenerative disk and brain injury (lost senses of balance, taste, and smell)
P13	M	51	osteoarthritis in all joints
P14	F	39	eating disorder and post-traumatic stress disorder
P15	M	37	bipolar depression and anxiety
P16	M	48	post-traumatic stress disorder
P17	M	48	arthritis in hands and knees
P18	F	38	degenerative arthritis in foot and ankle, anemia, and depression
P19	M	50	HIV due to an accidental needle poke
P20	M	50	depression and chronic pain from broken neck
P21	F	35	herniated disks (L4-L5) with chronic pain

B. SUPPLEMENTARY MATERIALS FOR THE CONFIRMATORY STUDY (ONLINE SURVEY)

B.1 Survey Items

Our survey questionnaire is presented in Appendix B.2. The survey consists of the following parts:

B.1.1 Demographics and Background

As presented in Appendix B.2, we collected general demographic information that were used to characterize different groups of participants (Q.1–Q.5). We also asked participants to identify their IT background and computer experiences (Q.6). Finally, we collected information about participants' Facebook usage (Q.7–Q.11), and asked them to describe their Facebook friends (Q.11).

B.1.2 Health Conditions and Perceived Health Status

We asked participants' to report their health conditions background (Q.12–Q.13). We also asked participants to indicate their perceived health status (Q.14). The demographic characteristics of SNS users and their health status might be highly predictive of their attitudes. For instance, younger SNS users, who did not have health problems, were assumed to have different HI sharing preferences and perceptions than older SNS users who suffered from a number of chronic health conditions.

B.1.3 Previous HI Sharing Experiences

We asked participants to indicate their HI sharing experiences with health-related SNSs (Q.15–Q.18). We also surveyed participants' previous HI sharing experiences on Facebook (Q.19–Q.20). Furthermore, we asked participants to evaluate their previous HI sharing experiences on Facebook (Q.21–Q.22). We aimed at comparing the attitudes and behaviors of participants who experienced sharing their HI on Facebook with others who did not have any experiences. It was assumed that prior experiences might affect participants' future HI sharing behaviors, especially if they had gone through good/bad experiences (e.g., gained benefits, privacy breaches, information misuse).

B.1.4 Motivation to Share HI on Facebook

Participants were asked to indicate their overall willingness to share their HI on Facebook by rating their choice on a 5-points Likert scale (Q.23). We also asked participants to identify the factors that might motivate or stop them from sharing their HI on Facebook (Q.24–Q.25).

B.1.5 Preferred Recipient(s) of the Shared HI

Participants were asked to indicate their motivation to share their HI with different user groups by rating their level of agreement on a 5-points Likert scale (Q.26). We also asked participants to indicate their willingness to use a search feature to find certain online social peers through Facebook (Q.27).

B.1.6 Anonymous Online Identity

We asked participants to consider an option for creating anonymous online identities and indicate their willingness to use it whenever sharing their HI with other Facebook users (Q.28–Q.29). Participants were also asked to identify the personal information that they were likely to hide from other online social peers if they were to create an anonymous online identity for HI sharing purposes (Q.30).

B.1.7 Trusted SNSs Provider(s)

We identified possible SNS providers and asked participants to identify their level of trust in each SNS providers (Q.31). We also ask participants to indicate their level of trust in an SNS if it was recommended to them by either a close friends/family member, friends who had medical expertise, friends who had mutual health experiences, or their doctors (Q.32).

B.1.8 Attitudes Toward Privacy

The following are the statements used in the Westin privacy index: (1) (Consumers) have lost all control over how personal information is collected and used by (companies); (2) Most (companies) handle the personal information they collect about consumers in a proper and confidential way; and (3) Existing laws and organizational practices provide a reasonable level of protection for (consumers) privacy today.

We modified the above statements by replacing the words in the parentheses with context specific words. Inspired by Westin, we asked participants to rate their level of agreement on a 4-points Likert scale for the modified statements, as shown in Q.33 (Appendix B.2). Participants who agreed (strongly or somewhat) with the first statement and disagreed (strongly or somewhat) with the second and third statements were classified as to have high privacy concerns. Participants with low privacy concerns were those who disagreed with the first statement and agreed with the second and third statements. The remaining participants were considered to have medium privacy concerns.

B.2 Survey Questionnaire

By volunteering to take part in this study, participants declare that they are at least 19 years old and that they maintain an active Facebook profile that they visit regularly. To complete the survey, participants were required to answer the following questions:

1. What is your gender?
   - Male
   - Female
   - Decline to answer
2. How old are you: [Select from list between 19 and 99]
3. What is your highest level of completed education?
   - Less than High School
   - High school (secondary school)
   - Some college/university courses
   - Diploma (post secondary courses)
   - Undergraduate University degree (Bachelor's)
   - Graduate University degree (Masters's or PhD)
   - Other (Please specify)
4. What is your employment category?
   - Administrative support (e.g., secretary, assistant)
   - Art, writing, or journalism (e.g., author, reporter)
   - Business, management, or financial (e.g., manager, accountant, banker)
   - Computer engineer or IT professional (e.g., systems administrator, programmer, IT consultant)
   - Education (e.g., teacher)
   - Engineer in other fields (e.g., civil engineer, bio-engineer)
   - Legal (e.g., lawyer, law clerk)
   - Medical (e.g., doctor, nurse, dentist)
   - Scientist (e.g., researcher, professor)
   - Service (e.g., retail clerks, server)
   - Skilled labor (e.g., electrician, plumber, carpenter)
   - Student
   - Unemployed
   - Other (Please specify)
5. What is your current country of residence? [Select from the list]
   - United States of America
   - Canada
   - Afghanistan
   - ... Additional choices hidden ...
   - Zimbabwe
   - Other
6. Do you have a college degree or work experience in computer science, software development, web development or similar computer/IT related fields?
   - Yes
   - No
   - I don't know
7. Approximately how many hours do you spend on the Internet each day? [Select between 0 and 24 hours]
8. When did you start using Facebook? [Select between 2004 and 2016]
9. How often do you check your Facebook?
   - At least once a day
   - At least once a week
   - Every month
   - Less often than every month
   - Don't use it at all
10. Please check your Facebook profile and tell us how many friends you have on Facebook?
11. How do you describe your Facebook friends? [Select all that applies]
    - Family members and relatives
    - Offline friends (e.g., childhood friends, school friends)
    - My friends' friends (online and offline)
    - Colleagues and co-workers
    - People whom I met online for the first time (e.g., people with common interests)
    - Celebrities and public figures
    - People with specific expertise/profession (e.g., lawyers, doctors, engineers)
    - Others (please specify)
12. Do you currently suffer from any chronic health conditions? [Please select all that applies]
    - Allergies
    - AIDS/ HIV
    - Asthma
    - Heart disease
    - Stroke
    - Cancer
    - Diabetes

- Arthritis and chronic pain
- Eating disorder
- Obesity
- Stress
- Depression
- Anxiety
- None
- Others (please specify)

13. How long have you had the above mentioned health conditions (if any)?
    - I don't have any chronic health conditions
    - Less than a year
    - About two years
    - About three years
    - About four years
    - More than four years
14. In general, would you say your health is:
    - Poor
    - Fair
    - Good
    - Excellent
15. Have you ever joined health-related social networking sites?
    - Yes
    - No
    - I don't know
16. Why did you join the health-related social networking sites?
17. Are you still using the health-related social networking sites?
    - Yes
    - No
    - I don't know
18. If you are not using the health-related social networking site anymore, then why did you decide to do so? [Type "NA" if you are still using the health-related social networking sites]
19. Have you ever shared details of your health information with anyone of the following people on Facebook? [Select all that applies]
    - Everyone on my Facebook friends list
    - Some close friends or family members
    - Select friends who had medical expertise and/or mutual health experiences
    - Other Facebook users (Non-friends) who had medical expertise and/or mutual health experiences
    - No one (Never shared my health information with others on Facebook)
    - Other people (Please specify)
20. Why did you share (or didn't share) your health information on Facebook?
21. How do you evaluate your prior experience with sharing your health information on Facebook?
    - Positive
    - Negative
    - Both positive and negative
    - Neither positive nor negative
    - I don't know or does not apply to me
22. What was positive and/or negative about your prior experience of sharing your health information on Facebook? [Leave blank if does not apply to you]
23. How likely would you share details of your health information with other people on Facebook? [Participants are asked to rate their response on a 5-points likert scale with responses varying from "Very Unlikely" to "Very Likely"]
24. What might motivate you to use Facebook for sharing your health information details with other people? [Please select all that applies]
    - My previous positive experiences
    - Lack of knowledge about my health issues (if any)
    - My passion to help others by sharing my health-related experiences with them
    - The need to learn from other people's expertise and experiences
    - Facebook provides me with the ability to hide my personal information and real identity from others
    - Seeking social support
    - Facebook can help me find other people with similar health issues
    - Facebook helps me to communicate with other people without having to meet them in real life
    - Nothing motivates me to share my health information on Facebook
    - Other (Please specify)
25. What might stop you from using Facebook to share your health information details with other people? [Please select all that applies]
    - My previous negative experiences
    - I don't see any benefits of sharing my health information with others
    - I am a healthy person and I do not have anything to say about my health
    - My health issues are personal and I do not want to share them with other people on Facebook
    - Others don't understand my health conditions
    - I don't have any Facebook friends that have expertise and/or experiences in the medical field
    - I don't want others to worry about my health
    - I have different people on my Facebook and I prefer not to talk about my health to all of them
    - My health condition(s) are completely manageable
    - I don't like to cry for help or feel week, my friends might misunderstand me
    - I don't feel protected online, my shared information might be misused against me
    - Other (Please specify)
26. I would consider sharing my health information details with the following Facebook users: [For each user group, participants must rate their response on a 5-points likert scale with responses varying from "Strongly disagree" to "Strongly agree"]
    - All my Facebook friends
    - Some close friends and/or family members

- Friends and/or family members who might have medical expertise and/or mutual health experiences
- Other Facebook users (Non-friends) who might have medical expertise and/or mutual health experiences

27. Facebook provides a "search" feature that can help you in finding people with specific interests, expertise, and/or experiences. Suppose that you have a chronic health condition, how likely would you use the "search" feature to find people with: [For each user group, participants must rate their response on a 5-points likert scale with responses varying from "Very Unlikely" to "Very Likely"]
    - Expertise in the medical field (e.g., Doctors, nurses, health professionals)
    - Mutual health experiences (e.g., people with similar health conditions)
28. Suppose that Facebook allows you to create an anonymous online identity. How likely would you use an anonymous online identity if you want to share your health information with other people on Facebook? [Participants are asked to rate their response on a 5-points likert scale with responses varying from "Very Unlikely" to "Very Likely"]
29. Why would you use (or not use) an anonymous online identity when sharing your health information on Facebook?
30. Suppose you want to create an anonymous identity in order to share your health information with strangers on Facebook. How likely would you "hide" each of the following personal information? [For each item, participants must rate their response on a 5-points likert scale with responses varying from "Very Unlikely" to "Very Likely"]
    - First name
    - Last name
    - Identifiable profile picture
    - Residential address
    - City where I live
    - Occupation and employment information
    - Hobbies and interests
    - Current location information (e.g., I am in "restaurant name" now)
    - Future location information (e.g., I will be in "restaurant name" at 6 PM)
    - My health condition(s)
    - Email address
    - Phone number
    - Age and date of birth
    - Gender
31. In general, I would trust a social networking site with my health information if it is operated/owned by: [For each provider, participants must rate their response on a 5-points likert scale with responses varying from "Strongly disagree" to "Strongly agree"]
    - A governmental agency (non-health related)
    - A governmental health authority (e.g., city, state/province, federal/national)
    - A recognized private company
32. In general, I would trust a social networking site with my health information if it is recommended by: [For each group, participants must rate their response on a 5-points likert scale with responses varying from "Strongly disagree" to "Strongly agree"]
    - My close friends and/or family members
    - Friends who might have medical expertise
    - Friends who might have mutual health experiences
    - My doctor(s)
33. Please rate your level of agreement with each given statement below [4-points likert scale with the given responses: "Strongly disagree", "Somewhat disagree", "Somewhat agree", and "Strongly agree"]
    - Internet users have lost all control over how personal information is collected and used by social networking sites
    - Most social networking sites handle the personal information they collect about consumers in a proper and confidential way
    - Existing laws and organizational practices provide a reasonable level of protection for internet users' privacy today
34. Have you ever performed any of the following actions on Facebook? [For each action, participants must answer with "Yes", "No", or "I don't know"]
    - Modified the privacy settings to specify the people who can see your photos, likes, comments, and other posts
    - Deleted some shared photos, comments, and/or other posts
    - Changed profile visibility (profile information that others can see)
    - Hid your friends' list from other Facebook friends
    - Modified the privacy settings to specify the people who can post on your Timeline
    - Deleted and/or blocked friends
    - Refused to provide some profile information or used fake information because it was too personal or unnecessary
    - Modified the way people can search your information on Facebook
    - Hid a specific post from others and shared it only with select friends
    - Modified the privacy settings to specify the people who can comment on and/or like your posts

B.3 Supplementary Results: Online Survey

Participants' age distribution and employment categories are presented in Figures 4 and 5. Also, Table 5 presents a list of health-related sites that were used by participants (note that these sites were not considered to be SNSs).

B.3.1 Positive and Negative Experiences

We also asked participants to explain in their words why they think their experiences were *Positive*. As presented in Table 6, a total of 272 text responses were analyzed and coded to represent participants' positive experiences. Positive emotional and social support in the form of sympathy, empathy, and prayers, were identified as the most common positive experiences among participants. Useful recommendations and advice came second

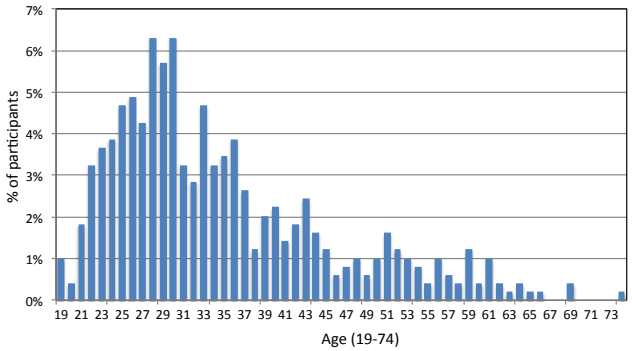


Figure 4: Participants' age distribution.

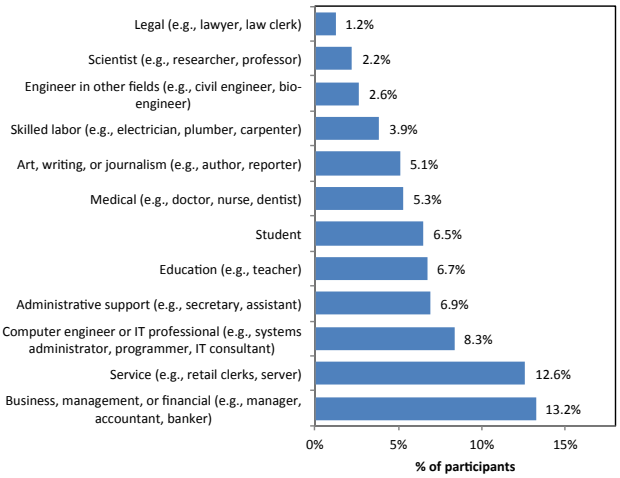


Figure 5: Participants' employment categories distribution.

in the list, with participants receiving feedback that positively helped them toward managing their health conditions. Participants also described their positive experiences by indicating that Facebook was used as an effective communication channel for broadcasting information related to their health, while receiving timely feedback from other social peers. Furthermore, participants benefitted from their conversations with others in order to bring awareness to their health issues and justify their behaviors whenever necessary. By sharing their HI on Facebook, participants were able to find other social peers who had mutual health experiences. Communicating with these social peers provided participants with valuable information/experiences while making them feel that they belong to a group of understandable and easy to communicate people. Finally, the two-way benefits of sharing HI on SNSs was easy to identify by going through participants' positive experiences in trying to help other people whenever possible.

As shown in Table 7, participants identified a number of reasons for describing their prior HI sharing experiences to be *Negative*. Participants were frustrated by the responses they received from their social peers who overreacted to their health problems and showed overwhelming and unnecessary concerns. Participants were also agitated by the social peers who used their shared HI in order to make judgments, spread rumours, gossip,

Table 5: Health-related sites used by participants that are not considered as SNSs.

Name/Description	Name/Description
Insulin Pump forum (www.insulinpumpforums.com)	PBC Group
Lymphomation.org	Hypothyroid Mom
www.community.breastcancer.org	Post traumatic stress self help group
JDRF (T1 Diabetes)	Understood.org (Kids learning)
Wrongplanet	Inspire (www.inspirehealth.ca)
Achalasia support group	Reddit communities
Weight Watchers	Healthy Brain Network
IBS Groups (ibsgroup.org)	MS Society (beta.mssociety.ca)
MS World (www.msworld.org/)	mdjunction (www.mdjunction.com/)
Mitoaction (www.mitoaction.org/)	Myelomabeacon (www.myelomabeacon.com/)
fibromyalgia of Ireland lupus and me (Facebook group)	enotalone (www.enotalone.com/)
Parenting/Breastfeeding	MedHelp (www.medhelp.org/)

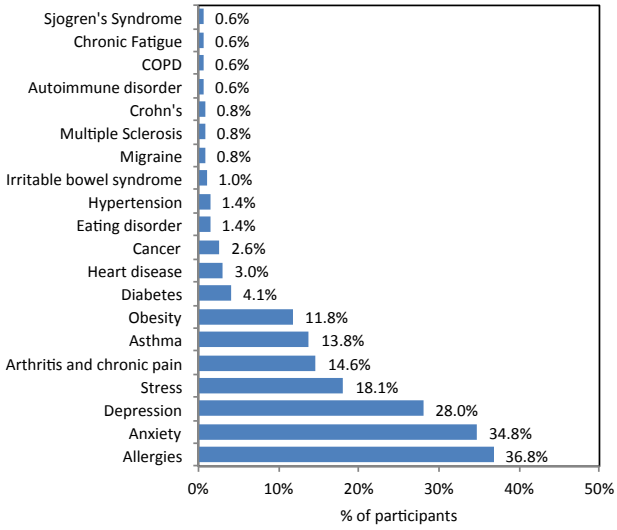


Figure 6: Reported health conditions frequencies (cumulative percentage frequency=95%).

or participated in insulting discussions. Furthermore, participants raised some privacy concerns with respect to discussing their health issues in a semi-public environment lie Facebook, which occasionally led to oversharing their health information without their permissions. Finally, while participants did not appreciate the impractical recommendations and advice given to them by some social peers, they felt lonely and unimportant when they received no support/replies from other social peers.

Table 6: Positive HI sharing experience. The first two columns represent the coded category and related sub-categories. The last two columns represent the total number/percentage of positive coded events under each category (272 total references).

Category	Sub-categories	Coded events	(%)
Positive support	sympathy, empathy, prayers, emotional and social support	107	39
Useful recommendation and advice	new medication, alternative medicine, health condition management tips, shared experiences and information resources	74	27
Communication with other peers	start conversations, quick/practical way to broadcast health information, bring attention to health conditions, receive quick feedback, justify behaviors	49	18
Mutual experiences	finding others with similar health issues, easy communication, mutual understanding, useful feedback and advice, sense of belonging	27	10
Two-way benefits	others helped me, I tried helping others	15	6

Table 7: Negative HI sharing experience. The first two columns represent the coded category and related sub-categories. The last two columns represent the total number/percentage of negative coded events under each category (86 total references).

Category	Sub-categories	Coded events	(%)
People don't understand	people overreact on health issues, feel pity, create unnecessary worry, provide responses that may increase anxiety	27	31
Negative social impact	gossips, rumours, insulting discussions and trolls, judgements, condescending responses	22	26
Privacy concerns	public/open environment, people get too involved/nosy, over sharing one's health information, receive spam/junk	17	20
Impractical advice	impractical recommendations, advice, and information	12	14
Ignored post	no replies to posts, no social support/interactions, feel lonely/unimportant	8	9

How Short is Too Short? Implications of Length and Framing on the Effectiveness of Privacy Notices

Joshua Gluck, Florian Schaub, Amy Friedman, Hana Habib,
Norman Sadeh, Lorrie Faith Cranor, Yuvraj Agarwal
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA
{jgluck, fschaub, amyfried, htq, lorrie, ns1i, yuvraja}@andrew.cmu.edu

ABSTRACT

Privacy policies are often too long and difficult to understand, and are therefore ignored by users. Shorter privacy notices with clearer wording may increase users' privacy awareness, particularly for emerging mobile and wearable devices with small screens. In this paper, we examine the potential of (1) shortening privacy notices, by removing privacy practices that a large majority of users are already aware of, and (2) highlighting the implications of described privacy practices with positive or negative framing. We conducted three online user studies focused on privacy notice design for fitness wearables. Our results indicate that short-form privacy notices can inform users about privacy practices. However, we found no effect from including positive or negative framing in our notices. Finally, we found that removing expected privacy practices from notices sometimes led to less awareness of those practices, without improving awareness of the practices that remained in the shorter notices. Given that shorter notices are typically expected to be more effective, we find the lack of increased awareness of the practices remaining in the notice surprising. Our results suggest that the length of an effective privacy notice may be bounded. We provide an analysis of factors influencing our participants' awareness of privacy practices and discuss the implications of our findings on the design of privacy notices.

1. INTRODUCTION

The purpose of a privacy policy is to make users aware of a system's or company's practices related to collection, sharing, use, and storage of personal information. In theory, a company's privacy policy contains all the information that users need to be aware of a company's privacy practices and to make informed decisions about which companies to entrust with their personal information. In practice, privacy policies are too long, leading to user fatigue and users ignoring privacy policies [12, 33, 40]. Recognizing this problem, the Federal Trade Commission (FTC) has called for clearer and shorter privacy notices [16].

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22–24, 2016, Denver, Colorado.

