Understanding & Controlling User Privacy in Social Media via Exposure

A dissertation submitted towards the degree
Doctor of Engineering
of the Faculty of Mathematics and Computer Science
of Saarland University

by

Mainack Mondal

Saarbrücken
October 2017

Date of Colloquium:	20^{th} November, 2017
Dean of Faculty:	Univ.-Prof. Dr. Frank-Olaf Schreyer
Chair of the Committee:	Prof. Dr. Christian Rossow
Reporters	
First Reviewer:	Prof. Dr. Krishna P. Gummadi
Second Reviewer:	Prof. Dr. Peter Druschel
Third Reviewer:	Prof. Dr. Blase Ur
Academic Assistant:	Dr. Rijurekha Sen

ABSTRACT

The recent popularity of Online Social Media sites (OSM) like Facebook and Twitter have led to a renewed discussion about user privacy. In fact, numerous recent news reports and research studies on user privacy stress the OSM users' urgent need for better privacy control mechanisms. Thus, today, a key research question is: how do we provide improved privacy protection to OSM users for their social content? In this thesis, we propose a systematic approach to address this question.

We start with the access control model, the dominant privacy model in OSMs today. We show that, while useful, the access control model does not capture many theoretical and practical aspects of privacy. Thus, we propose a new model, which we term the exposure control model. We define exposure for a piece of content as the set of people who actually view the content. We demonstrate that our model is a significant improvement over access control to capture users' privacy requirements. Next, we investigate the effectiveness of our model to protect users' privacy in three real world scenarios: (1) Understanding and controlling exposure using social access control lists (SACLs) (2) Controlling exposure by limiting large-scale social data aggregators and (3) Understanding and controlling longitudinal exposure in OSMs, i.e., how users control exposure of their old OSM content. We show that, in each of these cases, the exposure control-based approach helps us to design improved privacy control mechanisms.

KURZDARSTELLUNG

Die Popularität von sozialen Netzwerken (SN), wie Facebook, haben zu einer erneuten Diskussion über die Privatsphäre geführt. Wissenschaftliche Publikationen untersuchen die Privatsphäre und zeigen wie dringend SN Benutzer besseren Datenschutz benoötigen. Eine zentrale Herausforderung für in diesem Bereich ist: Wie kann der Schutz der Privatsphäre von SN Benutzern und ihren Inhalten garantiert werden? Diese Doktorarbeit schlägt Ansätze vor, die diese Frage beantworten.

Wir untersuchen das Privatsphäremodel, das Access Control Modell, in SN. Wir zeigen auf, dass das Access Control Modell theoretische und praktische Aspekte der Privatsphäre nicht erfasst. Deshalb schlagen wir das Expositionssteuerunsgmodell vor und definieren Exposition für einen Inhalt als die Menge der Personen, die einen Beitrag ansieht. Unser Modell stellt eine bedeutende Verbesserung zu dem Access Control Modell dar. Wir untersuchen die Effektivität unseres Modells, indem wir den Datenschutz der Benutzer in drei realen Szenarien schützen: (1) Verständnis und Steuerung der Exposition von Inhalten mit Sozialen Access Control Listen (SACLs), (2) Steuerung der Exposition durch Begrenzung der umfassenden sozialen Datenaggregation und (3) Verständnis und Steuerung von Langzeitexposition in SN, z.B. wie Benutzer Exposition alter Inhalte begrenzen. In diesen Fällen fürt Expositionssteuerungsmethoden zu einem verbesserten Privatsphäresteuerungsmechanismus.

PUBLICATIONS

Parts of this thesis have appeared in the following publications.

- "Longitudinal Privacy Management in Social Media: The Need for Better Controls". Mainack Mondal, Johnnatan Messias, Saptarshi Ghosh, Krishna P. Gummadi and Aniket Kate. In *IEEE Internet Computing* vol. 21, no. 3, pp. 48-55, May-June 2017.
- "Forgetting in Social Media: Understanding and Controlling Longitudinal Exposure of Socially Shared Data". Mainack Mondal, Johnnatan Messias, Saptarshi Ghosh, Krishna P. Gummadi and Aniket Kate. In *Proceedings of the 12th Symposium on Usable Privacy and Security (SOUPS)*, Denver, CO, USA, June 2016.
- "Understanding and Specifying Social Access Control Lists". Mainack Mondal, Yabing Liu, Bimal Viswanath, Krishna P. Gummadi and Alan Mislove. In *Proceedings of the 10th Symposium On Usable Privacy and Security (SOUPS)*, Menlo Park, CA, USA, July 2014. **Distinguished Paper Award**.
- "Beyond Access Control: Managing Online Privacy via Exposure". Mainack Mondal, Peter Druschel, Krishna P. Gummadi. and Alan Mislove. In *Proceedings of the Workshop on Usable Security (USEC)*, San Diego, CA, USA, February 2014.
- "Defending against large-scale crawls in online social networks". Mainack Mondal, Bimal Viswanath, Allen Clement, Peter Druschel, Krishna P. Gummadi, Alan Mislove and Ansley Post. In *Proceedings of the 8th International Conference on emerging Networking EXperiments and Technologies (CoNEXT)*, Nice, France, December 2012.