Prior research has examined short-form privacy notices, which are condensed versions of privacy policies that include the main practices, but may remove some degree of nuance or detail. Research studies have found that standardized short-form privacy notices can increase user awareness of privacy practices [15, 28, 29]. Other research and reports have suggested that focusing privacy notices on unexpected practices may increase awareness and effective transparency, reducing the potential for user surprise, and reducing the burden on users [6, 17, 41]. Prior work has also shown that presenting information with a positive or negative framing can also change users' perceptions and awareness of privacy practices [1, 2, 3, 22]. Our research builds upon prior work, examining three important questions.

Our first research question is whether removing from notices those privacy practices that most participants already expect to occur, would lead to greater overall awareness of an organization's privacy practices. We hypothesize participants will have higher awareness of privacy practices remaining in notices, since the notices will be shorter and more focused. In addition, participants should have similar awareness of practices that were removed, as these would be practices most participants would already expect without a notice.

Our second research question examines the effect of notice framing on user awareness about privacy practices. We compare positively and negatively framed notices against a neutral baseline.

Our third research question examines the effectiveness of short-form privacy notices in the context of fitness wearables. The effectiveness of short-form notices on increasing user awareness has been shown in several contexts [29, 31]. However, while the fitness wearable companies we surveyed (Fitbit, Misfit, Jawbone) have made some attempt to use clear language in their privacy policies, none utilized short-form privacy notices at the time of our study [19, 26, 36]. Fitbit had a plain-language illustrated version, but it was still fairly long when fully expanded. We picked fitness wearables for this study given their increasing popularity [25] and the fact that they typically collect a number of privacy-sensitive data items for their functionality (e.g., detailed physical activity of the user) leading to security and privacy concerns [24].

We conducted three online user studies to analyze notice design format, participants' baseline knowledge, and notice length and framing for the Fitbit Surge watch (shown in Figure 1). We conducted the design format study to compare the effectiveness of four candidate short-form notice designs. The baseline knowledge study served to determine which privacy practices a large majority of users would already be aware of. The notice framing and length

Figure 1: A Fitbit Surge Watch, which we used as a representative Fitness wearable.

study was a 3 (lengths) x 3 (framing) study, with a control condition, to answer the research questions outlined above. All studies were approved by Carnegie Mellon University's Institutional Review Board.

The results from our design format study showed our four short-format notice designs resulted in similar awareness of privacy practices, so we chose a format loosely based on the format of Fitbit's existing online privacy policy. The results from our second study showed a wide range of awareness rates about Fitbit's individual privacy practices and allowed us to identify 6 practices expected by at least 85% participants to remove from the medium and short version of the policy, and an additional 7 practices expected by at least 70% of participants to also remove from the short version of the policy.

Our final study, examining the effects of short-form notice length and framing on privacy awareness, provided a number of interesting results. We found that participants in the medium short-form notice conditions were similarly aware of privacy practices as those in the long short-form notice conditions. Removing expected practices from the medium notices did not impact awareness significantly of either the removed or remaining practices. However, participants in the shortest short-form notice conditions were less aware of the practices removed only from the shortest notices, with no significant change in awareness of the practices also removed from the medium notice or those that remained. We also found no significant difference in awareness from positive or negative framing in the notices. While not finding an effect does not prove that such an effect does not exist, it does suggest that the effect, at least in this context, is likely to be small. We discuss the implications of our results at the end of this paper.

2. RELATED WORK

Here we discuss prior work on privacy notice design in three areas: short-form privacy notices, framing, and delivery methods.

2.1 Short-form Privacy Notices

It is fairly rare for individuals to read a privacy policy in its entirety. Prior work has shown two key reasons for this: the complexity of privacy policies, and their length. Privacy policies are generally written in complex legalese or are purposefully vague, making it hard for readers to understand them [12, 27]. In fact, research has shown that not only do users struggle to make sense of privacy policies, but that even experts can disagree on the meaning of certain statements [42]. In addition, prior work has suggested that an individual would have to spend 244 hours each year to read the privacy policies of websites they visit [33]. As a result, the FTC and others have called for privacy notices to be made both clearer and shorter, in order to increase comprehension [16, 17].

Prior work has shown that short-form notices summarizing the key privacy practices of an organization can provide significant benefits to user awareness over a traditional privacy policy [28, 29]. However, including all of the relevant information in a privacy notice, even in a compact form, may still result in overly long notices, and leaving out unexpected privacy practices can hide information and impair transparency [34].

Others have suggested that focusing on unexpected practices is important for user understanding. A recent FTC staff report suggested that when "data uses are generally consistent with consumers' reasonable expectations, the cost to consumers and business of providing notice and choice likely outweighs the benefits" [17]. Rao et al. studied mismatches between user privacy expectations and practices disclosed in privacy policies. They found that mismatches (e.g. unexpected practices) comprise a relatively small set of practices described in privacy policies, and that creating privacy notices focusing on these practices could reduce user burden [41]. Ayres and Schwartz proposed warning labels to highlight unexpected terms in contracts [6]. Ben-Sahar and Chilton found that a warning label focusing on unexpected privacy practices benefited user comprehension, although they did not find any behavioral change associated with this increase in user comprehension [9].

Layered notices, short notices that link to a full policy containing more information, may allow for the benefits of a short-form notice, as well as avoiding the appearance of hiding unexpected practices [13, 35, 37]. However, users may consent to the first layer of the notice they encounter, without delving into the following layers [34, 43].

Other work has examined the potential of using machine learning and natural language processing to extract answers to specific questions from privacy policies and display it using a web browser plugin [47, 49]. Similarly, browser plugins have been developed to display summaries of computer-readable privacy policies [14].

We seek to reach a compromise between length and inclusion of relevant information in a short-form privacy notice. Our approach is to determine the privacy practices that are unexpected by most participants, and ensure that those are included in even the shortest privacy notice, while removing practices that users generally expect. We hypothesize that doing so will provide the benefits of a shorter notice without the downsides of leaving out unexpected privacy practices or relegating them to a secondary layer.

2.2 Framing

In addition to the content of a privacy notice, the way in which privacy practices are explained can also have a major effect on users' perception and retention of those practices. Perception of the relative importance, or sensitivity, of certain types of information can strongly affect a users' willingness to share it. Prior work has shown that providing reasons for privacy practices [44, 45], or communicating risks and implications [20], can grab users' attention, change their level of concern over practices, and cause them to reflect on privacy practices more deeply. Research has shown that including personal examples, such as the number of data accesses associated with mobile permissions, can lead to even greater concern, and therefore reflection [5, 7, 22].

Framing can also ease users' concerns over privacy. Studies have found that framing notices with more positive, misleading, or misdirecting statements can direct users' attention away from the implications of privacy practices, and thus decrease their awareness of these practices [1, 2, 3].

2.3 Delivery Methods

There has been substantial prior work examining the way in which privacy notices are delivered, including the timing [8], channel, and

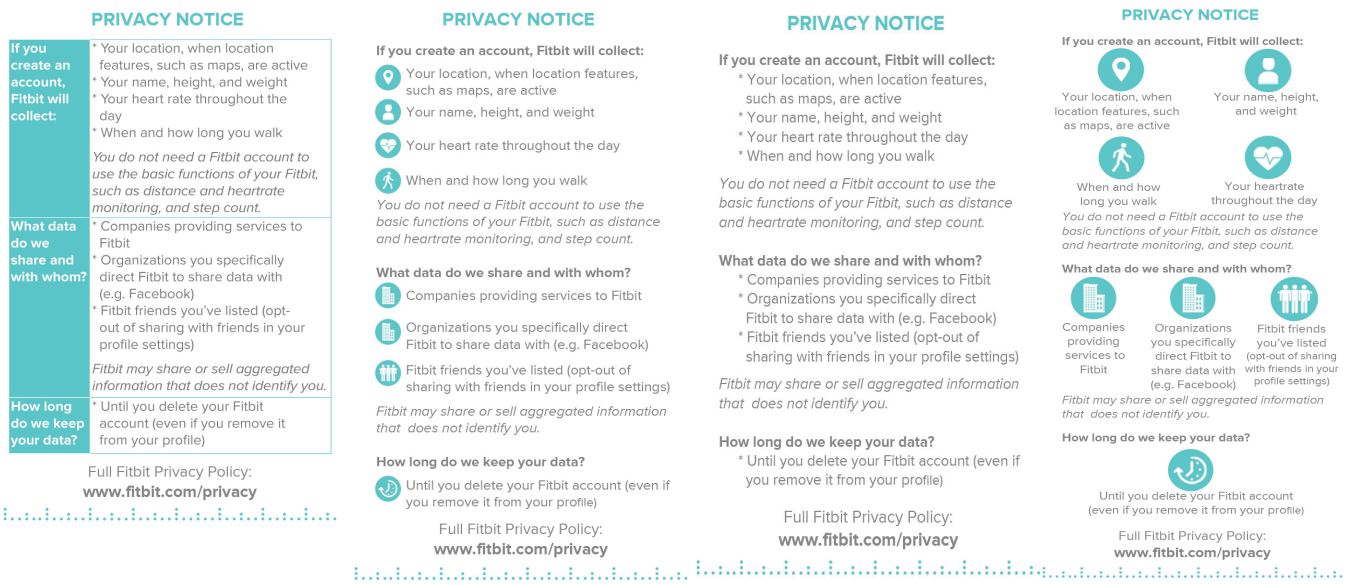


Figure 2: Privacy notice design formats tested in the first survey (*left to right*): Table format, bulleted icon format, bulleted format, and icon format. The privacy notices all show the same practices based on Fitbit's privacy policy. The bulleted icon format was used in the second and third study.

modality of privacy notices [43]. Rather than displaying a single privacy notice when a device is first purchased or activated, prior work has examined the potential for showing privacy notices at regular frequencies, or in the form of 'just-in-time' notices that are sent just before a privacy sensitive activity is about to occur [4, 5, 7, 38, 39]. Other research has focused on making privacy notices integral to the function of the device, for example playing sounds when a photo is taken or data is sent to the cloud [10].

Finally, there has been significant research into formats for privacy policies and other notices [34]. Research on standardization of privacy policies [15, 18, 31], and privacy 'nutrition labels' [28, 29] has found that standardized tabular formats are beneficial. Good et al. found that users were more likely to notice short versions of end user license agreements (EULAs), but the notice format did not impact installation rates significantly [21]. Waddel et al. found that paraphrasing EULA content and splitting it into multiple pages increased comprehension [46]. However, it is not clear whether the change can be attributed to the paraphrasing or the multiple pages. In our studies, we isolate specific aspects to reduce confounding factors in order to gain deeper insights into notice effectiveness.

3. PRIVACY NOTICE DEVELOPMENT

We focused our research on the Fibit Surge watch due to Fitbit's leading market share in fitness wearables (22%) [25]. The Surge was the newest Fitbit device at the time we began our study. The content of the privacy notices we developed and tested are based on an analysis of Fitbit's privacy policy from Dec. 9, 2014 [19], which was still Fitbit's current privacy policy at the time of this writing. We included Fitbit's collection, sharing, selling, and storage practices in our privacy notice designs. We did not include any practices relating to online tracking for individuals who visit Fitbit's website, as these practices did not relate directly to the Fitbit device. Note that while our research was focused on a single fitness wearable's privacy policy, we examined the privacy policies of other fitness wearable vendors (namely, Jawbone [26] and Misfit [36]) and found them to describe similar practices.

In the following sections, we describe our privacy notice development process. Our first step was to determine an effective privacy notice design format for the Fitbit device. Our second step was to determine which practices participants expected, even without a privacy notice. This informed our decisions about which practices to remove from the shorter versions of our notices in order to emphasize unexpected privacy practices.

3.1 Short-form Notice Design

We created four prototype short-form privacy notice designs, and conducted a survey to assess the effect of design on awareness of Fitbit's privacy practices. The designs are shown in Figure 2: table format, bulleted icon format, bulleted format, and icon format. Table formats have been used successfully in standardizing bank privacy policies [18, 31] and in privacy nutrition labels [29]. Fitbit's illustrated privacy notice uses icons with text and Fitbit's full legal privacy policy includes bulleted text [19]. While our four formats had different layouts and graphical elements, they all contained the same text. We designed our first study to test which of these formats led to the greatest awareness of Fitbit privacy practices.

3.1.1 Study Design

In summer 2015 we conducted a 200-participant survey on Amazon Mechanical Turk, using a between-subjects design with 50 participants per format. We chose 200 participants after conducting a power analysis using Cohen's medium effect size to ensure that we achieved 80+% power, even with study drop outs. Participants were paid $0.60 for completing the survey. Only US Turkers with 95% or higher HIT acceptance were recruited. To reduce bias, the survey was marketed as a survey on fitness wearables: no recruitment information indicated the survey was related to privacy.

After being asked a set of demographic questions, participants were shown one of the four short-form privacy notice designs and instructed to read it carefully as they may be asked questions about it. The goal was to create a best-case scenario in which all participants would pay attention to the notice, so that we could assess differences in awareness based on notice design, rather than due to

Question	Correct(%)	Incorrect (%)	Unsure (%)	In Short Notice	In Medium Notice
Collect					
Steps	94	3	3		
Distance	94	4	1		
Info Posted to Profile	93	6	1		
When Exercising	93	6	1		
Heartrate	93	6	1		
Stairs Climbed	88	11	1		
Name	81	16	3		*
Sleep	76	20	4		*
Exercise Comp. to Friend	73	22	5		*
Weight	72	24	4		*
Height	70	25	5		*
Location Specific (Q. 20 in Appendix)	31	56	13	*	*
Share With					
Fitbit Friends (Q. 16 in Appendix)	76	20	4		*
Companies Providing Services	72	22	6		*
Directed Organizations (e.g. Facebook)	67	26	7	*	*
Government	29	66	5	*	*
Misc.					
Where to Find Privacy Policy	88	12	0	*	*
Use Fitbit Without an Account	31	53	16	*	*
Selling Data Conditions	23	57	20	*	*
Data Retention Policy	22	47	31	*	*

Table 1: Results from our second MTurk study (70 participants). Shows the % correct/incorrect/unsure for Fitbit privacy practices without any form of privacy notice. Using Fitbit without an account denotes the functionality a Fitbit maintains without a connection to a Fitbit account (and thus without any form of data collection). We use asterisks to indicate which practices we displayed in our short and medium notices; all practices were displayed in our long notice.

different levels of attention. Participants could move on to the next survey page as soon as they wanted but were not able to return to the notice after that. They were then asked questions to test their awareness of the Fitbit privacy practices. After answering these questions, participants were again shown the assigned privacy notice format, and asked to rate its helpfulness on a 5-point Likert scale (not very helpful to very helpful), and to evaluate how comfortable they were with Fitbit's collection of location data, storage practices, and sharing practices on a 7-point Likert scale (very uncomfortable to very comfortable). We asked these questions to get a sense of a participant's feelings towards the privacy notices, as well as their feelings towards some of Fitbit's privacy practices.

3.1.2 Study Results and Conclusions

We found no statistically significant differences between formats in awareness of Fitbit's privacy practices. Additionally, using Kruskal-Wallis tests, we found no difference between privacy notice format in terms of how helpful participants found notices ($H(3,197)=.3326$ $p=.95$), or how they felt about collection of location data ($H(3,197)=.7017$ $p=.87$), storage practices ($H(3,197)=.0816$ $p=.99$), or sharing practices ($H(3,197)=.4961$ $p=.51$). In past studies that have found differences in the performance of privacy policy format variants [29, 34], the tested formats varied in wording, length, and layout, while our formats varied only in layout.

We selected the bulleted icon format (second from the left in Figure 2) for our final study because it was in line with Fitbit's general design motif of mixing icons and text [19].

3.2 Baseline Knowledge of Privacy Practices

One of our key hypotheses was that removing commonly expected pieces of information from a privacy notice would increase awareness of the information contained in the privacy notice, since there would be less information for people to read and understand. We conducted a study to identify which privacy practices described in the Fitbit privacy policy were commonly expected.

3.2.1 Study Design

We designed a survey asking participants questions about Fitbit's privacy practices without showing them any privacy notice. In addition, we let participants know at the beginning of the survey that they would not be penalized for wrong answers, so as to discourage them from searching for this information in Fitbit's privacy policy.

We recruited 70 Turkers from the US with 95% or higher HIT acceptance during Fall 2015. The survey was marketed as a survey on fitness wearables, with no recruitment information indicating the survey was related to privacy. Participants were paid $0.60 for completing the survey. After answering basic demographic questions, participants were directed to visit the Fitbit Surge page on Fitbit's website and could not move on from this page for 2.5 minutes. We included this provision because a potential buyer of a Fitbit device would likely spend some time looking at its webpage before purchasing the device. However, we did not enforce that participants look at the Fitbit Surge page, only that they wait 2.5 minutes before advancing in the survey.

Participants were then asked questions about 30 collection, sharing, selling, and data retention practices, specifically pertaining to the Fitbit Surge watch. These questions included 20 practices actually included in Fitbit's privacy policy (shown in first column of Table 1), as well as questions regarding ten additional fictitious practices. Examples of fictitious practices include collecting perspiration, altitude, and mood; and sharing with researchers, Facebook friends, and the public. We included fictitious practices in order to ensure that participants did not believe that all practices mentioned were performed by Fitbit. Participants were then asked

a series of multiple choice questions related to Fitbit policy details. Because we were interested in baseline knowledge of actual privacy practices, we report only these results (see columns 2-4 of Table 1).

3.2.2 *Study Results and Conclusions*

As shown in Table 1, there was a wide range of participant awareness. 94% of participants knew that the Fitbit Surge collected steps, whereas only 22% were aware of Fitbit's data retention policy. Many of these questions were based on a likert scale, as can be seen in questions 12 and 14 in Appendix. For our results, we aggregated any choice (from might to definitely) to a binary collect/did not collect. Overall, participants were more knowledgeable about data collection practices, somewhat less knowledgeable about sharing practices, and least knowledgeable about specific policies such as data retention or using the Fitbit Surge without a Fitbit account.

We used our results to inform our decisions about what practices to omit in our shorter notices. We wanted to remove practices only when a strong majority could answer questions relating to those practices correctly. We determined that removing items that 70% or more and 85% or more of participants were able to answer correctly allowed for the removal of two clear clusters of information. The data practices that were retained in the medium- and short-length notices are shown in the right two columns of Table 1.

4. NOTICE FRAMING AND LENGTH

Our two preliminary studies informed the design of the short-form privacy notices that we used to test our hypotheses relating to effects of the framing and length of the notice. We considered three forms of framing (positive, negative, neutral), and three notice lengths (short, medium, and long). This led to a 3x3 experimental design, with a tenth condition as control.

In the positive framing conditions we included positive reasons for Fitbit to engage in some of its practices, namely sharing and data retention. In the negative framing conditions we included potential drawbacks/risks related to the same practices. Figure 3 provides the positive and negative framing text. The neutral condition did not include any framing. What practices were included in which notice length can be seen in Table 1. Figures 4, 5, and 6 show the long, medium and short length notices. The figures show examples from different framing conditions. All use the bulleted with icons design from our design format study.

In addition to notice content, for all notice lengths we included at the end of the first two sections of the notice the phrase "Find further [collection/sharing] practices at Figbit.com/privacy." At the bottom of the policy we included the text "Full Fitbit Privacy Policy www.fitbit.com/privacy." We did this to avoid the perception that the absence of well-known practices from the shorter notices indicates that these practices do not occur.

In January 2016 we recruited 400 Turkers from the US with 95% or higher HIT acceptance, approximately 40 per condition in a between-subjects study design. We chose 400 participants as a result of a power analysis using Cohen's medium effect size to ensure that we achieved 95+% power, even with study drop outs. Due to randomized condition assignment and some participants failing to complete the survey after being assigned to a condition, actual conditions ranged in size from 33 to 42 participants (Mean=38.7 SD=3.71). The survey was marketed as a survey on fitness wearables, with no recruitment information indicating the purpose was related to privacy. Additionally, we noted within the survey that participants would not be penalized for incorrect answers, as we were more interested in their opinions and knowledge level than achieving the best answers. This was done to reduce the likelihood of Turkers looking up answers in the survey.

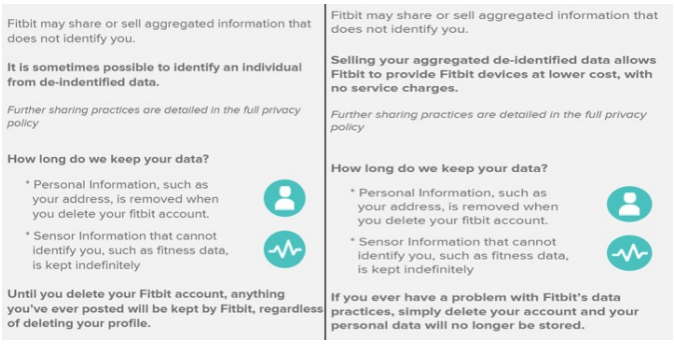


Figure 3: Negative (*left*) and positive (*right*) framing statements (in bold) for data sharing and retention practices.

This survey followed the same study design as our baseline knowledge survey until after participants were directed to view the Fitbit Surge's webpage (survey can be found in Appendix). At that point, participants were shown a notice (or shown nothing in the control), based on their condition. In order to make participants' interaction with the notices realistic, participants were allowed to skip to the next page of the survey without spending any time looking at the notice. We recorded the time participants spent on the notice page. We then presented questions relating to Fitbit privacy practices as we had in the baseline knowledge survey. Participants were not able to return to the notice while answering these questions.

In addition to the questions relating to Fitbit's privacy practices, we also asked participants to rate their general concern with Fitbit's privacy practices, as well as to answer the 10-item variant of the IUIPC privacy concerns scale [32]. We did this to measure what we expected to be the mechanism (concern) by which our framing conditions affected participant awareness of Fitbit privacy practices. To account for this longer survey length, we compensated the participants $1.50.

5. RESULTS

In the following sections, we aggregate our results by type of practice and overall awareness. For the purposes of aggregation, we count a participant's correct answer about each of 20 data practices as 1, an incorrect answer as -1, and unsure answers as 0. The question categories are shown in Table 1 and the questions are shown in the Appendix. Our metric led to a non-normal distribution of awareness for certain conditions. As a result, we performed non-parametric statistical tests (Kruskal-Wallis).