Additional publications during doctoral studies.

- "A Measurement Study of Hateful Messages in Social Media". Mainack Mondal, Leandro Araújo Silva and Fabrício Benevenuto. In *Proceedings of the 28th ACM*

Conference on Hypertext and Social Media (ACM Hypertext), Prague, Czech Republic, July 2017.

- "The Many Shades of Anonymity: Characterizing Anonymous Social Media Content". Denzil Correa, Leandro Araújo Silva, Mainack Mondal, Fabrício Benevenuto and Krishna P. Gummadi. In *Proceedings of The 9th International AAAI Conference on Weblogs and Social Media (ICWSM)*, Oxford, UK, May 2015.
- "Deep Twitter Diving: Exploring Topical Groups in Microblogs at Scale". Parantapa Bhattacharya, Saptarshi Ghosh, Juhi Kulshrestha, Mainack Mondal, Muhammad Bilal Zafar, Niloy Ganguly, and Krishna P. Gummadi. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*, Baltimore, MD, USA, February 2014.
- "Simplifying Friendlist Management" Yabing Liu, Bimal Viswanath, Mainack Mondal, Krishna P. Gummadi, and Alan Mislove. In *Proceedings of the 21st International World Wide Web Conference (WWW)*, Lyon, France, April 2012. **(Demo Paper)**
- "Canal: Scaling Social Network-based Sybil Tolerance Schemes". Bimal Viswanath, Mainack Mondal, Krishna P. Gummadi, Alan Mislove, and Ansley Post. In *Proceedings of the 7th European Conference on Computer Systems (EuroSys)*, Bern, Switzerland, April 2012.
- "Exploring the Design Space of Social Network-based Sybil Defenses". Bimal Viswanath, Mainack Mondal, Allen Clement, Peter Druschel, Krishna P. Gummadi, Alan Mislove, and Ansley Post. In *Proceedings of the 4th International Conference on Communication Systems and Networks (COMSNETS)*, Bangalore, India, January 2012. **Invited Paper**.

Dedicated to the journey so far and to the people who made it possible...

ACKNOWLEDGEMENTS

This part of my thesis should have been the easiest to write, but surprisingly I find it quite difficult to list everybody who helped me in their own unique way (both technically and non-technically) during ups and downs of graduate life. But, nonetheless, I will try; I might miss some and to them, I express my most sincere gratitude for helping me out during my hours of need.

First of all, a big thanks to my advisor, Krishna Gummadi. I am grateful to him for providing me the opportunity to start my academic journey, mentoring me and not giving up on me. Krishna consistently inspired me with his amazing insights and ideas. I learned a great deal from him about good research and about traits of a good researcher. I am thankful to Bimal Viswanath and Alan Mislove for fantastic work collaboration experience, without their help this thesis would not have been possible. I am grateful to Balachander Krishnamurthy for providing me the opportunity to spend a summer at AT&T Labs Research and learn a lot. I am also grateful to Peter Druschel for his valuable guidance and support.

I have been very fortunate to collaborate with several fantastic researchers during my graduate career. I would like to thank my co-authors, Fabricio Benevenuto, Allen Clement, Niloy Ganguly, Aniket Kate, Ansley Post and Ingmer Weber. I specially want to mention the collaboration experience with Saptarshi Ghosh and Denzil Correa. I learned a lot from both of them and it was an absolute pleasure to brainstorm during our long discussion sessions.