Our short-form notices increased awareness of privacy practices over the control condition (no notice). However, framing did not have a statistically significant effect on privacy practice awareness or concerns. Additionally, the shortest notices performed worse in terms of privacy practice awareness than the medium and long notices, particularly on practices removed from the short notices. Age and Gender were related to awareness, but there was no interaction effect between these factors and condition. Participants who visited the Fitbit website during the survey had significantly higher awareness scores than those who did not, and those in the control condition benefited most from visiting the website. Additionally, we found no significant difference in time spent reading notices between conditions. However, we found that longer reading times, concern about Fitbit privacy practices, and high IUIPC scores were associated with greater awareness of privacy practices. We discuss the results in detail below.

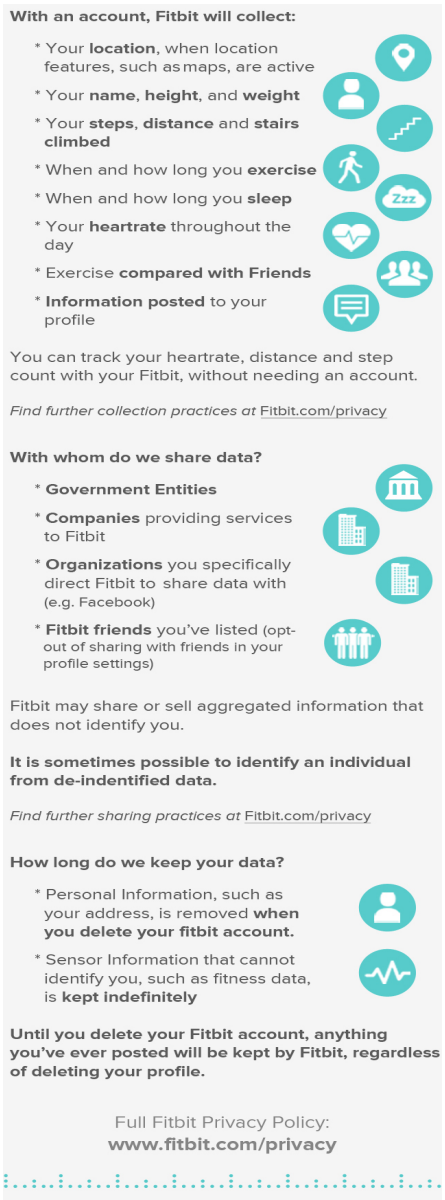


Figure 4: Long length notice (negative framing): Includes all Fitbit privacy practices relevant for using a Fitbit Surge, as well as negative framing statements for certain practices.

5.1 Participants

We initially recruited 400 participants through Amazon Mechanical Turk. Nine participants were removed when our survey tool (SurveyGizmo) indicated they were connecting from outside the US, despite being identified as US MTurkers.

Our sample was fairly diverse. The median age was 29, with a range of 18-69. 193 (49.4%) of our participants were male, 196 (50.1%) female, with two participants not reporting their gender. As shown in Table 2, most of our participants reported currently or previously using a fitness wearable device.

5.2 Effectiveness of Notices

Our short-form privacy notices led to increased participant awareness of privacy practices. Performing a Mann-Whitney U test, we found participants who saw one of our short-form privacy notices had significantly higher overall privacy practice awareness (Mean=12.06, SD= 5.89) than control participants (M=9.54, SD= 5.86), with (U(1,390)=-3.03, p=.002, r=.153).

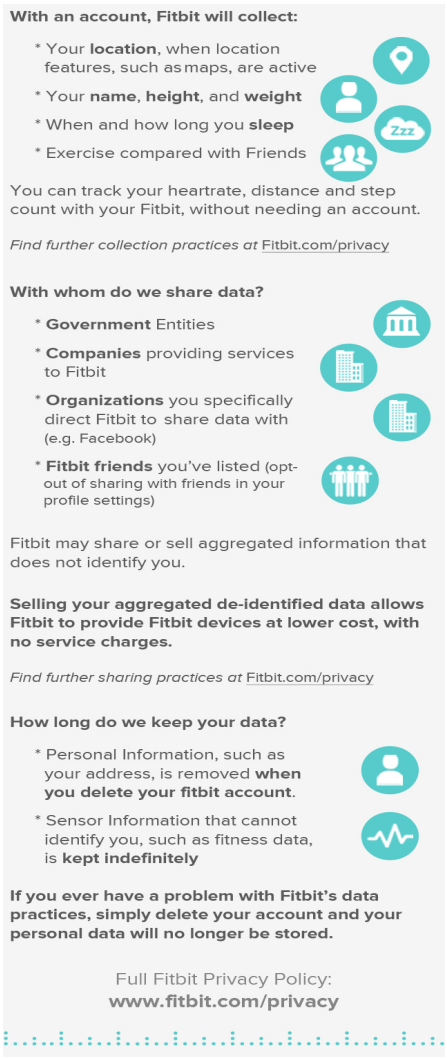


Figure 5: Medium length notice (positive framing): Has had relevant Fitbit privacy practices which 85% or more individuals assume are true removed.

We examined whether our hypotheses relating to framing and length of the notice led to significant changes in awareness. Performing a Kruskal-Wallis test, we found there was no statistically significant interaction between the framing and length conditions (H(8, 343)=14.26, p=0.08) on overall privacy practice awareness. Therefore, when conducting further analysis on each of these variables individually, we aggregate conditions by their framing or length.

5.2.1 *Framing*

Our positive and negative framing statements (shown in Figure 3) had no noticeable effect on participants' awareness of Fitbit's privacy practices. Performing a Kruskal-Wallis test, we found no significant differences in overall privacy practice awareness based on the framing of the notice (H(2,349)=2.643, p=.267).

5.2.2 *Length of Notice*

We found that our shortest notice resulted in lower privacy practice awareness than longer notices, and that this was particularly

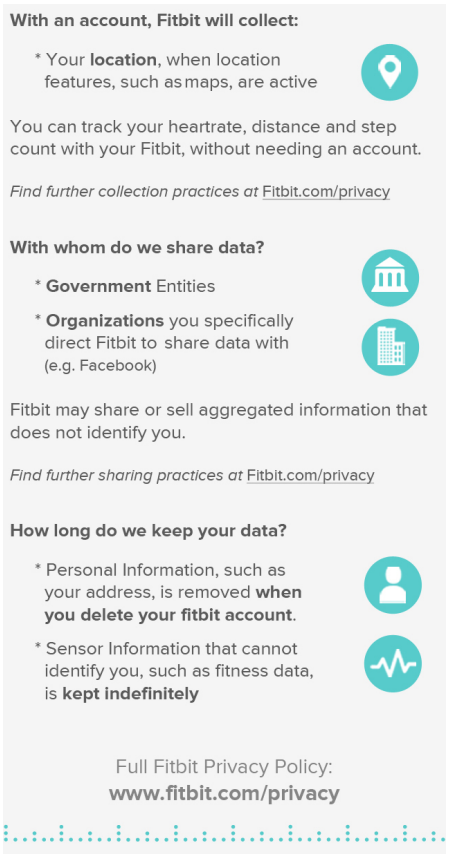


Figure 6: Short length notice (no framing): Has had relevant Fitbit privacy practices which 70% or more individuals assume are true removed. No framing statements included.

Category	Percent
I currently use a wearable Fitness device	30.2
In the past, I regularly used a wearable fitness device, but I no long do so	10.5
I have tried out a wearable fitness device, but have never regularly used one	17.1
I have never a wearable fitness device, but am familiar with the concept	40.2
I was unfamiliar with wearable fitness devices, before taking this survey	2.0

Table 2: Participant experience with fitness wearables.

true in the case of practices removed from the shorter notices (see 7). Demonstrating this, we ran a Kruskal-Wallis test and found significant differences in awareness of privacy practices ($H(2,349)$ =10.42, p=.005) based on length.

Performing post-hoc Mann-Whitney U tests with Tukey correction, we found that long notices (Mean =12.52, SD =5.98) and medium length notices (Mean=12.65, SD=5.14) outperformed short notices (Mean=11.05, SD=5.82) in terms of overall awareness of privacy practices and collection practices with ($U(1,232)$ =-2.909, p =.012, r =.191)and ($U(1,238)$ =-2.604, p =.027, r =.168), respectively. We found no significant difference between long and medium length notices in terms of overall awareness of privacy practices.

While important in aggregate, we also examined whether the

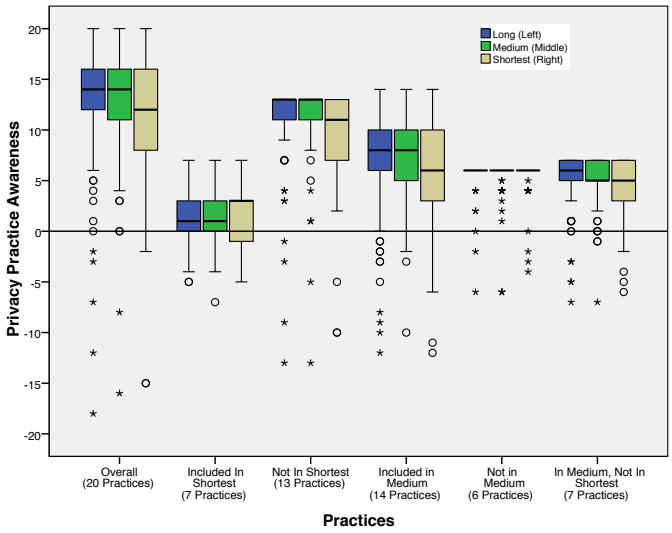


Figure 7: Privacy practice awareness by Length. Strong similarity in performance between long and medium length notices. Significantly worse performance for the shortest notice on privacy practices overall, and specifically on practices removed from the shortest notice. Medium length notices performed similarly to long length notices for practices both left in and removed from the medium length notice.

change in awareness between notice lengths was focused on practices that remained in the shortest version of the notices, or practices that were removed from the shortest version of the notice. We originally postulated that participants in shorter length conditions would perform less well on practices removed from their notices, and potentially better on practices that remained in their notices.

Performing a Kruskal-Wallis test, we found significant differences in awareness by length when considering practices that had been removed from the shortest notices ($H(2,349)$ =22.439, p <.00 05). Performing post-hoc Mann-Whitney U tests with Tukey correction, we found long and medium length notices (Long: Mean =11.05, SD=4.14; Medium: Mean=11.27, SD=3.66) outperformed short length notices (Mean=9.50, SD=4.36) in terms of practices removed from the shortest notice, with ($U(1,232)$=-3.891, p<.0015, r=.255) and ($U(1,238)$=-4.127, p<.0015,r=.267) respectively. We found no significant differences between long and medium length notices.

Additionally performing a Kruskal-Wallis test, we found no significant difference in awareness of practices remaining in the shortest notice by length.

While we found no difference in the performance of long and medium length notices overall, we also analyzed whether there was a difference in performance when considering practices left in and removed from the medium notices independently. Performing a pair of Kruskal-Wallis test, we found a significant difference in awareness of practices remaining in the medium notice (with $H(2,349)$=10.126, p=.005, and not significant difference in awareness of practices removed from the medium notice. Performing post-hoc Mann-Whitney U tests with Tukey correction, we found no significant difference between long and medium notices in awareness of practices remaining in the medium length notice (p=.882). Instead, we found that both the medium and long notices outperformed the shortest notice when considering practices remaining in the medium length notice, with (Long vs. Short: $U(1,232)$ =-2.726 p=.018, r=.181) and (Medium vs. Short: $U(1,238)$ =-2.756 p=.018 r=.178).

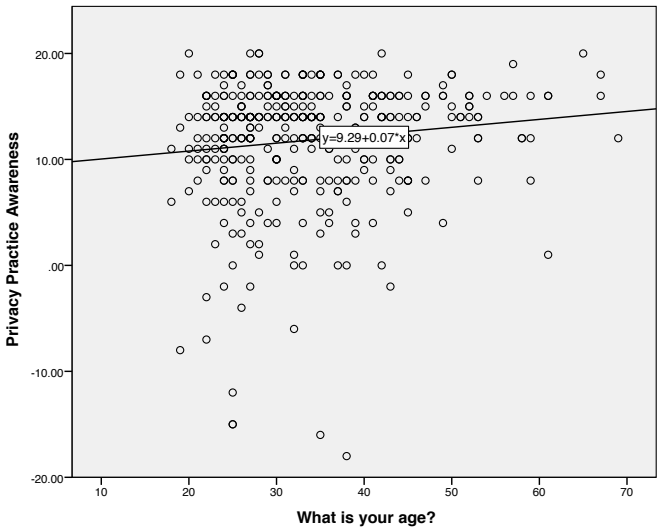


Figure 8: A statistically significant relationship between age and privacy awareness, with each year of age being associated with a .07 increase in awareness score.

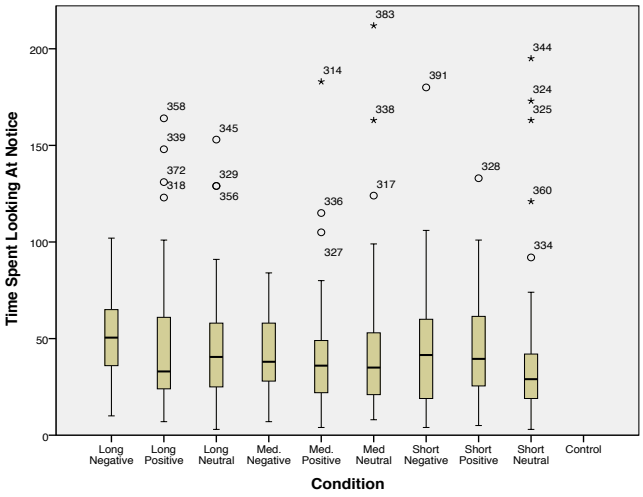


Figure 9: Time spent on notice by condition. No significant difference in time spent on notice by condition, with an average between 45–55 seconds for each condition.

These results prompted us to examine the performance of the various notice lengths on awareness of those 7 practices that were removed from the shortest notice, but were retained in the medium and long notices. Note that these practices were expected by between 70 and 85% of participants in our baseline knowledge study, while the other 6 removed practices were expected by over 85% of participants in that study. Performing a Kruskal-Wallis test, we found that there was a significant difference in awareness of these practices by notice length, with $H(2,349)=14.268$ p=.001. Performing post-hoc Mann-Whitney U tests with Tukey correction, we found significant differences in awareness of these practices between both the long and shortest notice lengths, and the medium and shortest notice lengths with (Long vs. Short: $U(1,233)=-3.037$ p=.006 r=.199) and (Medium vs. Short=$U(1,238)=-3.435$ p=.003 r=.222). Indeed the participants in the shortest notice conditions performed similarly to those in the control condition on these 7 practices, while participants in the medium and long conditions became more aware of these practices. This suggests that 70 to 85% awareness of practices may not be high enough for successful removal from a privacy notice.

5.3 Impact of Demographic Factors

Interestingly, age and gender both had significant effects on participants' overall privacy practice awareness, although we did not find any interaction between these factors and participant condition. This means that our conclusions regarding our notice conditions are generally applicable across these demographic factors.

We performed a linear regression, and found that for each year of age, participants had a .075 higher awareness score (see Figure 8), with t=2.558, p=.011. Performing a Mann-Whitney U test, we found that women (Mean= 11.13, SD=6.31) had higher overall privacy practice awareness than men (Mean=10.50, SD=7.77), with ($H(1,387)=-2.104$, p=.035, r=.109). We removed two participants who chose to not share their gender from this analysis.

5.4 Impact of Participant Behavior

We examined the relationship between participant behavior in our survey and privacy practice awareness in two ways. First, as mentioned in the methodology, we indicated to participants that they should visit the Fitbit Surge page on the Fitbit website as if they were shopping for a Fitbit. However, we did not force our participants to do so. While we did not record whether participants visited the website, we asked participants how much the Fitbit Surge costs ($250). This acted as a knowledge check to determine who had at the very least visited the page, as the cost was prominently displayed at the top right corner of the page. We did this to get a measure of participants' commitment to researching the device, and the extent to which a privacy notice would help those more or less likely to examine a fitness wearable's details on their own.

Additionally, we tracked how long participants spent on the page of the survey that showed them our privacy notice before moving on. We hypothesized that participants could spend less time on our shorter notices while maintaining at least similar performance.

5.4.1 Knowledge Check

Our analysis showed a strong majority, 339 (86.7%) participants, knew the cost of the Fitbit Surge, as compared to 52 who didn't know(13.3%). We performed a Mann-Whitney U test showing that participants who knew the cost of the Fitbit Surge had significantly higher overall privacy practice awareness (Mean=11.70, SD=5.47) than participants who did not (Mean=8.25, SD=6.85) with ($U(1,390)=-3.719$, p=<.0005,r=.188).

Examining the data more closely, we found that there was a major jump in overall privacy practice awareness for participants in our control condition, from (Mean=1.50, SD=8.22) to (Mean=10.46, SD=4.90) for those who passed the knowledge check, whereas the increase in awareness for those who passed this knowledge check in the treatment conditions (with notices) was not as dramatic going from (Mean=9.58, SD=6.41) to (Mean=12.46, SD=5.47). This may be due to the fact that the Surge page contained information about its functionality, which included mention of the data it collects. Participants in the control condition were not presented with information about data collection except on this page, whereas participants in the other conditions received this information both on the Surge page and in the privacy notice.

5.4.2 Time Spent on Notice

We found a number of interesting results regarding time spent looking at our notices. We found no significant differences between time spent reading the notices in each condition. However,

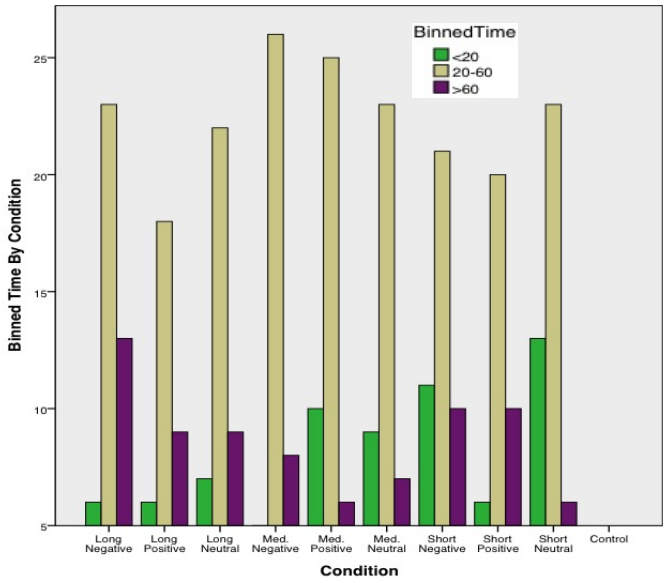


Figure 10: Binned time spent on notice by condition. We found no significant difference in binned time by condition.

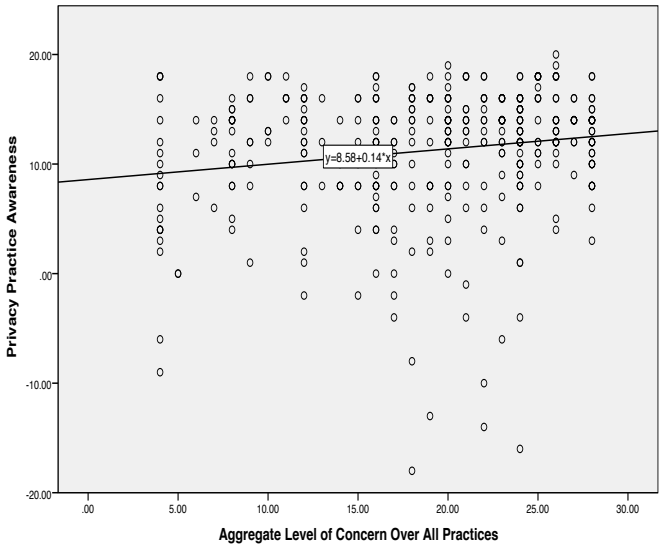


Figure 11: Relationship between awareness of privacy practices and overall concern with Fitbit practices: for every point of concern, there is an increase of .14 in awareness of privacy practices (or .14 more questions answered correctly).

we did find some relationships between time spent and overall privacy awareness.

In addition to analyzing time as a continuous variable, we also binned time into three segments: less than 20 seconds, between 20 and 60 seconds, and more than 60 seconds. We chose these bins as we did not think someone could read through the entire privacy notice in less than 20 seconds, but that almost anyone could read through the notice in 60 seconds, and would be examining it closely (or were distracted by another task) if they looked at it for longer.

We found that regardless of whether time was measured as continuous or binned, there was no difference in time spent on notice between length conditions. The distribution of participants by condition in each bin is shown in Figure 10. Using Pearson's Chi-Square test, we found no relationship between binned time and length of notice. The overall length of time spent on notice by condition is shown in Figure 9. Performing a Kruskal-Wallis test, we found no statistically significant differences in the length of time spent on the notice, and condition.

Performing a linear regression, we did not find a statistically significant relationship between time spent on the notice and overall privacy practice awareness. However, performing a Kruskal-Wallis test, we did find that binned time had an effect on overall privacy practice awareness ($H(2,349)= 26.89$, p<.0005). Using Tukey correction for multiple testing, we compared each binned time with Mann-Whitney U tests. We found that bin 0 (<20 seconds) was significantly outperformed (Mean=8.59, SD=8.09) by both bin 1 (20-60 seconds, Mean=12.87, SD=4.61) and bin 2 (>60 seconds, Mean =13.26, SD=4.03), with ($U(1,274)$=-4.839, p<.001, r=.292) and ($U(1,150)$=-4.431, p<.001, r=.361) respectively, in terms of privacy practice awareness. We found no significant difference between bins 1 and 2. This suggests that there is a difference between glancing at a notice and reading the notice, but how much time is spent reading or studying the notice may not matter as much.

5.5 Impact of Privacy Concern

We measured participants' privacy concern in two ways. First, we asked participants to rate their concern with Fitbit's privacy practices at the aggregate levels of collection practices, sharing practices, selling practices, and storage practices on a 7-point Likert scale from not very concerned to very concerned (see Q. 30 in Appendix). Second, participants completed the 10-item IUIPC questionnaire [32], which results in scales for awareness, collection, and control on a 7-point Likert scale from strongly disagree to strongly agree (see Q's 44-53 in Appendix A).