I am grateful to all the students who worked with me on several exciting projects. I would specially like to thank Bilal and Yabing. I am also fortunate to have the opportunity to mentor some very talented students: Johnnatan and Leandro. The MPI-SWS socsys group members helped me over the years through thick and thin. I am eternally grateful for their support. I specifically want to thank Juhi and Bilal for the numerous technical and

non-technical discussions while putting things into perspective. Furthermore, I would also like to thank current and former members in the various research groups at MPI-SWS, in particular, Riju, Przemek, Reza, Abhijnan, Paarijaat, Ekin, Reinhard, Eslam, Viktor, Arpan, Cheng, Aastha, Reinhard, Felipe, David, Vineet, Manohar, Ezgi, Thanos, Nina and Till for making MPI-SWS a fantastic place to work. I would also like to thank Rose for her fantastic English communication classes and her help with proof reading the papers and this thesis.

I am especially thankful to Alexander and Anjo; Alexander, for his support, encouragement, friendship and wisdom during the ups and downs of graduate school and my life. Anjo for being an awesome friend and my guide to German culture forever.

A special thanks to the technical stuff at MPI-SWS, and in particular to Carina and Chris for their fantastic efforts and technical know-hows. I would like to thank Brigitta Hansen, Claudia Richter and Annika Meiser for their help with administrative matters as well as in general making my time in Germany significantly easier. Also, thanks to Mary-Lou for providing all the information regarding submitting the dissertation and scheduling the defense.

I would like to thank my family for their support. Without them I could not have accomplished this work. Finally, I am deeply thankful to Junita, my wife, for all her love and support. Kudos to her for not losing hope, sticking with me through the darkest of times and help me to keep my sanity. I could not have completed this work without her help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Civilization is the progress toward a society of privacy. The savage's whole existence is public, ruled by the laws of his tribe. Civilization is the process of setting man free from men.

—Ayn Rand, *The Fountainhead*

CHAPTER 1

Introduction

1.1 The fundamental shift in content sharing due to Online Social Media sites (OSM)

Online social media sites (OSM) are systems like Facebook or Twitter where users share their personal content (e.g., photos, status updates, and social contacts). These systems have become hugely popular in the last few years. In fact, these OSMs have become the de facto portal to access the internet for millions of users. According to a recent Nielson survey [142], an average 35 to 49 years old OSM user visits these sites 7 hours a week and a heavy user visits these OSMs 3 hours per day.

Since OSMs have made it easy for any user to create, publish, distribute and consume content, the amount of data shared on OSMs has also grown massively since the inception of OSMs. For instance, Facebook users uploaded 4.75 billion pieces of content daily (on average) in May 2013 [54]. Much of the personal user generated content shared in OSMs—which we call *social content* in this thesis—is intended to have a limited audience. For example, a user's vacation photos uploaded on Facebook are often intended to be viewed primarily by family and close friends.

However, the huge popularity of these sites (which facilitate the sharing of social content) has also resulted in a fundamental shift in the patterns of content exchange in the online world. Previously, a small number of organizations (e.g., newspapers, companies,

governments or universities) published content on the web and in general users were *content consumers*. The content was mostly public, with open sharing and universal access permissions. In contrast, in OSMs, today, social content is created and shared by end users. The result of this fundamental shift is that, instead of just being *content consumers*, individual end users are now required to be *content managers*. Thus, in addition to creating and uploading social content, OSM users are also expected to manage the privacy of content themselves. Consequently, today OSM users are facing a privacy management crisis.

1.2 Privacy management crisis in OSMs

Today for every piece of data shared in OSMs—every photo, status update, friend request and video—a user must decide which of her friends, group members, and other Facebook users should be able to access the data. Given the per-Facebook-user average of 130 friends and 80 groups in recent years [103]—compounded with the average 90 pieces of content uploaded per Facebook user per month—it is unsurprising that OSM users are in the midst of a privacy management crisis, wherein the task of simply managing access to their content has become a significant mental burden for many users [103].

The privacy management crisis is illustrated by user privacy concerns and anecdotal privacy horror stories covered in numerous popular media and well as in research studies [35, 38, 87, 106]. Researchers have even tried to measure the extent of the privacy management crisis in OSM users. Liu et al. [103] conducted a user survey with 200 users to check the degree to which a user's intended privacy settings match the actual privacy settings of the uploaded content. In the survey, for each user, Liu et al. presented ten random photos the user previously posted on Facebook; then for each photo they ask—"who, ideally, would you like to be able to view and comment on the photo?" The responses from users constitute the intended privacy setting for each photo. Finally, Liu et al. checked the actual privacy settings (programmatically collected using the Facebook API) of those pictures and compare

the actual privacy settings with the users' intended privacy settings (as revealed by the survey). They found that the actual and intended privacy settings match for only 37% of photos in their survey, and when there is a mismatch, the actual privacy settings almost always (in 77% of cases) reveal content to more users than intended. Their survey results further suggested that while poor privacy defaults cause photos to be shared with more users than expected, users who are cognizant enough to modify their default privacy settings still have significant difficulty ensuring their actual privacy settings match their intended privacy settings.