5.5.1 Concern With Fitbit Privacy Practices

We found that participants were most concerned with Fitbit's sharing practices, participants with greater concern had greater privacy practice awareness, and there was no significant relationship between framing and concern.

We performed a Friedman test, finding participant concern was greatest for sharing practices (Mean =5.06, SD=1.89), compared to collection (Mean=4.52, SD= 1.84), selling (Mean=4.72, SD=1.93), and storage (Mean=4.70, SD = 1.93) with ($\chi^2(3,388)$= 74.32, p <.0005). Performing pair-wise Wilcoxon tests with post-hoc correction, we found that concern with sharing practices was significantly higher than collection practices ($Z(1,390)$=-7.968, p< .003), concern with storage practices was significantly higher than collection practices ($Z(1,390)$=-2.782, p=.030), concern with sharing practices was significantly higher than concern with selling practices (Z=-5.051, p<.0030), concern with sharing practices was significantly higher than concern with storage practices (Z=-5.696, p<.0030).

We had originally hypothesized that framing would lead to greater concern, causing participants to pause to reflect on the practices in the policies to a greater extent. The second part of this hypothesis appears to be correct, as can be seen in Figure 11. Performing a linear regression, we found that for every increase in overall concern over privacy practices, there was a .146 increase in overall privacy practice awareness, with (t=3.488, p=.001). However, performing a Kruskal-Wallis test, we found no relationship between condition and concern, as can be seen in Figure 12. We additionally tested whether aggregating notices by their framing and excluding the control condition made any difference. However, a Kruskal-

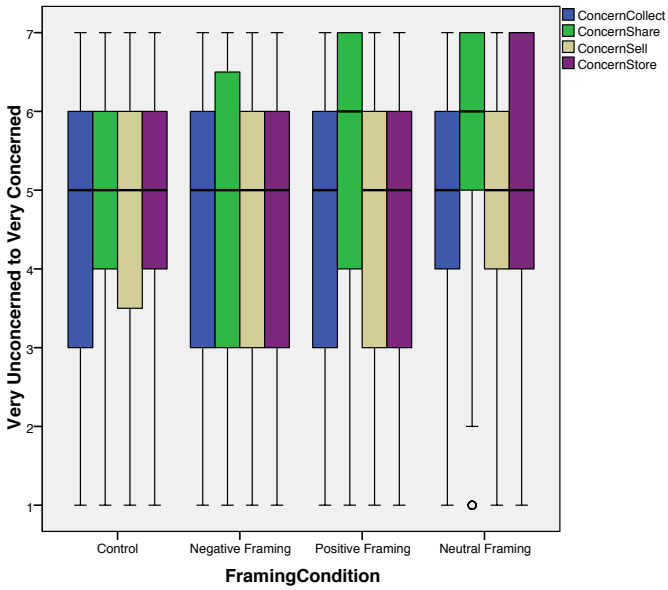


Figure 12: Concern with Fitbit privacy practices by framing condition. While concern over sharing personally identifiable information was slightly greater, we found no statistically significant differences between conditions.

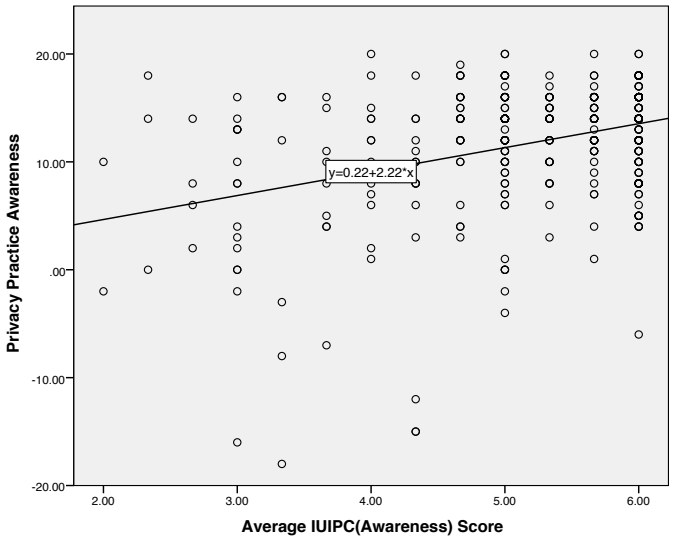


Figure 13: Relationship between the IUIPC awareness scale and awareness of Fitbit's privacy practices. For every point of agreement with an IUIPC awareness question, awareness of Fitbit privacy practices increased by 2.22.

Wallis test did not reveal significant differences. It seems that our framing conditions fail to impact overall privacy practice awareness because they fail to impact participant concern.

5.5.2 IUIPC Concern

Prior work has shown a significant relationship between IUIPC scores and putative online privacy behavior [32]. Therefore, we examined the relationship between the IUIPC scales and participants' awareness. Performing a linear regression, we found the IUIPC awareness scale was positively associated with awareness of privacy practices, with every point of agreement with the IUIPC awareness questions leading to (on average) an increase of 2.221 in overall awareness of privacy practices ($p<.0005$), see Figure 13.

However, performing a Kruskal-Wallis test, we found no relationship between condition and any of the IUIPC scales, suggesting agreement with IUIPC variables was not noticeably affected by notices, framing, or length of notices, see Figure 14.

On the whole this confirms the effectiveness of the IUIPC scales to predict overall privacy concerns of participants. It also demonstrates that our framing did not affect participant concern about online privacy in general, as measured by IUIPC questions.

6. DISCUSSION

We explored the idea that shorter short-form privacy notices focusing on less expected privacy practices might lead to greater awareness of privacy practices. We specifically investigated this approach in the context of fitness wearables' privacy practices. We measured success by participant awareness of Fitbit's privacy practices.

We first discuss potential limitations of our study design. We then discuss the effectiveness of privacy notices, the specific effects (or lack thereof) of our enhancements, as well as explanations for these effects from the data, and implications for notice design.

6.1 Limitations

It is unclear how generalizable our results are, as our surveys focused on a single context, the privacy policy of a single company, one specific wording of that policy, and one specific device. We chose to focus on Fitbit as it is the market leader in fitness wearables [25], and the Fitbit Surge as it was Fitbit's newest product at the time our research commenced. Our examination of other fitness wearable manufacturers (e.g. Jawbone, Misfit), found their policies to be functionally similar to Fitbit's [26, 36]. More importantly, the focus of our research – improving privacy notice design through framing or length – is not specific to Fitbit or even fitness wearables, except for the privacy practices we displayed in the tested notice formats. Our notice-development process could be applied to any company's privacy policy.

It is possible that some of our results can be attributed to the wording of the short-form privacy notice we tested. For example, we included wording intended to inform participants that the short-form notices did not contain all of Fitbit's data collection or sharing practices. However, we did not directly investigate whether participants understood that. In addition, we tested only one set of words for our framing conditions. It is possible that other approaches to framing might have produced different results.

Another potential limitation is the use of MTurk for conducting surveys. Some prior work has shown that MTurkers can differ from the general population, and that individuals may interact with a survey differently than they would in reality [23]. Other research has shown that MTurkers constitute a reasonably good sample of the general population [11]. We addressed this potential problem in two ways: first, our survey was consistently designed to elicit natural reactions to privacy notices. Our recruitment materials did not mention privacy or security, participants were informed at the beginning of the survey that they would not be penalized for wrong answers, and at no point did we force participants to look at privacy notices, but instead we let them click through to the next page of the survey if they so chose. These design decisions were meant to, as closely as possible, mirror a participant's actual interaction with privacy policies and privacy notices. Secondly, we examined relative effectiveness of our various design decisions, with a control group included, which should mitigate biasing effects.

A related potential limitation is the direct confrontation of participants with a privacy notice. We chose this approach to reduce

variations in participants' attention. This provided us with a best case scenario for a comparative assessment of how notice length, framing, and other characteristics impact participants' awareness of privacy practices. We expect that under real conditions, participants would likely perform worse, due to distractions and lack of attention to the notice. Since we did not observe framing effects in our study, it is unlikely that they would surface in a field study with the type of privacy notice we focused on.

6.2 Privacy Notices Can Be Effective

An important result from our work is demonstrating that short-form privacy notices uniformly led to significantly higher awareness than the control. This result, while a reconfirmation of the basic effectiveness of privacy notices [43] is important for two further reasons. First, fitness wearables generally collect data that is inherent to their function (e.g., steps, distance, heart-rate). It was therefore possible that since many of Fitbit's privacy practices would be linked to the function of the device, participants might have had a higher awareness of such practices without seeing a privacy notice. This was not the case.

Second, Fitbit does not currently have comparable short-form privacy notices. Our results show a practical method by which Fitbit and other fitness wearable manufacturers could increase user awareness of their privacy practices by integrating privacy notices similar to ours into their mobile companion apps or websites.

6.2.1 Framing Did Not Affect Concern

The results from our analysis of participants' reported concern over Fitbit's privacy practices provide a potential explanation for the lack of significant difference we found between framing conditions. We found no significant difference in concern with Fitbit's privacy practices or general privacy concern (IUIPC) and the framing conditions. In other words, framing some practices in a positive or negative light did not seem to make a difference in how concerned participants were about them. However, the lack of change in level of concern suggests that this was due to a lack of effectiveness of our chosen framing technique, and not a failure in the underlying concept of framing itself. Including framing statements that lead to greater or lesser participant concern might very well lead to greater or lesser awareness of policies. This could be done through heightening the focus on risk and implications, or by including personalized information, such as the data that could be re-identified for the particular user receiving the notice.

6.2.2 Shortest Notices Led to Less Awareness

Our results show that removing well-known privacy practices to make short-form notices even shorter actually led to similar or worse participant awareness of privacy practices. Our intuition was that further condensing a short-form privacy notice would lead to even better performance, provided that the practices removed were well known. However, this intuition proved false, as our results show no increase in awareness of the practices remaining in the notice when some practices are removed.

Our medium length notices did not result in significantly different performance compared to our longest notices. This suggests that removing some of the most known practices had little effect on participant awareness, but that there may be some benefits of using such medium notices when space is constrained.

Our shortest notices performed significantly worse than our longest notices, suggesting that there may be a lower bound to the length of an effective privacy notice. In addition, the awareness threshold we selected for removing practices from the shortest notice may have been too low.

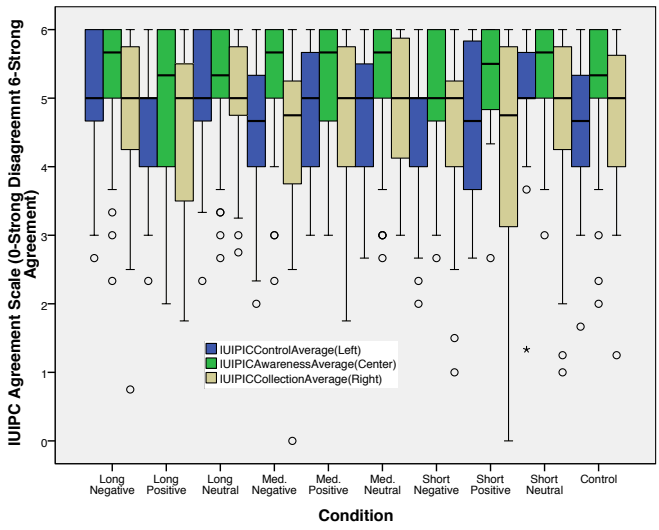


Figure 14: Average agreement with IUIPC scales. No significant difference between conditions. Awareness questions had the highest agreement, followed by control and collection.

Analyzing the time spent on notices does not make the picture clearer. As part of our study design, we did not force participants to look at our notices for a set period of time, instead they could click through to the next page immediately if they so chose. We made this design decision to increase ecological validity, since in the real world users can generally quickly click through a privacy policy. We did, however, record the amount of time participants stayed on the page with the notice. We examined this time as both a continuous variable, as well as binning it into three time lengths based on our estimation of the time necessary to read through the notice. We found that there was no significant relationship between continuous time and participant awareness. However, binned time showed that participants in the larger time bins had significantly higher awareness of privacy practices. Our results also showed that there was no significant difference in time spent on notices by condition (either length or framing). Given the length disparity between our long and short notices (see Figures 4 and 6), we expected participants to be able to spend far longer on the remaining privacy practices in the short notice, and therefore have better awareness of these practices. However, we did not find such a difference.

It is possible that participants spent their time looking at those practices that were unknown or alien to them, with only very brief confirmation of those practices which they assumed or were well known. Participants therefore would have spent a roughly equivalent amount of time on the lesser known privacy practices regardless of length, and participants in the short length notice conditions did not have the benefit of quick confirmations of practices they were already aware of, leading to worse awareness of these practices. It is also possible that even the long privacy notice we created was short enough to achieve all of the gains from condensing a privacy notice, and that shortening a notice for a more complex and lengthy privacy policy could achieve better results.

6.3 Importance of Participant Factors

Our sample was diverse with respect to age, gender, and experience with fitness wearables. Interestingly, we found that each of these participant factors had at least some statistically significant effect on awareness of privacy practices; with older participants and

women having significantly higher awareness of Fitbit's privacy practices. While not the focus of our study, these results are important as they showcase that awareness of privacy practices varies based on demographic factors. This demonstrates that user studies on the effectiveness of a privacy notice should be conducted with a diverse sample in order to account for demographic differences or should target specific audiences with a specific notice design.

7. CONCLUSIONS AND FUTURE WORK

We presented in this paper a series of three MTurk user studies. Our first survey was focused on determining an effective design format for a Fitbit short-form privacy notice. Our second survey focused on determining participant awareness of each of 20 Fitbit privacy practices. Our final study examined the potential for removing generally expected privacy practices from notices, as well as including framing statements in notices, to increase participant awareness of privacy practices.

Our results reconfirmed the utility of short-form privacy notices, as all notice conditions outperformed the control. However, we also found that while condensing long legalistic privacy policies into succinct privacy notices does increase awareness, taking this a step further by further condensing privacy notices to succinctly include only practices that users are not generally aware of, had the opposite effect. Participants with shorter notices had similar performance on practices that were left in the notice, but performed significantly worse on practices that were removed. Additionally, incorporating positive and negative framing statements into our privacy notices did not bear fruit, with no statistically significant difference in performance. Our analysis of participant concern over Fitbit privacy practices suggests that this lack of effect was due to insufficient differences in the level of concern between framing conditions to elicit significant changes in awareness.

Given these results, we suspect that a lower bound for the potential to compress privacy notices exists, and that further research should focus on personalization of privacy notices [5, 22, 30, 48], or in the timing of the notices (e.g. just-in-time notification, or notification on a regular basis rather than on purchase/install) [7, 17, 30]. That said, further studies investigating the effectiveness of generic short-form privacy notices may be able to address some of the limitations of our study and shed additional light on ways notices may be shortened effectively.

8. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under grants CNS-1012763, CNS-1330596, SBE-1513957 and CSR-1526237, as well as by DARPA and the Air Force Research Laboratory, under agreement number FA8750-15-2-0277. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, the Air Force Research Laboratory, the National Science Foundation, or the U.S. Government.

The authors would like to thank Blase Ur and Ariel Polakoff for their input and feedback.

9. REFERENCES

[1] A. Acquisti, I. Adjerid, and L. Brandimarte. Gone in 15 seconds: The limits of privacy transparency and control. *IEEE Security & Privacy*, (4):72–74, 2013.

[2] A. Acquisti, L. Brandimarte, and G. Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015.

[3] I. Adjerid, A. Acquisti, L. Brandimarte, and G. Loewenstein. Sleights of privacy: Framing, disclosures, and the limits of transparency. In *Proc. SOUPS '13*. ACM, 2013.

[4] Y. Agarwal and M. Hall. Protectmyprivacy: Detecting and mitigating privacy leaks on iOS devices using crowdsourcing. In *Proc. MobiSys '13*, pages 97–110. ACM, 2013.

[5] H. Almuhimedi, F. Schaub, N. Sadeh, I. Adjerid, A. Acquisti, J. Gluck, L. F. Cranor, and Y. Agarwal. Your location has been shared 5,398 times!: A field study on mobile app privacy nudging. In *Proc. CHI '15*, pages 787–796. ACM, 2015.

[6] I. Ayres and A. Schwartz. No-Reading Problem in Consumer Contract Law, The. *Stanford Law Review*, 66:545, 2014.

[7] R. Balebako, J. Jung, W. Lu, L. F. Cranor, and C. Nguyen. Little brothers watching you: Raising awareness of data leaks on smartphones. In *Proc. SOUPS '13*. ACM, 2013.

[8] R. Balebako, F. Schaub, I. Adjerid, A. Acquisti, and L. Cranor. The impact of timing on the salience of smartphone app privacy notices. In *Proceedings of the 5th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices*, SPSM '15, pages 63–74, New York, NY, USA, 2015. ACM.

[9] O. Ben-Shahar and A. S. Chilton. 'Best Practices' in the Design of Privacy Disclosures: An Experimental Test. SSRN ID 2670115, Oct. 2015.

[10] R. Calo. Against Notice Skepticism In Privacy (And Elsewhere). SSRN ID 1790144, Mar. 2011.

[11] K. Casler, L. Bickel, and E. Hackett. Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, 29(6):2156–2160, 2013.

[12] F. H. Cate. The limits of notice and choice. *Security & Privacy, IEEE*, 8(2):59–62, 2010.

[13] Center for Information Policy Leadership. Ten Steps to Develop a Multilayered Privacy Notice. White paper, Mar. 2007.

[14] L. F. Cranor, P. Guduru, and M. Arjula. User interfaces for privacy agents. *ACM Trans. Comput.-Hum. Interact.*, 13(2):135–178, June 2006.

[15] J. B. Earp, Q. He, W. Stufflebeam, D. Bolchini, C. Jensen, and others. Financial privacy policies and the need for standardization. *IEEE Security & privacy*, (2):36–45, 2004.

[16] Federal Trade Commission. Protecting Consumer Privacy in an Era of Rapid Change, Mar. 2012.

[17] Federal Trade Commission. Internet of Things: Privacy & security in a connected world, 2015.

[18] Federal Trade Commission and Kleimann Communication Group. Evolution of a Prototype Financial Privacy Notice: A Report on the Form Development Project, Feb. 2006.

[19] Fitbit inc. Fitbit Privacy Policy, December 2014. Available at https://www.fitbit.com/legal/privacy-policy.

[20] C. Gates, J. Chen, N. Li, and R. Proctor. Effective risk communication for Android apps. *IEEE Trans. Dependable and Secure Computing*, 11(3):252–265, May 2014.

[21] N. Good, R. Dhamija, J. Grossklags, D. Thaw, S. Aronowitz, D. Mulligan, and J. Konstan. Stopping spyware at the gate: A user study of privacy, notice and spyware. In *Proc. SOUPS '05*, pages 43–52. ACM, 2005.

[22] M. Harbach, M. Hettig, S. Weber, and M. Smith. Using personal examples to improve risk communication for security & privacy decisions. In *Proc. CHI '14*, pages 2647–2656. ACM, 2014.

[23] D. J. Hauser and N. Schwarz. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods*, pages 1–8, 2015.

[24] A. Hilts, C. Parsons, and J. Knockel. Every Step You Fake: A Comparative Analysis of Fitness Tracker Privacy and Security. 2016. Available at http://citizenlab.org/2016/02/fitness-tracker-privacy-and-security/.

[25] IDC. Worldwide Wearables Market Soars in the Third Quarter as Chinese Vendors Challenge the Market Leaders, Dec 2015. Available at http://www.idc.com/getdoc.jsp?containerId=prUS40674715.

[26] Jawbone. UP privacy policy, December 2014. Available at https://jawbone.com/up/privacy.

[27] C. Jensen and C. Potts. Privacy policies as decision-making tools: An evaluation of online privacy notices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 471–478, New York, NY, USA, 2004. ACM.

[28] P. G. Kelley, J. Bresee, L. F. Cranor, and R. W. Reeder. A nutrition label for privacy. In *Proc. SOUPS '09*. ACM, 2009.

[29] P. G. Kelley, L. Cesca, J. Bresee, and L. F. Cranor. Standardizing privacy notices: An online study of the nutrition label approach. In *Proc. CHI '10*, pages 1573–1582. ACM, 2010.

[30] A. Kobsa and M. Teltzrow. Contextualized communication of privacy practices and personalization benefits: Impacts on users' data sharing and purchase behavior. In *Privacy Enhancing Technologies*, pages 329–343. Springer, 2004.

[31] A. Levy and M. Hastak. Consumer Comprehension of Financial Privacy Notices. *Interagency Notice Project, http://ftc. gov/privacy/privacyinitiatives/Levy-Hastak-Report. pdf*, 2008.

[32] N. K. Malhotra, S. S. Kim, and J. Agarwal. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research*, 15(4):336–355, 2004.

[33] A. M. McDonald and L. F. Cranor. Cost of reading privacy policies, the. *ISJLP*, 4:543, 2008.

[34] A. M. McDonald, R. W. Reeder, P. Kelley, and L. Faith. A Comparative Study of Online Privacy Policies and Formats. *Privacy Enhancing Technologies*, 2009.

[35] Microsoft. Privacy Guidelines for Developing Software Products and Services. Technical Report version 3.1. 2008.

[36] Misfit Inc. Misfit Privacy Policy, June 2015. Available at http://misfit.com/legal/privacy_policy.

[37] OECD. Making Privacy Notices Simple. Digital Economy Papers 120, July 2006.

[38] S. Patil, R. Hoyle, R. Schlegel, A. Kapadia, and A. J. Lee. Interrupt now or inform later?: Comparing immediate and delayed privacy feedback. In *Proc. CHI '15*. ACM, 2015.

[39] A. Patrick and S. Kenny. From privacy legislation to interface design: Implementing information privacy in human-computer interactions. In *Proc. PET '03*. Springer, 2003.

[40] President's Concil of Advisors on Science and Technology. Big data and privacy: A technological perspective. Report to the President, Executive Office of the President, May 2014.

[41] A. Rao, F. Schaub, N. Sadeh, A. Acquisti, and R. Kang. Expecting the Unexpected: Understanding Mismatched Privacy Expectations Online. In *Proc. SOUPS '16*. USENIX Assoc., 2016.

[42] J. R. Reidenberg, T. Breaux, L. F. Cranor, B. French, A. Grannis, J. T. Graves, F. Liu, A. McDonald, T. B. Norton, R. Ramanath, N. C. Russell, N. Sadeh, and F. Schaub. Disagreeable Privacy Policies: Mismatches Between Meaning and Users' Understanding. *Berkeley Tech. LJ*, 30:39, 2015.

[43] F. Schaub, R. Balebako, A. L. Durity, and L. F. Cranor. A Design Space for Effective Privacy Notices. In *Proc. SOUPS '15)*, pages 1–17. USENIX Assoc., 2015.

[44] F. Shih, I. Liccardi, and D. Weitzner. Privacy Tipping Points in Smartphones Privacy Preferences. In *Proc. CHI '15*, pages 807–816. ACM, 2015.

[45] J. Tan, K. Nguyen, M. Theodorides, H. Negrón-Arroyo, C. Thompson, S. Egelman, and D. Wagner. The effect of developer-specified explanations for permission requests on smartphone user behavior. In *Proc. CHI '14*. ACM, 2014.

[46] T. F. Waddell, J. R. Auriemma, and S. S. Sundar. Make It Simple, or Force Users to Read?: Paraphrased Design Improves Comprehension of End User License Agreements. In *Proc. CHI '16*, pages 5252–5256. ACM, 2016.

[47] S. Wilson, F. Schaub, R. Ramanath, N. Sadeh, F. Liu, N. Smith, and F. Liu. Crowdsourcing annotations for websites' privacy policies: Can it really work? In *Proc. WWW '16*, 2016.

[48] M. S. Wogalter, B. M. Racicot, M. J. Kalsher, and S. N. Simpson. Personalization of warning signs: the role of perceived relevance on behavioral compliance. *International Journal of Industrial Ergonomics*, 14(3):233–242, 1994.

[49] S. Zimmeck and S. M. Bellovin. Privee: An Architecture for Automatically Analyzing Web Privacy Policies. In *USENIX Security Symposium*. USENIX Assoc., 2014.

APPENDIX
Length and Framing Survey Questions

Opinion Survey on Fitness Wearables

1) What is your age? (type "0" if you prefer not to answer)*
_____

2) What is your gender?*
( ) Male
( ) Female
( ) Other
( ) Prefer Not to Answer

3) Have you earned a degree in or held a job in computer science, IT, electrical engineering, or a related field?*
( ) Yes
( ) No

4) What is your prior experience with wearable fitness devices (devices you wear which collect fitness information), such as Fitbit, Jawbone, Garmin, or Misfit devices?*
( ) I currently use a wearable fitness device
( ) In the past I regularly used a wearable fitness device, but I no longer do so
( ) I have tried out a wearable fitness device, but have never regularly used one
( ) I have never used a wearable fitness device, but am familiar with the concept
( ) I was unfamiliar with wearable fitness devices before taking this survey

5) Which of the following fitness devices have you used in the past? *
[ ] A Fitbit Product
[ ] A Jawbone Product
[ ] A Garmin Product
[ ] A Misfit Product
[ ] I've used a fitness wearable, but not one listed
[ ] I've never used a fitness wearable before
[ ] I don't remember

6) What specific model(s) of fitness wearable(s) did you use?*
_____

7) How trustworthy or untrustworthy are the following fitness wearable companies in your opinion?*

	Very trust worthy	Trust worthy	Somewhat trust worthy	Neutral	Somewhat untrust worthy	Untrust worthy	Very untrust worthy	I don't know
Jawbone	( )	( )	( )	( )	( )	( )	( )	( )
Misfit	( )	( )	( )	( )	( )	( )	( )	( )
Garmin	( )	( )	( )	( )	( )	( )	( )	( )
Fitbit	( )	( )	( )	( )	( )	( )	( )	( )

8) How protective or unprotective of your privacy do you think the following fitness wearable companies and their privacy policies are?*

	Very protect tive	Protec tive	Somewhat protective	Neutral	Somewhat unprotect ive	Unprotec tive	Very unprotect ive	I don't know
Jawbone	( )	( )	( )	( )	( )	( )	( )	( )
Misfit	( )	( )	( )	( )	( )	( )	( )	( )
Garmin	( )	( )	( )	( )	( )	( )	( )	( )
Fitbit	( )	( )	( )	( )	( )	( )	( )	( )

Action: Page Timer 2 minutes 30 seconds before able to move on
9) How much does a Fitbit Surge Watch Cost?*
( ) $100
( ) $150
( ) $200
( ) $250
( ) I Don't Know
10) Would you consider using a Fitbit Surge Watch?*
( ) Yes
( ) Yes, except it costs too much
( ) No
( ) I don't know
( ) I prefer not to answer
11) Can you explain your answer to the question above?*
_____

Privacy Notice Shown here (page skipped if in control)

---

12) Imagine you are using a Fitbit Surge, which of the following types of information do you think Fitbit would collect about you?*

	Definitely Collects	Probably Collects	Might Collect	Might not Collect	Probably Does not Collect	Definitely Does not Collect	I'm Un sure
Your perspiration rate	( )	( )	( )	( )	( )	( )	( )
Your mood	( )	( )	( )	( )	( )	( )	( )
Your altitude	( )	( )	( )	( )	( )	( )	( )
Your shoe size	( )	( )	( )	( )	( )	( )	( )
How many steps you've taken	( )	( )	( )	( )	( )	( )	( )
How far you've walked	( )	( )	( )	( )	( )	( )	( )
Information you've posted to your Fitbit profile	( )	( )	( )	( )	( )	( )	( )
Your name	( )	( )	( )	( )	( )	( )	( )
A list of your Facebook friends	( )	( )	( )	( )	( )	( )	( )
When you exercise	( )	( )	( )	( )	( )	( )	( )
Your heartrate	( )	( )	( )	( )	( )	( )	( )
Your height	( )	( )	( )	( )	( )	( )	( )
Your weight	( )	( )	( )	( )	( )	( )	( )

How many sets of stairs you've climbed	( )	( )	( )	( )	( )	( )	( )
The part of your body you wear the Fitbit	( )	( )	( )	( )	( )	( )	( )
How often you exercise compared to friends who also have Fitbit devices	( )	( )	( )	( )	( )	( )	( )
How often you sleep	( )	( )	( )	( )	( )	( )	( )
How often you exercise in the dark	( )	( )	( )	( )	( )	( )	( )
Who was using the Fitbit Device	( )	( )	( )	( )	( )	( )	( )
Your location	( )	( )	( )	( )	( )	( )	( )

13) Are there any other types of data that you think the Fitbit Surge collects about you?

_____

14) Which of the following groups do you think Fitbit will share your personally identifiable information (information that could be used to identify you) with by default?*

	Definitely Shares	**Probably Shares**	**Might Share**	**Might not Share**	**Probably Does not Share**	**Definitely Does not Share**	**I'm Unsure**
Government entities	( )	( )	( )	( )	( )	( )	( )
Your Fitbit Friends (friends you've added on Fitbit)	( )	( )	( )	( )	( )	( )	( )
Companies providing services to Fitbit	( )	( )	( )	( )	( )	( )	( )
Researchers studying fitness and health aspects of wearables	( )	( )	( )	( )	( )	( )	( )
No one, Fitbit does not share your personally identifiable information with anyone	( )	( )	( )	( )	( )	( )	( )
Anyone who requests your information	( )	( )	( )	( )	( )	( )	( )

Organizations you direct Fitbit to share your information with	( )	( )	( )	( )	( )	( )	( )

15) Are there any other groups that you believe that Fitbit shares your personally identifiable information with by default?

_____

16) Do you think Fitbit allows you to control how information is shared with your Fitbit friends? *

( ) No, anyone you add as a Fitbit friend can see all of your Fitness data
( ) Yes, you can opt-out of sharing specific forms of data with your Fitbit friends on the Fitbit website
( ) Yes, you can opt-out of sharing ANY data with your Fitbit friends on the Fitbit website, but it is all or nothing
( ) No, Fitbit doesn't share your information with Fitbit friends
( ) None of the above
( ) I don't know

17) How confident are you in your answer to the question above? (Do you think Fitbit allows you to control sharing information with your Fitbit friends)*

( ) Very Unconfident ( ) Unconfident ( ) Somewhat Unconfident ( ) Neutral ( ) Somewhat Confident ( ) Confident ( ) Very Confident

18) Under what conditions do you think Fitbit may sell your data?*

( ) Whenever they want, with no restrictions
( ) Whenever they want, as long as your real name and address are not attached to the data profile
( ) They can sell aggregated, de-identified data that does not identify you
( ) They can sell aggregated, de-identified data that does not identify you, but only if you opt-in (choose to let them do it)
( ) Never; they cannot sell your data
( ) None of the Above
( ) I don't know

19) How confident are you in your answer to the question above? (Under what conditions do you think Fitbit may sell your data)*

( ) Very Unconfident ( ) Unconfident ( ) Somewhat Unconfident ( ) Neutral ( ) Somewhat Confident ( ) Confident ( ) Very Confident

20) When do you think Fitbit can collect your location?*

( ) Fitbit can never collect my location
( ) Fitbit can only collect my location if I choose to let them (opt-in)
( ) Fitbit will collect my location when location features, such as maps, of my Fitbit device are active
( ) Fitbit always collects my location
( ) None of the Above
( ) I don't know

21) How confident are you in your answer to the question above? (When do you think Fitbit can collect your location?)*

( ) Very Unconfident ( ) Unconfident ( ) Somewhat Unconfident ( ) Neutral ( ) Somewhat Confident ( ) Confident ( ) Very Confident

---

22) For how long do you think Fitbit keeps the data it collects?*

( ) Until that data item has not been accessed for 6 months
( ) Until you remove an item from your profile or Fitbit device
( ) Until you fully delete your Fitbit account
( ) Forever; it never deletes the data even if you delete your account
( ) None of the above
( ) I don't know

23) How confident are you in your answer to the question above? (For how long do you think Fitbit keeps the data it collects)*

( ) Very Unconfident ( ) Unconfident ( ) Somewhat Unconfident ( ) Neutral ( ) Somewhat Confident ( ) Confident ( ) Very Confident

24) In the event of a data breach of some of its consumer data, how soon do you think Fitbit will contact its users to let them know that their data has been stolen?*

( ) Within 1 week
( ) Within 1 month
( ) Within 3 months
( ) As specified by law
( ) Never
( ) I don't know

25) How confident are you in your answer to the question above? (In the event of a data breach of some of its consumer data, how soon do you think Fitbit will contact its users to let them know that their data has been stolen)*

( ) Very Unconfident ( ) Unconfident ( ) Somewhat Unconfident ( ) Neutral ( ) Somewhat Confident ( ) Confident ( ) Very Confident

26) Do you think you can use a Fitbit device without having a Fitbit account?*

( ) Yes and the device will function the same way as with an account
( ) Yes, but only basic functions will work, such as distance, heartrate and step count.
( ) Yes, but without an account to maintain calibration data, it won't count steps correctly

( ) No. The Fitbit device won't function if it's not connected to an account
( ) None of the above
( ) I don't know

27) How confident are you in your answer to the question above? (Do you think you can use a Fitbit device without having a Fitbit account)*

( ) Very Unconfident ( ) Unconfident ( ) Somewhat Unconfident ( ) Neutral ( ) Somewhat Confident ( ) Confident ( ) Very Confident

28) Where do you think can you find details about Fitbit's Privacy Policy? (select all options that apply)*

[ ] In the Fitbit mobile phone app
[ ] On the Fitbit website
[ ] On a paper insert in the box Fitbit devices comes in
[ ] Fitbit does not have a privacy policy
[ ] None of the above
[ ] I don't know

29) How confident are you in your answer to the question above? (Where do you think can you find details about Fitbit's Privacy Policy)*

( ) Very Unconfident ( ) Unconfident ( ) Somewhat Unconfident ( ) Neutral ( ) Somewhat Confident ( ) Confident ( ) Very Confident

---

30) Please rate the extent to which the following Fitbit practices concern you.*

	Very Unconcerned	Unconcerned	Somewhat Unconcerned	Neutral	Somewhat Concerned	Concerned	Very Concerned
What data Fitbit collects	( )	( )	( )	( )	( )	( )	( )
With whom Fitbit shares data that	( )	( )	( )	( )	( )	( )	( )
With whom Fitbit sells my de-identified	( )	( )	( )	( )	( )	( )	( )
How and for how long Fitbit stores my data	( )	( )	( )	( )	( )	( )	( )

31) Please explain your answer(s) to the question above.

_____

32) Please indicate the extent to which Fitbit's collection of the following forms of data concerns you.*

	Very Unconcerned	Unconcerned	Somewhat Unconcerned	Neutral	Somewhat Concerned	Concerned	Very Concerned
Your location, when location features of your Fitbit are active	( )	( )	( )	( )	( )	( )	( )
Name	( )	( )	( )	( )	( )	( )	( )
Height	( )	( )	( )	( )	( )	( )	( )
Weight	( )	( )	( )	( )	( )	( )	( )
Amount of	( )	( )	( )	( )	( )	( )	( )

exercise							
Time of exercise	( )	( )	( )	( )	( )	( )	( )
Sleeping habits	( )	( )	( )	( )	( )	( )	( )
Information posted to your Fitbit profile	( )	( )	( )	( )	( )	( )	( )

33) Please explain your answer(s) to the question above

_____

34) Please indicate the degree to which you are concerned with Fitbit sharing your personally indentifiable information with the following groups.*

	Very Unconcerned	**Unconcerned**	**Somewhat Unconcerned**	**Neutral**	**Somewhat Concerned**	**Concerned**	**Very Concerned**
Government entities	( )	( )	( )	( )	( )	( )	( )
Organizations providing services to Fitbit	( )	( )	( )	( )	( )	( )	( )
Your Fitbit Friends	( )	( )	( )	( )	( )	( )	( )
Organizations you specifically direct Fitbit to share data with (e.g. Facebook)	( )	( )	( )	( )	( )	( )	( )

35) Please explain your answer(s) to the question above

_____

36) How would you feel about Fitbit collecting and sharing your location while using the device?*
( ) Completely uncomfortable ( ) Uncomfortable ( ) Somewhat uncomfortable ( ) Neutral ( ) Somewhat comfortable ( ) Comfortable ( ) Very Comfortable

37) How would you feel about Fitbit keeping a copy of all your data, including data you deleted, until you fully delete your entire Fitbit account?*
( ) Very uncomfortable ( ) Uncomfortable ( ) Somewhat uncomfortable ( ) Neutral ( ) Somewhat comfortable ( ) Comfortable ( ) Very comfortable

38) How would you feel about Fitbit sharing all of your fitness data by default, such as exercise and food consumption, with your Facebook friends?*
( ) Very uncomfortable ( ) Uncomfortable ( ) Somewhat uncomfortable ( ) Neutral ( ) Somewhat comfortable ( ) Comfortable ( ) Very comfortable

39) How would you feel about Fitbit sharing all of your fitness data, such as exercise and food consumption, with friends you add on Fitbit?*
( ) Very uncomfortable ( ) Uncomfortable ( ) Somewhat uncomfrotable ( ) Neutral ( ) Somewhat comfortable ( ) Comfortable ( ) Very comfortable

40) How would you feel about Fitbit sharing your personally identifiable information with companies providing services to Fitbit, with no limit to what those companies can do with your information, provided they don't share it?*
( ) Very uncomfortable ( ) Uncomfortable ( ) Somewhat uncomfortable ( ) Netural ( ) Somewhat comfortable ( ) Comfortable ( ) Very comfortable

41) How would you feel about Fitbit selling your personally identifiable information (information that identifies you) to other companies?*

( ) Very uncomfortable ( ) Uncomfortable ( ) Somewhat uncomfortable ( ) Neutral ( ) Somewhat comfortable ( ) Comfortable ( ) Very comfortable

42) How would you feel about Fitbit selling your information as part of a de-identified, aggregated block (does not identify you) to other companies?*

( ) Very uncomfortable ( ) Uncomfortable ( ) Somewhat uncomfortable ( ) Neutral ( ) Somewhat comfortable ( ) Comfortable ( ) Very comfortable

43) How would you rate your desire to buy and use a Fitbit product in the future?*

( ) No Desire ( ) Little Desire ( ) Some Desire ( ) A lot of Desire ( ) I already own and use another fitness wearable device

---

44) Consumer online privacy is really a matter of consumers' right to exercise control and autonomy over decisions about how their information is collected, used, and shared. *

( ) Strongly Disagree ( ) Disagree ( ) Mildly Disagree ( ) Neutral ( ) Mildly Agree ( ) Agree ( ) Strongly Agree

45) Consumer control of personal information lies at the heart of consumer privacy. *

( ) Strongly Disagree ( ) Disagree ( ) Mildly Disagree ( ) Neutral ( ) Mildly Agree ( ) Agree ( ) Strongly Agree

46) I believe that online privacy is invaded when control is lost or unwillingly reduced as a result of a marketing transaction.*

( ) Strongly Disagree ( ) Disagree ( ) Mildly Disagree ( ) Neutral ( ) Mildly Agree ( ) Agree ( ) Strongly Agree

47) Companies seeking information online should disclose the way the data are collected, processed, and used.*

( ) Strongly Disagree ( ) Disagree ( ) Mildly Disagree ( ) Neutral ( ) Mildly Agree ( ) Agree ( ) Strongly Agree

48) A good consumer online privacy policy should have a clear and conspicuous disclosure. *

( ) Strongly Disagree ( ) Disagree ( ) Mildly Disagree ( ) Neutral ( ) Mildly Agree ( ) Agree ( ) Strongly Agree

49) It is very important to me that I am aware and knowledgeable about how my personal information will be used. *

( ) Strongly Disagree ( ) Disagree ( ) Mildly Disagree ( ) Neutral ( ) Mildly Agree ( ) Agree ( ) Strongly Agree

50) It usually bothers me when online companies ask me for personal information.*

( ) Strongly Disagree ( ) Disagree ( ) Mildly Disagree ( ) Neutral ( ) Mildly Agree ( ) Agree ( ) Strongly Agree

51) When online companies ask me for personal information, I sometimes think twice before providing it. *

( ) Strongly Disagree ( ) Disagree ( ) Mildly Disagree ( ) Neutral ( ) Mildly Agree ( ) Agree ( ) Strongly Agree

52) It bothers me to give personal information to so many online companies.*

( ) Strongly Disagree ( ) Disagree ( ) Mildly Disagree ( ) Neutral ( ) Mildly Agree ( ) Agree ( ) Strongly Agree

53) I'm concerned that online companies are collecting too much personal information about me.*

( ) Strongly Disagree ( ) Disagree ( ) Mildly Disagree ( ) Neutral ( ) Mildly Agree ( ) Agree ( ) Strongly Agree

---

Addressing Physical Safety, Security, and Privacy for People with Visual Impairments

Tousif Ahmed Patrick Shaffer Kay Connelly David Crandall Apu Kapadia
School of Informatics and Computing
Indiana University
Bloomington, IN, USA
{touahmed, patshaff, connelly, djcran, kapadia}@indiana.edu

ABSTRACT

People with visual impairments face a variety of obstacles in their daily lives. Recent work has identified specific *physical privacy* concerns of this population and explored how emerging technology, such as wearable devices, could help. In this study we investigated their *physical safety and security* concerns and behaviors by conducting interviews (N=19) with participants who have visual impairments in the greater San Francisco metropolitan area. Our participants' detailed accounts shed light on (1) the safety and security concerns of people with visual impairments in urban environments (such as feared and real instances of assault); (2) their behaviors for protecting physical safety (such as avoidance and mitigation strategies); and (3) refined design considerations for future assistive wearable devices that could enhance their awareness of surrounding threats.

1. INTRODUCTION

Maintaining privacy, security, and safety in both physical and online domains are major challenges that almost everyone faces. For certain populations, however, these challenges are especially acute. For example, people with visual impairments (ranging from complete blindness to an inability to read a book when wearing corrective lenses [45]) may not be able to perceive their surroundings as easily as sighted people and are thus less able to effectively monitor for potential privacy, security, and safety risks.

Recent work has begun to study the unique concerns of people with visual impairments. Most of this work has focused on privacy and security related to technology, especially in using online services [7, 33, 42]. Other recent work, like that of Ahmed *et al.* [2], has studied this population's privacy concerns in more general settings, including in the physical world (e.g., eavesdropping on conversations). As these concerns and risks are better understood, the next logical step is to develop assistive devices to help people address them, potentially using new technologies like wearable cameras and other sensors.

As a first step in this direction, we initially set out to study design considerations for potential assistive technologies by conducting interviews focusing on key privacy related scenarios identified by past work [2]. However, as we began our interviews, we were surprised to discover that a recurring major theme that nearly every participant mentioned was physical safety and security, whereas Ahmed *et al.*'s study revealed significant privacy concerns but little concern about physical safety and security. We believe this difference arose because our study significantly broadened the target population; while Ahmed *et al.* conducted their interviews in a small, relatively safe college town (Bloomington, IN with a crime index of 229.8 versus the U.S. average of 294.7),[1] ours was conducted in a major metropolitan area (greater San Francisco, with an average crime index of 497.9). Statistically, our participants are probably right to be concerned given that people with visual disabilities in the U.S. have a higher risk of victimization than the overall population (17.8 versus 14.0 per 1,000 as of 2013) [25].

In this paper, we report on this broader understanding of the safety, security, and privacy concerns of people with visual impairments in an urban context and report design considerations for assistive wearable technology for addressing them.[2] Our findings on physical privacy concerns and behaviors largely confirm previous studies, but they give new insight into the physical safety and security challenges of people with visual impairments. Our findings also shed new light on design considerations for potential technological solutions for all three types of challenges (safety, security, and privacy).

Specifically, we focus on the following three research questions:

R1: *What are the privacy, safety, and security concerns of people with visual impairments in urban environments?* In particular, we seek to identify concerns in contrast or in addition to those expressed in the study of the small college town.

R2: *How do people with visual impairments manage their privacy, safety, and security in urban environments?* We seek to understand the behaviors and coping mechanisms of people with visual impairments in this broader environment.

R3: *How could wearable cameras and sensors address privacy, safety, and security concerns of people with visual impairments?* We aim to identify more detailed design considerations than previous studies through more focused questions on several key scenarios.