In summary, these results demonstrate that the privacy management crisis is an acute problem for OSM users today, emphasized by the widespread mismatch of user's actual and intended privacy settings [103]. This crisis is further aggravated by the tremendous increase in the user base of OSMs, as well as the continuous increase in the amount of content shared on these sites. Thus, there is a dire need to address this privacy management crisis.

1.3 Challenge: Providing better privacy controls to users

An obvious way to alleviate the privacy management burden for OSM users is to provide them better privacy controls. However, managing privacy is a complex issue and OSM users have different privacy management requirements depending on the *threat model*, i.e., from *whom* she is protecting privacy of her content. More specifically, the privacy management requirements of users can be divided into two broad categories based on the threat model:

1. Managing privacy of social content from OSM operators and their associates: OSM users need to manage data privacy from OSM operators themselves or their associates (e.g., advertisers in OSMs who leverage OSM's infrastructure to serve personalized ads). Prior research investigated multiple dimensions of this requirement: Guha et al. [82] and Baden et al. [23] consider hiding content from OSM operators themselves via encryption. Meng et al. [113] investigated privacy leak for OSM users via mobile in-app advertisements. Goga et

al. [78] even considered privacy leaks via cross correlation of data from multiple OSMs. In other words, this privacy management requirement is an active research topic pursued by other researchers and their research is complementary to the goal of this thesis.

2. Managing privacy of social content from other users: When an OSM user uploads a piece of content, she needs to manage privacy of the uploaded social content from other users in OSM. Failure to do so (e.g., due to setting wrong access permissions) leads to privacy violations. For example, a privacy violation happens when office colleagues of a user view and react to a rant about the workplace (actually intended for close friends) or even when an inappropriate but harmless post from a user is viewed and misinterpreted by strangers and consequently causes serious damage to the user as retribution [139]. In addition to the cases where users wrongly set privacy of their content, failure in managing privacy of social content can happen for multiple other reasons, e.g., strangers digging up a user's old content, automatic aggregation of content from OSMs or even system provided features like Facebook News Feed or Facebook timeline.

In this thesis, we are particularly interested in addressing this privacy management requirement—how to improve the current privacy management mechanisms in OSMs which aim to control the privacy of social content from other OSM users. These mechanisms can be leveraged either by OSM users or can be deployed proactively by OSM operators themselves. Note that it is very unlikely that the current privacy management crisis will be solved automatically by the privacy management mechanisms deployed today (which have been in place for years now). Thus the key research challenge in this space is: *How do we provide improved privacy controls to OSM users for managing privacy of their social content from other users?*

1.4 Thesis research: Understanding & controlling user privacy in OSMs via exposure

The high-level goal of this thesis research is to improve privacy controls in OSMs by developing a systematic, expressive and practical framework. First, we review privacy violations systematically in order to come up with a broad model which can then be used for investigating and improving privacy control. We accomplish this goal by introducing *exposure control*, a model to capture privacy violations in OSMs and in the general online world. Second, we show that our model is expressive enough to capture multiple important privacy violation scenarios from real-world OSMs. We review theoretical privacy definitions as well as anecdotal privacy violation examples from the real world to show that exposure control significantly extends access control, the dominant privacy model in OSMs, in two ways—(i) Exposure control captures more facets of existing theoretical privacy definitions than access control and (ii) Exposure control captures more privacy violations from the real world compared to exposure control. Finally, our model is practical: it is feasible to use exposure control to design and deploy privacy management mechanisms for controlling exposure. We demonstrate the practicality of exposure control in this thesis by investigating three important privacy management scenarios from the real world and proposing improved privacy management mechanisms to better control exposure for each of these scenarios.

In the rest of this chapter, first, we will briefly review the related work in this space and then give a high-level overview of the specific contributions of this thesis.

1.5 Related work

A rapidly growing number of reports in both academia and the popular press emphasizes the problem of managing privacy in popular OSMs. As we mentioned earlier, the reported cases include examples of users wrongly setting the privacy of their content, people digging up

old content, automatic aggregation of content from OSMs and even privacy concerns with system-provided features like the Facebook News Feed or the Facebook timeline. However many of these incidents are anecdotal and from a technical standpoint we need to first define *what is privacy* and how is privacy enforced in OSMs today.