[1] http://www.city-data.com/crime/

[2] In our context, we mean 'security' to refer not only to protecting information, but also to physical protection of personal property and spaces. We do not offer a precise definition of the difference between safety *vs.* security, but informally we mean 'physical safety' to refer more to protection from bodily harm (e.g., being assaulted), and 'physical security' to refer to protection from less violent harm (e.g., theft of one's smartphone or ATM passcode).

Symposium on Usable Privacy and Security (SOUPS) 2016, June 22–24, 2016, Denver, Colorado.

To answer these questions, we conducted semi-structured interviews with a diverse group of people with visual impairments (N=19) living in the greater San Francisco metropolitan area, including people with a range of impairments and of different ages. Using scenarios from Ahmed *et al.*, our participants described significant physical safety and security *concerns* not reported earlier, e.g., in public transit, at automated teller machine (ATM) booths, and even in private spaces that sighted people may consider safe. We identified various coping *behaviors* that people with visual impairments currently use to address these concerns, such as avoidance, repositioning, and technology use. The interviews revealed several new and refined design *considerations* for assistive devices that could provide alternatives for addressing the behaviors. For instance, a majority of our participants described wanting to know about the presence and intentions of other people in their immediate physical vicinity, as well as an ability to collect forensic evidence (e.g., imagery to share with law enforcement) of a physical assault.

2. BACKGROUND

Before describing work related to our study, we begin by introducing background related to visual impairments in general, including key terminology and a brief overview of existing assistive devices.

2.1 Key Terminology

The estimated 285 million people living with visual impairments worldwide experience a variety of difficulties with their sense of sight [46]. Clinically, 'visual impairment' is defined as a "visual acuity of 20/70 or worse in the better eye with best correction, or a total field loss of 140 degrees" [4]. 'Severe visual impairment' usually implies a corrected visually acuity of 20/200 or worse. 'Low vision' is sight "that may be severe enough to hinder an individual's ability to complete daily activities such as reading, cooking, or walking outside safely, while still retaining some degree of usable vision" [4]. Finally, 'total blindness' describes a person's inability to see anything with either eye.

Visual impairments also come in a variety of forms. The most common causes of visual impairment stem from the inability to correct refractive issues (43% of cases) and diseases including cataracts (33%) and glaucoma (2%) [46]. Other cases are caused by accidents, other diseases, or a reduction in vision or vision processing such as the loss of central vision, peripheral vision, contrast sensitivity, or depth perception [3]. Only about 15% of people with visual impairments are totally blind, while the majority (65%) of people are over age 65 and live in developing countries (90%) [46].

2.2 Current Assistive Technology

There are many assistive technologies currently available to aid people with visual impairments in their daily activities. Hersh and Johnson [27] provide a comprehensive discussion of these technologies, e.g., for tasks such as personal care (e.g., Braille labels for clothing[3]), reading (e.g., with video magnifiers [5]), navigation (e.g., Miniguide[4]), financial management (e.g., Note Teller[5] for currency detection), healthcare monitoring (e.g., talking bathroom scales[6]), and food preparation (e.g., talking microwave ovens[7]).

Smartphones and PCs are popular with people with visual impairments [37] in part because they have helped them achieve greater independence [14], but they introduce their own challenges since modern visual mouse and touch based user interfaces are often not accessible to people with visual impairments. Blind people generally use audio screen reading software, such as JAWS (Job Access with Speech),[8] Window-Eyes,[9] and VoiceOver,[10] all of which generate synthesized speech to relay information from the screen. People with low vision often use screen magnifying software, such as ZoomText[11] and MAGic,[12] to enlarge a part of the screen to make it more readable. Some people use refreshable Braille displays, although the use of this technology is becoming less common because the number of people who read Braille is decreasing [44] (e.g., only 10% of blind children are learning Braille [43]).

The ubiquity of smartphones and other portable computing devices has motivated research into more advanced assistive devices that can better help people sense their environment for tasks such as identifying and finding objects [8], taking photographs [24, 31], and navigating new spaces [16] and transportation networks [6, 13]. Although some of this work has explored using automated computer vision techniques, other projects such as VizWiz [8] and GoBraille [6] leverage crowdsourcing where remote users view photos taken by people with visual impairments and help identify content in the scene. Crowdsourcing has also been applied to let people with visual impairments take photos and share them on social media [53, 55]. Recently, assistive technology research has shifted towards wearable devices [22, 51] as wearable cameras are becoming more affordable and practical in the form of Google Glass,[13] Orcam,[14] and Narrative Clip[15] [56, 58].

3. RELATED WORK

We now summarize research work related to ours, specifically in better understanding the concerns, coping behaviors, and potential solutions for the security, safety, and privacy of people with visual impairments.

3.1 Privacy, Security, and Safety Concerns

While certain types of concerns like online security have been studied extensively, the physical safety and security concerns of people with visual impairments has not yet been adequately researched. Shinohara and Wobbrock [52] study how assistive devices may attract unwanted attention from friends and colleagues, possibly making users even more conspicuous to potential attackers. Azenkot *et al.* [6] report on the safety concerns of blind and deaf participants in unfamiliar locations in the context of designing a navigational tool. Cassidy *et al.* [15] design a haptic feedback mechanism for using ATMs in order to make assistive devices less obvious to potential attackers. Ahmed *et al.* [2] focus on the privacy concerns of people with visual impairments but also mention safety to the extent that feeling "safe from intrusion" in the home is an important aspect of privacy. Our work focuses on better understanding physical safety and security concerns, and adds significantly to this existing body of knowledge.

Recent work has also addressed the privacy and security concerns of people with visual impairments in the context of electronic device use [7, 20, 33, 42, 56], but again, it has not focused on security and safety in the physical world. Ahmed *et al.* [2] report on privacy

[3] www.labelsp.com/braille
[4] www.gdp-research.com.au
[5] www.brytech.com/noteteller/
[6] www.maxiaids.com/talking-bathroom-scale
[7] www.maxiaids.com/talking-microwave-oven
[8] www.freedomscientific.com/JAWS
[9] www.gwmicro.com/window-eyes/
[10] www.apple.com/accessibility/osx/voiceover/
[11] www.zoomtext.com
[12] www.freedomscientific.com/MAGic
[13] www.google.com/glass/start/
[14] www.orcam.com
[15] getnarrative.com

concerns expressed by people with visual impairments in both the virtual and physical worlds, but their study did not reveal significant concerns related to physical safety and security, which we believe to be an artifact of the fact that their participants all lived in a small, safe college town. In our work, we confirm their findings related to privacy and shed new light on physical safety and security issues, which our urban participants identified as their key concerns.

3.2 Coping Mechanisms

Caine [9] reports three categories of privacy behaviors across technology and age groups including 'avoidance', 'modification', and 'alleviatory'. Our study found evidence of these among our population in addressing not only privacy but also security and safety concerns. We found the 'avoidance' and 'modification' behaviors to be especially prominent, but our study also identifies several additional behaviors, such as 'adaptation' and 'acceptance', that occur specifically because of our participants' visual impairments (see Section 6). In addition, we further categorize the modification coping behaviors ('repositioning', 'mitigation', and 'human assistance') because of their prevalence and importance with a visually impaired population. We did not find any current coping behaviors that would fall under Caine's 'alleviatory' classification. This is likely due to our participants' inability to know if they had been victims of certain behaviors (eavesdropping) and inability to easily identify perpetrators of other crimes (assault).

Both Ahmed *et al.* [2] and Azenkot *et al.* [7] discuss strategies used by people with visual impairments to protect themselves from other people eavesdropping on their devices, including using headphones and screen occlusion software. We report similar defensive strategies but go beyond behaviors related to eavesdropping and report the coping strategies that people with visual impairments use to address privacy, security, and safety concerns.

3.3 Proposed Solutions

Several researchers have addressed the safety concerns of people with visual impairments in the context of navigation and transportation, especially through using mobile and wearable devices. Both Azenkot *et al.* [6] and Campbell *et al.* [13] introduce mobile device applications that provide information about buses and bus stops. Some researchers have addressed the navigational concerns of people with visual impairments through tools that can detect obstacles [18], help cane users with a wearable camera [22], and provide haptic feedback through a wristband [56] among others [17, 32]. We explore these safety concerns as well as others beyond navigational and transportation safety, although our work may also shed light on the design of such devices in the context of physical safety.

Other related work has explored using cameras to help people better monitor their surroundings. Wang *et al.* [54] consider how to alert sighted people who may be distracted by their mobile phones of potentially dangerous situations (e.g., while crossing the street). Abboud *et al.* [1] use sensory substitution device (SSD) cameras in their 'EyeMusic' prototype to convey an image in the form of music. Our work is complementary, and our findings could inform the future design of these devices.

4. METHOD

We interviewed visually impaired participants in an urban setting to investigate their physical safety and security concerns and behaviors, and to understand considerations for addressing their concerns through wearable technologies. The interviews were semi-structured and conducted either in person or by phone, and were conducted individually except when two participants were living with each other and consented to a joint interview (more information is provided in Section 4.3). We included both participants in the same interview in these cases because they were often able to improve their partner's recall of concerns and experiences. Participants were allowed to choose the location of the interview, including the option to be interviewed over the phone.

4.1 Interview Protocol

Our interviews consisted of two parts. First, we presented three hypothetical scenarios derived from the findings in Ahmed *et al.* [2] (in which a person with a visual impairment may experience security and privacy concerns because of people around them). We then introduced potential technological solutions to gather participant feedback and to inform future design choices.

Privacy and Safety Scenarios

We framed our interview discussion around three scenarios related to physical safety, security, and privacy [2]: (1) sharing health history at a doctor's office, (2) reading email in a public place, and (3) typing a password into a computing device or a PIN into an ATM. During the first several interviews, participants reported much greater concern about entering ATM PINs than about entering passwords on a personal computer and, in particular, on the safety aspects of ATM use (e.g., physical assault while withdrawing cash). We therefore tailored the third scenario to consider only ATMs in subsequent interviews in order to obtain more insight into physical safety concerns.

For some interviews, we skipped one or more scenarios depending on the specific impairment of our participants. For example, one participant was able to see nearby people, so we skipped the ATM scenario in her case. Some participants mentioned that they kept their computer screens off during use, so we did not ask them any further questions for the reading email scenario.

Deriving Design Considerations

We next interviewed participants about potential technological solutions to address their concerns. Ahmed *et al.* [2] presented several technology ideas that may possibly address the privacy concerns of people with visual impairments but discussed that we need further research to understand their requirements. Our goal here was to better understand the design considerations for such technologies. We specifically focused on camera based and wearable devices since progress in lightweight, low-cost mobile technology and computer vision has shown promising potential to assist people with visual impairments [2, 56]. Of course, any real-world devices will have to strike a trade-off between various factors, including accuracy, utility, cost, convenience, weight, and so on. We give additional insight into the preferences and behaviors of our participants, which we aggregate into design considerations, and which may form the input into an eventual functional analysis for more formal design requirements.

Our interviews first discussed the use of cameras to analyze the surroundings and help assess the environment for people with visual impairments. We then asked participants: (1) how such a system might help them; (2) how they would prefer such a system to relay feedback to them; (3) what information about the surroundings they would like to receive; and (4) what devices (e.g., wearable first-person cameras or stationary third-person cameras) would be most suitable. For participants unfamiliar with the concept, we gave a brief introduction to wearable camera technology.

We purposely adapted our interview questions when a participant indicated a strong desire for an assistive device that could enhance their safety. When this occurred, our follow-up questions sought to better understand their safety concerns and how assistive devices might be useful.

4.2 Study Procedure

Recruitment and Enrollment

To recruit participants, we contacted various organizations for people with visual impairments and asked them to distribute our recruitment email to their members and other organizations. Our recruitment process ran from February through August, 2015, although most of the interviews were conducted in July and August, 2015. We also used snowball sampling, by asking our participants to notify others about our study.

Ethical Considerations

Indiana University's institutional review board (IRB) approved our study. To obtain informed consent, we provided our information sheet via email so that participants could use accessibility tools to read the study sheet; we read the information aloud if needed. Participants could skip any question, and we recorded interviews only after obtaining verbal or written consent.

Compensation

In-person participants received $15 in cash (interviews lasted at most 100 minutes), and those participating over the phone received a $15 Amazon.com eGift Card.

4.3 Participants

During the six-month study period, we interviewed a total of 19 participants, including 11 in person and eight over the phone. Table 1 summarizes their demographic information. We categorized our participants into three groups based on the nature of their impairment and their personal history. Congenitally blind participants are denoted by 'T', congenitally low-vision participants are denoted by 'L', and late visually impaired participants are denoted by 'X'. The group included nine men and 10 women and a diverse age range from 18-to-65. A majority of our participants lived in either the greater San Francisco or greater Los Angeles metropolitan areas. Four participants were from small cities of relatively equal crime rates (Sonoma and Santa Rosa, California and Bloomington, Indiana). Our participants included two couples (one of which was married) of which both partners were visually impaired; these participants chose to have their interviews jointly. The interviews lasted between 25 and 100 minutes with most lasting about 35 minutes. In-person participants (N=11) chose where to be interviewed with most (N=6) choosing a public place. Others picked their home (N=3) or office (N=2).

4.4 Analysis Approach

All but one of the interviews were conducted by a single researcher (the other was conducted by a second researcher). The interviews were audio-recorded and transcribed. The transcribed interviews were later analyzed and coded using an iterative coding procedure with open coding where two researchers separately developed a list of concepts based on the interview transcripts [48, 49]. Later they created a codebook by combining the lists of concepts, and re-coded one interview. As the agreement (Cohen's κ=0.38) was not satisfactory on that interview, they again discussed and refined the codes. When agreement reached a satisfactory level (κ=0.79 with SD=0.04) on five re-coded interviews, the researchers divided the rest of the transcripts into two sets and re-coded the interviews based on the refined codebook. The final codebook had seven groups of concepts: 'Safety Concerns', 'Privacy Concerns', 'Feelings', 'Coping Behavior', 'Design Attributes', 'Desired Information', and 'Feedback Preference'.

5. FINDINGS: CONCERNS

In this section, we discuss our findings related to *concerns* of people with visual impairments. In Section 6 we report on their *coping behaviors* before discussing the *design considerations* revealed by our study in Section 7.

As mentioned above, we were surprised to find that physical safety and security were recurring themes of our interviews despite rarely arising in Ahmed *et al.*'s study. We attribute this difference to the fact that their study was conducted in a small, safe town, whereas ours was conducted in a major metropolitan area. We thus begin by describing these new findings related to physical safety and security concerns. We also more briefly discuss our findings related to physical privacy concerns, which largely mirror the findings of past studies, in Section 5.2.

5.1 Physical Safety and Security Concerns

Fifteen participants described at least one scenario in which they were concerned about their safety or security, and eight of these described more than one such scenario. In this section we report on these personal safety and security scenarios, which fell into four main groups: on the street, in public transit, in ATM booths, and in private spaces.

On the Street

Although safety on the street is a universal concern, people with visual impairments are at particular risk because they cannot fully assess their surroundings and cannot always recognize (un)safe situations. Moreover, in the case of an encounter such as assault or theft, they cannot describe the visual characteristics of their assailant to police officers, making them particularly attractive targets. Several of our participants expressed such concerns during our interviews.

One participant (T2) expressed heightened concerns about being followed at night, whereas another (T7) expressed a general sense of helplessness about not being able to assess the safety of her environment:

> *When I go for walks, I have been followed. And so basically because of how society is today, I don't go for walks with my guide dog because I don't know who is around me and I think that is much more debilitating for me than anything that we have discussed. Not knowing my environment, not knowing who is around me and if something happened to me I would not be able to tell anyone.* *(T7)*

Some participants described actual scenarios where such fears were realized. One participant (X6) shared a story about an attacker who tried to steal his guitar after a chase that lasted over five minutes. Another (X1) had been a victim of theft only a few days before the interview. Another participant relayed a story about not being able to flee from an unsafe situation:

> *I was across the street from a shooting once. So, I heard the shots — everybody sort of freaked out, of course. And I looked up and went "those weren't firecrackers, right?"*

ID	Sex	Age Group	Impairment type	History	Technology Usage	Interview Method	Participant's Location	Crime Index†
T1	F	24–30	Totally Blind	Since Birth	iPhone	In person	Oakland, CA	970.6
T2	F	24–30	Blind in one eye, light perception in other	Since Birth	iPhone, Laptop	Phone	San Pablo, CA	426.7
T3	F	30–35	Totally Blind	Since Birth	Windows Phone, Regular and Braille Laptop	Phone	Santa Rosa, CA	193.4
T4	F	30–35	Totally Blind	Since Birth	iPhone, Portable Braille Computer	Phone	Santa Rosa, CA	193.4
T5	F	35–40	Blind with Light perception	Since Birth	iPhone, Laptop	In Person	San Leandro, CA	405.0
T6	F	40–50	Totally Blind	Since Birth	Android	In person	Oakland,CA	970.6
T7	F	50–65	Totally Blind	Since Birth	iPhone	Phone	San Bernardino,CA	554.0
T8	F	50–65	Blind with Light perception, can see shapes	Since Birth	Regular phone, Laptop	In person	Berkeley, CA	387.9
L1	F	18–24	Low Vision	Since Birth	iPhone, Laptop	In person	Bloomington, IN‡	229.8
L2	M	18–24	Low Vision	Since Birth	iPhone, Laptop	In person	Berkeley, CA	387.9
L3	M	30–35	Low Vision	Since Birth	iPhone, Laptop	In person	Oakland, CA	970.6
L4	M	50–65	Low Vision	Since Birth	Smartphone, laptop	Phone	San Leandro, CA	405.0
L5	M	50–65	Low Vision	Since Birth	Regular phone, iPad, Laptop	Phone	San Bernardino, CA	554.0
X1	M	24–30	Low Vision	Last 5 years	iPhone, Laptop	In person	Oakland, CA	970.6
X2	M	30–35	Totally Blind	Last 11 years	iPhone, iPad, Macbook	In person	San Leandro,CA	405.0
X3	F	30–35	Totally Blind	Last 7 years	Android Smartphone, Tablet, Laptop	In person	El Cerrito, CA	377.3
X4	M	40–50	Blind in one eye, low vision in other	Last 3 years	Android Smartphone, Laptop	In person	El Cerrito, CA	377.3
X5	M	40–50	Blind with Light perception, can see shapes	Since childhood	Android Smartphone, Laptop	Phone	San Francisco, CA	487.9
X6	M	50–65	Totally Blind	Since 1963	iPhone	Phone	Sonoma, CA	192.9

†2013 city-data.com crime index (Higher means more crime, U.S. average=294.7)

‡This was the first study interview, conducted in the researchers' home city.

Table 1: Demographic information for our study participants. ID Key: T-congenitally blind; L-low vision; X-late visually impaired

> *And everybody is so freaked out they can't talk to me... I am standing on the corner and trying to figure out: What's going on? (T6)*

Public Transit

Most of our participants were heavily dependent on public transport, which gave rise to several safety concerns. One participant expressed concern about waiting for public transit for extended periods of time:

> *What I would like to see is more public transportation run more frequently because when you have public transportation running more frequently you are not standing out there waiting a long time period for help. Because when the bus comes nobody wants to mess around. They want to catch the bus. But if you are standing outside for a half hour to 45 minutes waiting for a bus, a lot of things can happen. (L5)*

Another participant had experienced suspicious activity while waiting for a van:

> *I was in a waiting spot to get a paratransit van, and somebody came into this area. I thought there was someone there but I wasn't sure, and then someone else came up and said: "Did he do anything?" And I was like "What?!" And so I was right and there was someone there. (T7)*

Another participant expressed similar concerns with handling money at transit stations:

> *Another big one that you need to consider especially in big cities: I am standing at a bus stop and I am about to pull out my wallet to get my bus fare ready. Is it safe to do that? Because there might be no one around and it might be okay or there might be 5 or 10 other people around and the minute you pull out your wallet they are going to pounce... I think that is a big, big one you cannot leave out. It is so important. (T4)*

ATM Booths

Although accessible ATMs enable people with visual impairments to perform banking more easily, the overt nature of using an ATM puts them at risk. This concern was expressed by a majority (N=10) of our participants. Some of them also expressed similar concerns for point-of-sale transactions where PINs must be entered and can be observed by others. One participant emphasized this threat when asked about shoulder surfing in the context of laptops, saying that theft of ATM PINs was a greater concern:

> *When I am at an ATM, like I am entering my PIN, those numbers are huge so I am wondering how to mitigate that somehow. Or if I am like at the cash register and I am buying groceries, because that is more of me putting information out in a public place. For me that is more of a concern — going*

to an ATM, person behind me, going to a grocery store and entering my PIN in. (X1)

Another participant highlighted the fact that many people with visual impairments cannot drive and cannot use drive-up ATMs that offer more security than the walk-up stations on the street:

I don't think it is safe to use ATM. We walk, so I can't get into a car. If I use an ATM to get $ 20, I could walk down the street and get mugged. So why should I go to an ATM showing everybody that I am getting money or if I am making a deposit? (L5)

As Cassidy *et al.* [15] reported, although headphone jacks are available at many ATMs to try to enhance privacy, using headphones can actually put people with visual impairments at greater risk by muffling their hearing and further impairing their ability to sense the surroundings [23, 39, 35]. T7 noted this issue, reporting that they need to be so engaged in the transactions that they tend to lose their focus on the surroundings.

Private Spaces

As noted by Ahmed *et al.* [2] in the context of privacy, people with visual impairments can have heightened safety and security concerns in enclosed spaces, including even in their own homes and offices. The main concern expressed during the interviews was an inability to identify others entering their personal space. At home, the safety risk can be reduced by installing home security systems, but these systems trade off security for convenience, as described by one participant:

I want to know who is coming up to my front door. I hate not knowing that because I feel very vulnerable when people knock at my door at home. We have a home security system on at night but we don't have it on, you know, all the time. That would be horrible to have to unset it to go out and in. We have motion detectors but that hasn't been very optimal either, and I would just like to be warned when somebody is coming up to my porch. (T7)

Although office spaces tend to be safer because of better security and the presence of coworkers, people with visual impairments cannot always rely on their coworkers to announce their presence:

There was one time when I was at my office and there was someone walking around. I assumed it was somebody that needed to be there, but the person refused to identify themselves. And they were kind like of creeping around in the middle of the night and stuff. And I think I knew who it was but they wouldn't tell me and it was kinda creepy. (T5)

Participants mentioned similar concerns arising in other enclosed spaces, including hotel lobbies (T1) and libraries (L1, L2).

5.2 Physical Privacy Concerns

Both Ahmed *et al.* [2] and Azenkot *et al.* [7] reported 'eavesdropping' and 'shoulder surfing' as concerns of people with visual impairments. To further explore these concerns and in order to inform the design of defensive technologies, we gave our participants the above-mentioned three scenarios in which eavesdropping and shoulder surfing concerns may arise. Our findings mostly confirm what was found by Ahmed *et al.* [2], so we provide only a brief summary here.

Eavesdropping Concerns

Ten of our participants reported that they would feel uncomfortable verbally sharing their health history with a staff member in the waiting room of a medical facility out of concern that others in the waiting room may overhear the information. Fourteen of our participants said that they had experienced similar situations. Two other participants mentioned that generally they do not feel uncomfortable in these situations, but it depended on the type of information requested, and that sharing Social Security or bank account numbers would make them feel uncomfortable. Two participants said that they felt embarrassed when they had to share their weight. As one participant put it:

It's not that I have anything to hide but I don't really want everybody in the waiting room thinking 'Oh she has this, or she has that.' It's nobody's business. (T6)

Participants also reported similar concerns while filling out forms at the bank (X4, T8), sharing personal information in an office (T3), having personal conversations with others (X4), or having to share their PIN when needing assistance at an ATM (X4).

Shoulder-Surfing Concerns

In response to our second and third scenarios, most participants (N=13) reported that they have shoulder-surfing concerns while using an ATM, and six reported concerns when using their laptops in public places. Two participants indicated that they are uncomfortable when they send text messages on their smartphones. One participant was a victim of shoulder surfing where her confidential information was stolen by one of her coworkers:

I was at work doing [sic] receptionist, sitting down, and a gentleman, a coworker, stood behind me reading my information and I was on the Internet at the time researching information about the company. And he stands behind me and reads off what I was researching on for the company, which was confidential... I felt a little embarrassed and then I had to talk to my manager about it because he took everything that he saw and basically ran with it and got credit for it and I didn't. (X3)

T7 expressed her concern about shoulder surfing as she has to deal with other people's medical information, and other people are put at risk by her lack of awareness of the actions of people around her. These concerns differed across people depending on their specific type of visual impairment; although people with total blindness may not need to turn their screens on, people with visual impairments often use the screen with text rendered in a large font, making them particularly vulnerable to shoulder surfing.

6. FINDINGS: COPING BEHAVIORS

Our participants reported various strategies to address their safety, security, and privacy concerns. We organized these strategies into seven different categories: 'avoidance', 'repositioning or relocation', 'mitigation', 'use of technology', 'help from an acquaintance', 'adaptation', and 'acceptance'.

Avoidance

Fifteen of our participants reported simply avoiding certain situations as a major coping mechanism. Examples of this behavior ranged from avoiding walking on the street when possible (T7), to avoiding the use of ATMs (L5), to sharing personal health information over the phone before a medical visit in order to avoid having to discuss it in the waiting room:

> *What I often do is that I tell doctor's staff before I even go into the office that I won't do things right in the waiting room. So, either we can do that over the phone so that there is a level of confidentiality that way or pull me into the examination room. (T6)*

A common strategy to avoid shoulder surfing as well as device theft was to not use devices outside of the home. Eight participants said that they try to avoid using their laptops outside of their home. One participant (X4) reported that he was advised not to use his phone outside his home to avoid theft, while Participant X6 turns off his devices (or features) to try to avoid using them excessively:

> *The way I address the concerns is just to refrain from using them. I tend to keep the WiFi disabled [on my iPhone], and I just listen to the music or let it be completely off, you know, where it is on standby mode or my laptop is turned off as I am carrying it around. It is in its container and it's off. And I only use them when I feel safe... When I have reservations about the safety of my behaviors, my default choice is just turn the device off. That way no one can have access to it. Because I am not even really using it. (X6)*

Relocation

Fourteen participants said they typically address their eavesdropping and shoulder-surfing concerns by changing their location. They indicated that if they could sense their environment, they might change their location as needed. One participant with low vision who is able to assess the environment usually moves to the corner of a room when sending text messages:

> *Usually I talk and then stop and go to a corner by myself and send it. Before doing magnification I usually sit somewhere or won't take [the text] right away – I will wait until I am by myself and at the same time put myself back-against-the-wall, so that I am holding my cell phone when I read the text, so that I can see everyone walking around. (L3)*

Repositioning was quite common in medical settings: 10 of the 19 participants reported that they had felt uncomfortable at the doctor's office and asked if they could move to a different room.

Mitigation

Fifteen participants mentioned various mitigation techniques to address safety, security, and privacy concerns. For the health information scenario, L2 reported trying to talk quickly so that only the receptionist could understand him, while some whisper or give their information in a softer voice (L3, T7, and T8), and others lean close to the counter (T1).

Our participants' most common defensive strategy to address shoulder surfing concerns at ATMs was to cover their hands so that others could not see their keypresses (T3, T6, and X5). One participant tries to confuse people who may be shoulder surfing:

> *With the phone password, sometimes I intentionally make mistakes so that my passwords are little bit secured. I will hit the keys, I will hit more keys, I will be hitting delete in rapid succession so that it's not easy to understand what the password was. I will be going back and forth between the actual password and use extra letters and numbers, and hitting delete quickly – eventually the password is put in but if it's a four character password, I actually typed 10 characters including the deletions. (X5)*

When asked how she addresses security concerns about her personal possessions, one participant shared her frustration in finding a solution for her phone, as a specific example:

> *I don't know what's safe and what is not. I don't know when it's safe to pull out my phone or not. I have been trying to figure out protocols so that I can kind of barely pull the speaker part of my phone out of my purse. (T6)*

Most of our participants addressed their shoulder surfing concerns either by turning off the screen or by lowering the laptop lid. However, some low-vision participants struggle with these defensive strategies, as they need to have bright lighting in order to see the keyboard and screen. A common solution for eavesdropping is to use headphones, although these have the disadvantage that they can interfere with the person's ability to monitor the environment. One participant (X2) reported using bone-conduction headphones, which allow him to continue listening to his surroundings as his ears are not obstructed by the headphones.

Help from Others

Ten participants said that they generally seek help from their acquaintances, especially for filling out paperwork (L1, X4), conducting transactions at ATMs (X6), or being aware of their surroundings (L2, L3). However, they reported some frustration on having to rely on friends' availability:

> *In our college campus, I was doing the financial aid information thing with a friend. We have to disclose like tax information or birth date or other information that's needed for financial aid. So, I was trying to discuss that information because someone was helping me fill out the information. We are like in the cafeteria type of setting. That was really uncomfortable trying to do that. That was the person's only time that they had to do that. And I pretty much didn't have a choice. (X4)*

Our interviews revealed that people with complete blindness are often helped by friends who themselves have low vision. Meanwhile, since blind people tend to have a better sense of hearing [23, 39, 35], the low-vision helpers sometimes rely on their blind friend. L3, who often helps guide his blind friend on the street, mentioned:

> *Since I have more friends who are totally blind, I usually am the one who is watching over everything else. A couple of times I had some scenarios with friends, I had friends who are totally blind, I am the one who is seeing them because I*

am guiding them. I am also the one who is usually watching around. Because they can hear, they have good hearing, before I see it they already heard it, but the majority of the time, I usually will be the eyes and they are also the ones who will be the ears. (L3)

Adaptation

Many participants reported developing strategies to use their sense of hearing to assess their environment. Six participants reported that they use their hearing or echo location to sense their surroundings. One participant described this as using his 'facial vision' to prevent shoulder surfing at ATMs and grocery store checkouts:

I will stop typing if anyone comes closer than three or four feet from me if I am in the grocery store. People tend to stand six or eight feet away from me, but if they approach close to me then I will stop my work and ask them what they are doing. I can tell that because they start to bump into me. I have like a territorial bubble around me and I hear people's footsteps and I hear the activities that are going on behind me. If anyone's presence is near then they are blocking the sound that I can hear from behind them. It's my 'facial vision,' they call it, when I hear the echoes. The person's presence blocks the ambient noise. (X6)

Others also described similar types of hearing senses. L3 said that he always tries to feel the situation and if he does not feel it is right to perform some activity, then he does not do it. Both T3 and T6 reported that they can "always" tell what others are doing based on the sounds people make.

Acceptance

Participants reported sometimes feeling that a situation was outside of their control and they had very little choice other than to accept the risks. Nine participants indicated such acceptance, for example, having no other choice than to get help from others:

Whenever we have difficulties we have to call someone in and that invades our privacy. We can't read my mail, don't even know who it is from. Most of the time [automatic scanning] doesn't work. Most of the time if you are trying to read bills, scan doesn't work. It works fine for block text, but if you are trying to read tables or anything like that so you are reading any of your personal material or bills, then no. So our privacy... we don't have any privacy. (T7)

Three participants expressed how they have come to accept their lack of privacy and have to always assume they are being eavesdropped upon. For example, one participant said:

I guess over my lifetime I have developed an assumption that someone is there. I kind of say to myself, "if I walk out my front door someone can hear me." (T6)

Those who feel uncomfortable sharing their health history in a waiting room sometimes have to do so unwillingly. One participant (X5) said that he had to do so in the interest of time.

7. FINDINGS: CONSIDERATIONS FOR DESIGN OF ASSISTIVE DEVICES

As the above findings show, people with visual impairments face considerable challenges in maintaining their physical safety, security, and privacy in everyday life, and they cope with these challenges in a variety of ways. Although some of these coping behaviors are effective and do not affect the quality of life, others (like avoidance and acceptance) either continue to put people at risk, or prevent people from realizing the same opportunities that fully-sighted people can enjoy.

Given these findings, a logical next step is to identify potential technological solutions that could help people with visual impairments better manage their physical safety, security, and privacy in various settings. Mobile or wearable devices could use cameras and other sensors to help perceive the environment around the user and then report information about potential threats nearby. Of course, before trying to design or implement such a system, we need to understand the preferences and requirements of people with visual impairments. To do this we asked our participants for feedback about what they would like to see in such devices, including what capabilities would be most important to them, what types of devices they would prefer, and what the important design considerations would be. We first report on the types of information people would like from such devices, and then on the important design considerations for these devices.

7.1 Desired Information

A key goal of our study was to identify what type of information people would like from an assistive device in order to enhance personal safety, security, and privacy. Our interviews uncovered that most participants were interested in answers to a small set of questions:

How Many People Are in My Vicinity?

Most participants (N=14) thought it would be useful to know how many people were nearby, as this information would help them assess their security and privacy and act accordingly. Although one participant (L4) already uses his hearing (adaptation) to infer the number of people around him, he indicated that higher quality information would help him better identify any suspicious activity. Two participants (X3 and X4) added that this information would help with navigation in general.

How Close Are People to Me?

Most participants (N=13) also wanted information about how close people in their vicinity were to them in order to better assess their surroundings. Several participants used the term "bubble" to mark the territory of their private space and a desire to know when people enter this bubble. The radius of the bubble varied between participants, e.g., 5–10 feet (T4), 5–15 feet (T1), 6 feet (X6), and 10 feet (L3). The size of the bubble depended on the situation, e.g., in public places the bubble was smaller than in private places, while X2 and T6 reported that they wanted to know in general if other people were within earshot.

Who Is in My Vicinity?

Almost all participants (N=18) said it would be useful to know if friends or acquaintances were nearby. This information could help them address their privacy concerns in private spaces or at an ATM by relying on trusted individuals. One participant (L2) specifically

wanted to know this information to prevent shoulder surfing by specific people. Another participant (L4) said this information could help him in the office to differentiate between strangers and trusted coworkers. T7 reported wanting to know the general properties of a person, such as age, gender, and other visual characteristics.

What Are the People in My Vicinity Doing?

Most participants (N=16) mentioned a desire to know what others around them are doing, and especially if anyone is paying attention to them or looking at them. For example, T1 wanted to know if people are being "nosy" and looking at her. X1 mentioned wanting to know if people are holding up a camera to capture or record his ATM interactions, while T5 wanted to know if someone was trying to hear her personal conversations. One participant suggested a way to provide this information:

> *Maybe a lot of people aren't paying attention to me at all. The device could say that you have a person two feet away from you watching TV or texting on their cell phone. (T4)*

Knowing what others are doing can be useful at the bus stop, in the doctor's office, or in public places. One participant (X3) reported that having this information would also help her maintain the privacy of other people, since she could avoid disturbing someone who was busy or engaged in a private activity.

Six participants mentioned that they would like information to help them infer people's *intentions*, since knowing their intentions would help address nearly all of the concerns that were reported by our participants. T3 wanted to know if someone is about to reach toward him, e.g., trying to touch him or trying to steal from him. L2 would like to know if someone is trying to read his texts, and if so, whether they seemed to be doing so on purpose or incidentally (e.g., out of boredom).

Forensic Capture: Who Was Around Me?

The interviews revealed an interesting application of cameras that we did not anticipate: four of the participants indicated a desire to record and preserve a video record of their interactions with other people in order to have evidence when their safety or others' safety was compromised. For example, one participant shared a recent incident on a train where this record would have been helpful:

> *I witnessed a scuffle. I witnessed at least a couple guys beating up a third guy. I went to my house, and I called 911 and said that I heard this. I described it as best I could, but I could not – they want visuals. If I would have a camera on me anyway, I would want control of that camera. If I could just get off the train and I would like to give my phone to the station master and say that here is my camera you can have it. (T6)*

This forensics capability identified in our interviews was always mentioned in the context of the visually impaired person's safety. One participant suggested using a body camera to witness potential tampering with her items when she is separated from them:

> *Every time I go through security I have to take my shoes off and put my computer in like separate bins. It would be helpful to have a camera like that is looking out for me to see, especially looking out for my [belongings] going through the security checkpoint. (T7)*

Finally, one participant had a novel idea regarding the use of cameras and enhancing their personal safety, expressing the desire to know *where* cameras are located:

> *I would like to know as a blind person, when other cameras are about. There are cameras on [public transit], at bus stops, and intersections. I would like to know where those cameras are because, for example, if I thought I was in kind of an icky neighborhood and I need to make a phone call or do something on my phone, if I know there is a camera up ahead at the corner, I would do whatever I did by the camera so that a cop could – if I was robbed – have a chance of figuring out who that person was. I will use those cameras as my friend. (X2)*

7.2 Design Considerations

During our interviews we presented participants with several scenarios that involved the use of camera technology to give them better awareness of their surroundings. They reflected on the technologies, interpreted how they could be used in their everyday lives, and often offered further suggestions and refinements. We performed a bottom-up coding of their responses, which resulted in three categories of design attributes they used to describe these technology preferences: 'Discreet', 'Wearable', and 'Forensic Considerations'. Our identification of these three categories provides some insight into design preferences for a potential system, but we note that this is just a starting point; more rigorous future work is required (including functional decomposition, requirements analysis, and prototyping) to derive formal design requirements.

Discreet

As Shinohara and Wobbrock reported, people with visual impairments do not want to be marked as different, so they prefer less noticeable assistive devices and are particularly sensitive about others' reactions towards them [52]. Similarly, most of our participants (N=12) mentioned they would prefer something discreet as they do not want to look "weird" (L1, L2, X1, T5, T7) or draw attention to themselves (T4, T6). Often the discreet and wearable design considerations were brought up together:

> *I like the idea of having something on your clothes because it is less noticeable... because people will start to wonder why is he wearing this weird eyeglass thing. If you want to do stuff low key, then you do it that way. (L2)*

Although many participants imagined a subtle and discreet device to avoid embarrassment, one particular user equated discretion with safety and security:

> *Clip on camera, something I could clip on my glasses or clip on to my cap or collar. Not too visible because it would make me an easy target to someone who might want to steal my camera. They might try to get my camera and knock me over. (X4)*

In order to maintain discretion, many of the participants expressed a desire for subtle feedback from the system:

> *It would need to be something that is not obvious to sighted people, like an app that would vibrate and not let a sighted person know of the alert. It would be very helpful if the notification was discreet. (T4)*

Wearable

Most participants were already familiar with the concept of head-mounted wearable cameras (e.g., Google Glass and Orcam), and we also gave a brief introduction to wearable devices to further familiarize them with the concept. Most of our participants (N=16) indicated a preference for wearable cameras over other types of devices for a variety of reasons: wearable devices are small and less noticeable (L2, T4, L3), they are more convenient as they do not require deliberate pointing like a smartphone camera (T3), and may require less time to activate compared to other mobile devices (L4). One participant (X2) suggested that the camera could be wearable as an earring and another (L5) suggested a lapel pin which could be attached to coats, shirts, or hats, similar to a broach.

Participants had mixed feelings about head-mounted wearable cameras. Some preferred them since they could be worn like sunglasses (X2) and would not affect their natural movement (T8). But most participants felt that these cameras would be more noticeable than other wearable cameras, and would prefer the more discreet devices.

Forensic Considerations

Some participants gave us specific design considerations about forensic capture, such as maintaining their own control of the camera in order to preserve documentation of an extreme situation (such as assault). In particular, one participant told us:

> *I'd like the notification tone and at that point, maybe when it gives that tone, start taking 30- or 15-second interval pictures of who is around. When the police do decide to help, they ask "oh well you didn't see them," we can't describe them. We'd have these pictures in every five, ten, fifteen or thirty seconds intervals of who is around at that point. (T5)*

By mentioning control of the camera, T5 differentiated the camera state from its normal operating mode supporting privacy awareness as posed by the interview questions. Indeed, the other participants who mentioned a forensics capability also desired a way to explicitly change the camera operating mode, either by a specific request from the user or automatically based on a policy specified by the user ahead of time (e.g., in certain predefined scenarios).

8. DISCUSSION

Our interviews yielded new information about visually impaired people's concerns and behaviors regarding physical safety and security, and confirmed past findings about physical privacy. In addition, the interviews explored their thoughts, perceptions, and preferences toward design concepts involving wearable cameras to enhance physical safety, security, and privacy. These findings represent a first step toward designing new assistive devices, and could provide useful input into future formal requirement processes including functional analysis. We term them "considerations" as they should be considered as user feedback much like use-case feedback available during the design process. Although these considerations are not complete, our study group identified them as major themes that could positively influence any potential design. Of course, as with the design of any new technology, there will be competing requirements, including practical limitations on device size, power usage, and cost, and some of the design considerations expressed by our participants conflict with one another (e.g., discreet but with the capability to accurately sense the whole environment). Nevertheless, our study is a starting point for future, more rigorous design processes.

Safety and Privacy 'Bubble'

Although all people share some concern about their private space, our target population of people with visual impairments were clear that their concerns extended beyond their immediate space (for example, within an arm's reach) to several feet away. Their concerns were largely motivated by wanting to sense the presence and intentions of others around them so that they could take action or modify their behavior to avoid risks to their personal safety, security, and privacy. Our interviews also made clear that participants' privacy concerns were preempted by any safety concerns until the latter were satisfied. However, the design considerations for a 'bubble' to enforce safety also apply to protecting privacy, so the potential exists for assistive technology to satisfy both concerns.

Offering Adequate Coping Mechanisms

In terms of supporting coping strategies, we hope our work could help shed light on how to create technologies that prevent people with visual impairments from having to completely avoid activities or completely accept their risks. Wearable cameras combined with computer vision techniques offer the hope of helping people with visual impairments become more aware of their physical surroundings, including when people enter their security and privacy 'bubble'. Knowing who and how many people are in the vicinity, how close they are, and what they are doing could help people with visual impairments better assess and manage their safety and security. This information combined with the knowledge of the layout of a physical space may allow users to better 'reposition' themselves to avoid shoulder surfing, or to adopt 'mitigation' strategies (such as speaking softly) to avoid eavesdropping.

Feasibility of Assistive Technology

After many years as just a research curiosity, wearable cameras such as GoPro Hero,[16] MeCam,[17] and Narrative Clip are already available on the consumer market. Some of these devices even give a near 360 degree view of the wearers' surroundings [40]. Head-mounted cameras like Orcam, Google Glass, and Microsoft HoloLens[18] are also on or nearing the market and may soon be more mainstream. Cameras that can sense in three dimensions (by measuring or estimating depth information), including Google's Project Tango[19] and dual- and multiple-lens sensors [50, 41], are likely to soon appear on these wearable devices. Low-cost infrared imaging sensors like FLIR One[20] may also be useful to more easily detect and recognize people based on their thermal signatures. All of this new camera technology is progressing rapidly and is likely to significantly improve a device's potential to monitor the area surrounding a user.

Meanwhile, impressive advances in computer vision technology have occurred over the last few years, driven in large part by deep learning [38], which can allow hundreds of objects to be accurately detected in near real-time [47], sometimes rivaling or even outperforming human accuracy [26]. Currently these techniques are computationally intensive and are not easy to implement on low-power, resource-constrained devices like wearable computers, but mobile processors are developing rapidly, and we expect deep learning will become feasible on mobile devices in the next few years. In the meantime, devices could rely on lower-cost, less accurate vision

[16] gopro.com
[17] mecam.me
[18] www.microsoft.com/microsoft-hololens/
[19] www.lenovo.com/projecttango/
[20] www.flir.com/flirone/

algorithms, or could send images off-board to remote cloud computing resources, or some combination of the two. Further work is needed to assess how well assistive devices based on current technology could perform given limitations on cost, weight, and power.

Wearable Cameras for Capturing Forensics

The use of cameras to monitor and record incidents of interest has begun to expand, most notably in the form of police officers wearing body cameras. Our work shows that people with visual impairments are also interested in the forensic collection of imagery to improve their physical safety and security. Kientz *et al.* [34] presented a system called CareLog which allows caregivers of autistic children to "document and analyze specific, unplanned incidents of interest" through the use of a wireless trigger. In the case of CareLog, the video and audio are archived for later review. Our research suggests such systems may be extended for people with visual impairments, as one participant suggested:

> *A device like that, to be honest, I think would help me to be less dependent on sighted people. That would be nice. It would allow me to do more things by myself. (T3)*

Networked Cameras

An interesting design consideration directly linked to forensics is *where* the record is maintained. If wearable cameras record and retain the video locally, then an assailant need only steal the visually impaired person's camera. If the record is preserved separately from the device, e.g., in the cloud, then stealing the camera does not destroy the forensics but raises questions about who may have access to private details captured by the camera. An alternative option might create a live video feed from a visually impaired person's camera to a trusted individual such as a friend or 911 operator, similar to current live-broadcasting smartphone apps like Meerkat,[21] Periscope,[22] and Facebook Live.[23] As one possible design template, LiveSafe's smartphone app for campus safety[24] provides direct connection to campus public safety, audio and video recording, discreet initiation, geo-position reporting, and geo-boundary control (to work only within the campus limits). Duncan *et al.* [21] describe networked cameras that monitor activity in residences to allow trusted agents to monitor elderly persons. Complementary to the concept of forensics is whether knowledge of the presence of a camera could be a suitable deterrent and is another interesting direction for future work.

Safety Risks to the Camera Wearer

Although we hope wearable cameras could enhance safety and security, participants expressed concerns that the devices themselves may draw additional attention and actually increase the risk of assault. The safety risks to the camera wearer need to be key design considerations for a camera based solution. Designing wearable devices to be as discreet as possible, combined with forensic capture capabilities, may help reduce this risk. Additionally, the cost of the sensor may contribute to the risk of theft. Based on input from our participants, assistive devices should be low cost and incorporate features such as store and forward of images, and perhaps technology that renders the device useless if separated from its owner.

[21] meerkatapp.co
[22] www.periscope.tv
[23] live.fb.com
[24] www.livesafemobile.com

Privacy Risks to the Camera Wearer

The privacy implications of wearable cameras to the wearer should be considered. Caine and others explored this subject in the context of senior citizens being monitored in their own homes [10, 12], and we expect similar privacy concerns may apply to people with visual impairments. Hoyle *et al.* [30, 29] study the privacy concerns of people wearing cameras with automatic data collection ('lifelogging') and discuss impression management issues as being a major privacy concern for the wearer. People with visual impairments may find it even more difficult to filter the images captured by such devices, requiring careful thought to where the images are stored and how and with whom they are shared.

The use of cameras also puts people with visual impairments at risk of accidentally sharing images or information with the wrong people, which Caine calls a "misclosure" [11]. One participant (P12) mentioned an embarrassing incident in which her friend accidentally shared a naked photo of herself while using the VizWiz app [8] to try to differentiate between conditioner and shampoo. Such incidents underscore the requirement for the camera and recorded data to be under the review and control of the visually impaired person or their trusted surrogate during normal operations. Alternatively, as computer vision continues to improve, automatic algorithms could be employed to scan for potentially sensitive information in images and alert the user accordingly [36].

Privacy Risks to Bystanders

Given that a system might include an outward-facing camera to detect people within the visually impaired person's 'bubble' of personal space, designers also need to take the privacy of bystanders into consideration. Although a future system may or may not store and forward images, the expectation of bystander privacy must be honored, or at a minimum, managed. Denning *et al.* [19] studied the reactions of bystanders towards wearable augmented reality cameras and proposed several design axes for privacy mediating technologies to respect the privacy of bystanders. Another suitable analogy for this design implication is found in lifelogging. Hoyle *et al.* [28] describe the legal difficulties in conducting user studies involving wearable devices because of bystander expectations of privacy. Taking a proactive approach such as privacy by design could help mitigate the privacy concerns of bystanders. Ye *et al.* [57], for instance, detail the use of privacy by design in lifelogging applications. In the case of storing and forwarding images to facilitate forensic capture, the tension between bystander privacy and preserving the image record would require appropriate attention from designers. For example, the transition to store and forward from merely object detection could imply a system state change that might be indicated to bystanders in some manner such as a flashing light. We leave managing this tension to future work.

Beyond Cameras

We also suggest that cameras could be used in conjunction with other rich sensing modalities. One participant (X1) mentioned the possibility of scanning for nearby cell phone signals to identify who was entering their privacy 'bubble'. It is possible to scan the local area for Bluetooth devices and make camera state decisions and user alerts based on received signal power, reported device type and unique ID. It may also be possible to monitor WiFi traffic to make inferences about the people (such as number and distance) carrying those devices nearby. Input from these non-visual sensors could then be used in combination with camera data, for instance, by turning on the camera when a new person is detected nearby.

Concerns Related to Impairment Types

Our participants had different types of visual impairments, and we grouped them based on their impairment history in order to observe any correlations between impairment type and security, privacy, and safety behaviors or concerns. One trend we observed is that the majority of the safety and security concerns we report were given by the congenitally blind participants, whereas only one low vision participant was extremely concerned about safety. However, given the small number of participants and the fact that our study did not investigate this correlation further (e.g. by asking follow-up questions), this observation would need to be confirmed in a future study. Our interviews also suggest that concerns may be correlated with one's own personal history and experiences. For example, the fact that L5 was the only congenitally low vision participant that was highly concerned about personal safety is likely because he experienced a robbery in the past. It is left to future work to understand any correlations between attitudes towards safety and participant demographics, impairments, and personal experiences.

9. CONCLUSIONS

In order to gain an understanding of the physical safety and security concerns of people with visual impairments, and how technological solutions such as wearable cameras can address such concerns, we conducted semi-structured interviews with 19 participants. Our sample was predominantly urban, represented a wide range of ages and visual impairments, and had a balanced gender distribution. We reported on various *concerns* that people with visual impairments have about their physical safety and security, their *coping mechanisms* to address these concerns, and *desired information and design suggestions* in the context of assistive solutions to address safety and security.

We found that people with visual impairments have significant concerns about their physical safety in the context of crime, as they feel not only vulnerable but also unable to fully assess their environment. People with visual impairments, as a result, must develop several coping mechanisms that range between, and include, the extremes of complete acceptance of risk and the complete avoidance of performing certain activities. In addition to finding wearable cameras as a helpful tool to provide feedback about the environment, our participants indicated that the forensic collection of imagery would be helpful in the case of assault. We hope that the results of this study will help illuminate the unique concerns, behaviors, and needs of people with visual impairments in the context of physical safety and security, and will motivate further research to address their needs.

10. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under awards CNS-1408730 and IIS-1253549. We thank Roberto Hoyle for helping us develop the interview protocol. We especially thank our participants, as well as Barbara Salisbury from the Bloomington Chapter of the American Council for the Blind, Michelle Tyler Lagunas from the Lions Center for the Blind, Erin Foley from the Hatlen Center (a program of Junior Blind), Beth Berenson from the LightHouse for the Blind and Visually Impaired, and Dorothy Lenard from IU Disability Services for Students for helping us recruit participants. Finally, we thank our paper shepherd, Simson Garfinkel, for his feedback and suggestions, and the anonymous reviewers for their helpful comments.

11. REFERENCES

[1] Sami Abboud, Shlomi Hanassy, Shelly Levy-Tzedek, Shachar Maidenbaum, and Amir Amedi. 2014. EyeMusic: Introducing a "visual" colorful experience for the blind using auditory sensory substitution. *issues* 12, 13 (2014), 14. DOI: http://dx.doi.org/10.3233/RNN-130338

[2] Tousif Ahmed, Roberto Hoyle, Kay Connelly, David Crandall, and Apu Kapadia. 2015. Privacy Concerns and Behaviors of People with Visual Impairments. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 3523–3532. DOI: http://dx.doi.org/10.1145/2702123.2702334

[3] American Academy of Ophthalmology. 2010. What is Low Vision? (2010). http://www.aao.org/eye-health/diseases/low-vision

[4] American Foundation for the Blind. 2008. Key Definitions of Statistical Terms. (2008). http://www.afb.org/info/blindness-statistics/key-definitions-of-statistical-terms/25

[5] American Foundation for the Blind. 2016. CCTV/Video Magnifier. (2016). http://www.afb.org/info/living-with-vision-loss/for-job-seekers/careerconnect-virtual-worksites/retail-worksite-for-low-vision-users/cctv-video-magnifier/12345

[6] Shiri Azenkot, Sanjana Prasain, Alan Borning, Emily Fortuna, Richard E. Ladner, and Jacob O. Wobbrock. 2011. Enhancing Independence and Safety for Blind and Deaf-blind Public Transit Riders. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 3247–3256. DOI:http://dx.doi.org/10.1145/1978942.1979424

[7] Shiri Azenkot, Kyle Rector, Richard Ladner, and Jacob Wobbrock. 2012. PassChords: Secure Multi-touch Authentication for Blind People. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '12)*. ACM, New York, NY, USA, 159–166. DOI:http://dx.doi.org/10.1145/2384916.2384945

[8] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. 2010. VizWiz: Nearly Real-time Answers to Visual Questions. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology (UIST '10)*. ACM, New York, NY, USA, 333–342. DOI: http://dx.doi.org/10.1145/1866029.1866080

[9] Kelly Caine. 2009a. *Exploring everyday privacy behaviors and misclosures*. Ph.D. Dissertation. Georgia Institute of Technology. http://hdl.handle.net/1853/31665

[10] Kelly Caine, Selma Sabanovic, and Mary Carter. 2012. The effect of monitoring by cameras and robots on the privacy enhancing behaviors of older adults. In *HRI*, Holly A. Yanco, Aaron Steinfeld, Vanessa Evers, and Odest Chadwicke Jenkins (Eds.). ACM, 343–350. DOI: http://dx.doi.org/10.1145/2157689.2157807

[11] Kelly E. Caine. 2009b. Supporting Privacy by Preventing Misclosure. In *Extended Abstracts on Human Factors in Computing Systems (CHI EA '09)*. ACM, New York, NY, USA, 3145–3148. DOI: http://dx.doi.org/10.1145/1520340.1520448

[12] Kelly E. Caine, Celine Y. Zimmerman, Zachary Schall-Zimmerman, William R. Hazlewood, Alexander C. Sulgrove, L. Jean Camp, Katherine H. Connelly, Lesa L. Huber, and Kalpana Shankar. 2010. DigiSwitch: Design and Evaluation of a Device for Older Adults to Preserve Privacy While Monitoring Health at Home. In *Proceedings of the 1st*

ACM International Health Informatics Symposium (IHI '10). ACM, New York, NY, USA, 153–162. DOI: http://dx.doi.org/10.1145/1882992.1883016

[13] Megan Campbell, Cynthia Bennett, Caitlin Bonnar, and Alan Borning. 2014. Where's My Bus Stop?: Supporting Independence of Blind Transit Riders with StopInfo. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS '14)*. ACM, New York, NY, USA, 11–18. DOI: http://dx.doi.org/10.1145/2661334.2661378

[14] Pete Carey. 2015. Smartphones, apps are liberating the blind and visually impaired. (2015). http://www.mercurynews.com/business/ci_28641561/smartphones-apps-are-liberating-blind-and-visually-impaired

[15] Brendan Cassidy, Gilbert Cockton, and Lynne Coventry. 2013. A Haptic ATM Interface to Assist Visually Impaired Users. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '13)*. ACM, New York, NY, USA, Article 1, 8 pages. DOI:http://dx.doi.org/10.1145/2513383.2513433

[16] Hsuan-Eng Chen, Yi-Ying Lin, Chien-Hsing Chen, and I-Fang Wang. 2015. BlindNavi: A Navigation App for the Visually Impaired Smartphone User. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15)*. ACM, New York, NY, USA, 19–24. DOI: http://dx.doi.org/10.1145/2702613.2726953

[17] James Coughlan and Roberto Manduchi. 2007. Color Targets: Fiducials to Help Visually Impaired People Find Their Way by Camera Phone. *J. Image Video Process.* 2007, 2 (Aug. 2007), 10–10. DOI: http://dx.doi.org/10.1155/2007/96357

[18] D. Dakopoulos and N.G. Bourbakis. 2010. Wearable Obstacle Avoidance Electronic Travel Aids for Blind: A Survey. *IEEE Transactions on Systems, Man, and Cybernetics* 40, 1 (Jan 2010), 25–35. DOI: http://dx.doi.org/10.1109/TSMCC.2009.2021255

[19] Tamara Denning, Zakariya Dehlawi, and Tadayoshi Kohno. 2014. In Situ with Bystanders of Augmented Reality Glasses: Perspectives on Recording and Privacy-mediating Technologies. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 2377–2386. DOI: http://dx.doi.org/10.1145/2556288.2557352

[20] Bryan Dosono, Jordan Hayes, and Yang Wang. 2015. "I'm Stuck!": A Contextual Inquiry of People with Visual Impairments in Authentication. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*. Ottawa, 151–168. https://www.usenix.org/conference/soups2015/proceedings/presentation/dosono

[21] John Duncan, L. Jean Camp, and William R. Hazelwood. 2009. The Portal Monitor: A Privacy-enhanced Event-driven System for Elder Care. In *Proceedings of the 4th International Conference on Persuasive Technology (Persuasive '09)*. ACM, New York, NY, USA, Article 36, 9 pages. DOI:http://dx.doi.org/10.1145/1541948.1541995

[22] Alexander Fiannaca, Ilias Apostolopoulous, and Eelke Folmer. 2014. Headlock: A Wearable Navigation Aid That Helps Blind Cane Users Traverse Large Open Spaces. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS '14)*. ACM, New York, NY, USA, 19–26. DOI: http://dx.doi.org/10.1145/2661334.2661453

[23] Frederic Gougoux, Franco Lepore, Maryse Lassonde, Patrice Voss, Robert J. Zatorre, and Pascal Belin. 2004. Neuropsychology: Pitch discrimination in the early blind. *Nature* 430, 6997 (15 07 2004), 309–309. http://dx.doi.org/10.1038/430309a

[24] Susumu Harada, Daisuke Sato, Dustin W. Adams, Sri Kurniawan, Hironobu Takagi, and Chieko Asakawa. 2013. Accessible Photo Album: Enhancing the Photo Sharing Experience for People with Visual Impairment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 2127–2136. DOI: http://dx.doi.org/10.1145/2470654.2481292

[25] Erika Harrell. 2015. *Crime Against Persons with Disabilities, 2009–2013 - Statistical Tables*. Technical Report. Bureau of Justice Statistics. http://www.bjs.gov/content/pub/pdf/capd0913st.pdf

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *CoRR* abs/1502.01852 (2015). http://arxiv.org/abs/1502.01852

[27] Marion Hersh and Michael A. Johnson. 2008. *Assistive Technology for Visually Impaired and Blind People* (1st ed.). Springer Publishing Company, Incorporated.

[28] Roberto Hoyle, Qatrunnada Ismail, David Crandall, and Apu Kapadia. 2015a. Challenges in Running Wearable Camera-Related User Studies. In *CSCW Workshop: The Future of Networked Privacy: Challenges & Opportunities*.

[29] Roberto Hoyle, Robert Templeman, Denise Anthony, David Crandall, and Apu Kapadia. 2015b. Sensitive Lifelogs: A Privacy Analysis of Photos from Wearable Cameras. In *Proceedings of The ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '15)*. 1645–1648. DOI: http://dx.doi.org/10.1145/2702123.2702183

[30] Roberto Hoyle, Robert Templeman, Steven Armes, Denise Anthony, David Crandall, and Apu Kapadia. 2014. Privacy Behaviors of Lifeloggers using Wearable Cameras. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. 571–582. DOI:http://dx.doi.org/10.1145/2632048.2632079

[31] Chandrika Jayant, Hanjie Ji, Samuel White, and Jeffrey P. Bigham. 2011. Supporting Blind Photography. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '11)*. ACM, New York, NY, USA, 203–210. DOI: http://dx.doi.org/10.1145/2049536.2049573

[32] Alekhya Jonnalagedda, Lucy Pei, Suryansh Saxena, Ming Wu, Byung-Cheol Min, Ermine A Teves, Aaron Steinfeld, and M Bernardine Dias. 2014. *Enhancing the Safety of Visually Impaired Travelers in and around Transit Stations*. Technical Report CMU-RI-TR-14-28. Robotics Institute, Pittsburgh, PA. https://www.ri.cmu.edu/publication_view.html?pub_id=7815

[33] Shaun K. Kane, Chandrika Jayant, Jacob O. Wobbrock, and Richard E. Ladner. 2009. Freedom to Roam: A Study of Mobile Device Adoption and Accessibility for People with Visual and Motor Disabilities. In *Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '09)*. ACM, New York, NY, USA, 115–122. DOI:http://dx.doi.org/10.1145/1639642.1639663

[34] Julie A. Kientz, Gillian R. Hayes, Tracy L. Westeyn, Thad Starner, and Gregory D. Abowd. 2007. Pervasive Computing and Autism: Assisting Caregivers of Children with Special Needs. *IEEE Pervasive Computing* 6, 1 (Jan. 2007), 28–35. DOI:http://dx.doi.org/10.1109/MPRV.2007.18

[35] Andrew J. Kolarik, Silvia Cirstea, Shahina Pardhan, and Brian C.J. Moore. 2014. A summary of research investigating echolocation abilities of blind and sighted humans. *Hearing Research* 310 (2014), 60 – 68. DOI: http://dx.doi.org/10.1016/j.heares.2014.01.010

[36] Mohammed Korayem, Robert Templeman, Dennis Chen, David Crandall, and Apu Kapadia. 2016. Enhancing Lifelogging Privacy by Detecting Screens. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*.

[37] Liat Kornowski. 2012. How the Blind Are Reinventing the iPhone. *The Atlantic* (May 2012).

[38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*. 1097–1105.

[39] N. Lessard, M. Pare, F. Lepore, and M. Lassonde. 1998. Early-blind human subjects localize sound sources better than sighted subjects. *Nature* 395, 6699 (17 09 1998), 278–280. http://dx.doi.org/10.1038/26228

[40] Mandy Mandelstein. 2016. Kodak's New 4K Camera Captures Beautiful 360 Video For the Price of a GoPro. (2016). http://gizmodo.com/kodaks-new-4k-camera-captures-beautiful-360-video-for-t-1751390086

[41] Tim Moynihan. 2015. How This Magical 16-Lens Camera Will Actually Work. (2015). http://www.wired.com/2015/10/light-l16-camera/

[42] Maia Naftali and Leah Findlater. 2014. Accessibility in Context: Understanding the Truly Mobile Experience of Smartphone Users with Motor Impairments. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS '14)*. ACM, New York, NY, USA, 209–216. DOI: http://dx.doi.org/10.1145/2661334.2661372

[43] National Federation of the Blind. 2016a. Braille Readers are Leaders. (2016). https://nfb.org/braille-campaign

[44] National Federation of the Blind. 2016b. Braille Usage. (2016). https://nfb.org/braille-usage-toc

[45] U.S. Department of Veterans Affairs. 2002. *Visual Impairment and Blindness*. Technical Report. Department of Veterans Affairs. http://www.publichealth.va.gov/docs/vhi/visual_impairment.pdf

[46] World Health Organization. 2014. Visual impairment and blindness, fact sheet 282. (2014). http://www.who.int/mediacentre/factsheets/fs282/en/

[47] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. 2015. You Only Look Once: Unified, Real-Time Object Detection. *arXiv* abs/1506.02640 (2015). http://arxiv.org/abs/1506.02640

[48] Marcus Renner and Ellen Taylor-Powell. 2003. *Analyzing Qualitative Data*. Technical Report. University of Wisconsin-Extension., Cooperative Extension, Madison Wisconsin, USA. http://learningstore.uwex.edu/assets/pdfs/g3658-12.pdf

[49] Johnny Saldana. 2009. *The Coding Manual for Qualitative Researchers*. Sage, Los Angeles, California.

[50] Vlad Savov. 2016. Dual-camera phones are the future of mobile photography. (2016). http://www.theverge.com/2016/4/11/11406098/lg-g5-huawei-p9-two-camera-smartphone-trend-apple

[51] Roy Shilkrot, Jochen Huber, Connie Liu, Pattie Maes, and Suranga Chandima Nanayakkara. 2014. A Wearable Text-reading Device for the Visually-impaired. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems (CHI EA '14)*. ACM, New York, NY, USA, 193–194. DOI:http://dx.doi.org/10.1145/2559206.2579520

[52] Kristen Shinohara and Jacob O. Wobbrock. 2011. In the Shadow of Misperception: Assistive Technology Use and Social Interactions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 705–714. DOI: http://dx.doi.org/10.1145/1978942.1979044

[53] Violeta Voykinska, Shiri Azenkot, Shaomei Wu, and Gilly Leshed. 2016. How Blind People Interact with Visual Content on Social Networking Services. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, CSCW 2016, San Francisco, CA, USA, February 27 - March 2, 2016*. 1582–1593. DOI: http://dx.doi.org/10.1145/2818048.2820013

[54] Tianyu Wang, Giuseppe Cardone, Antonio Corradi, Lorenzo Torresani, and Andrew T. Campbell. 2012. WalkSafe: A Pedestrian Safety App for Mobile Phone Users Who Walk and Talk While Crossing Roads. In *Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications (HotMobile '12)*. ACM, New York, NY, USA, Article 5, 6 pages. DOI: http://dx.doi.org/10.1145/2162081.2162089

[55] Shaomei Wu and Lada A. Adamic. 2014. Visually Impaired Users on an Online Social Network. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3133–3142. DOI:http://dx.doi.org/10.1145/2556288.2557415

[56] Hanlu Ye, Meethu Malu, Uran Oh, and Leah Findlater. 2014a. Current and Future Mobile and Wearable Device Use by People with Visual Impairments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3123–3132. DOI:http://dx.doi.org/10.1145/2556288.2557085

[57] Tengqi Ye, Brian Moynagh, Rami Albatal, and Cathal Gurrin. 2014b. Negative FaceBlurring: A Privacy-by-Design Approach to Visual Lifelogging with Google Glass. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*. 2036–2038. DOI: http://dx.doi.org/10.1145/2661829.2661841

[58] Yuhang Zhao, Sarit Szpiro, and Shiri Azenkot. 2015. ForeSee: A Customizable Head-Mounted Vision Enhancement System for People with Low Vision. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS '15)*. ACM, New York, NY, USA, 239–249. DOI: http://dx.doi.org/10.1145/2700648.2809865