1.5.1 What is privacy?

Privacy is a complex concept which has multiple dimensions. There are extensive discussions, especially in the legal, social and psychology research community, for defining exactly what is privacy. We briefly state the major definitions below; we will expand on each of these definitions in Chapter 2.1.

Warren and Brandeis [168], presented one of the very first definition of privacy in their 1890 law review article. They defined privacy as "right to be let alone" for individuals. However, their definition captures only a basic aspect of privacy and more modern definitions aimed to capture more complicated aspects. Westin's theory of privacy [169], proposed in 1967, captures how people protect their privacy by "temporarily limiting access of others to themselves". Westin defined privacy as "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others." Westin viewed privacy from a legal aspect and other researchers further tried to capture the aspects of privacy from the sociological point of view. Specifically, Altman, a social and an environmental psychologist, presented his definition of privacy [12, 13, 14] using social interactions. According to Altman privacy is a process of boundary control to ensure "the selective control of access to the self". Moreover, he identified that privacy is dialectic and dynamic. In recent times, with the advent of digital media and the networked world, researchers tried to extend Altman's theory. Specifically, around 2002-2003, Petronio (in her CPM theory [128]) viewed privacy as a rule-based mechanism for managing boundaries and Palen and Dourish [127] extended Altman's theory as a tension between three specific boundaries in the networked world.

More recently, Solove presented a data-centric view of privacy and presented a taxonomy of privacy violations [146, 147]. Finally, one of the most recent definition came from Nissenbaum [24, 124]. She presented privacy as the preservation of "contextual integrity".

All of these privacy theories are aimed at understanding what privacy is. A subsequent step is to design privacy preserving mechanisms in data sharing systems like OSMs in order to enforce privacy. Currently, the dominant model for privacy management in OSMs is the *access control* model.

1.5.2 How is privacy enforced in OSMs today?

The aforementioned privacy definitions are intended to systematically define the theory of what privacy is. However, definitions are not enough; a subsequent step is to actually design privacy control mechanisms for OSMs. Today, privacy is primarily enforced by *access control* mechanisms provided by OSM operators. When users upload content, they *a priori* specify who should be allowed to access their content. Options for allowing access can vary from allowing all of the internet users to access a piece of content (e.g., the "public" privacy setting in Facebook) to allowing no one but the uploading user to view the piece of content. We will discuss in detail in Chapter 2.2.1 how access control captures different facets of the theoretical privacy definitions.

Intuitively, OSM users feel that privacy is violated when some unexpected entity actually views a particular content. Today, the access control model requires an OSM user to specify the entities who can view their content. The OSM operator can commonly view all the content. Thus, the access control model captures privacy violation scenarios in which an entity from outside the list of specified entities accesses the content.

Recent work [93, 20, 32] argues through anecdotal examples that access control is insufficient to meet the definition of privacy. boyd et al. [32] presented an example where Facebook users felt their privacy was violated when Facebook launched their News Feed feature. However, the News Feed did not violate any access control policy, but users still

perceived a privacy violation since more people are actually viewing their content due to the News Feed. Facebook quickly reacted [101] after the surfacing of user concerns with News Feed, providing more fine-grained access control mechanisms on Facebook. However, researchers have shown that in spite of the availability of granular access control mechanisms, Facebook users still severely underestimate the number of users who actually view their content [27, 103]. Other work [20] discussed privacy violations in the context of data aggregators, showing that these aggregators were collecting only public data and were not violating any access control. However, these data aggregators violate user privacy. Thus, even though access control is an important and necessary component of privacy management mechanisms, it is not sufficient. To fill this gap, in Chapter 2 we propose and formalize a new privacy model which captures the aforementioned privacy violation scenarios as well as the ones captured by the access control model. We call our model *exposure control*.

1.6 An overview of thesis contributions

In the paragraphs below, we provide a brief overview of our thesis contributions. Our contributions can be divided into two broad parts, First, we propose *exposure control*, an improved and more inclusive privacy model to better capture users' privacy needs.

Then, We investigate three important real-world scenarios where either the users or the OSM operators try to manage privacy of the user generated content. Specifically, in this thesis we investigate controlling exposure using Social Access Control Lists (SACLs), controlling exposure from large scale third party data aggregators and controlling longitudinal exposure. We acknowledge that there might be other venues where exposure control is quite important, e.g., limiting the exposure of content in the face of search in OSMs. We leave investigating how to control exposure in those scenarios for future work.

We start with a brief overview of our exposure control model.

1.6.1 Exposure control: A new privacy model

We start with the observation that access control is the most widely used method by OSM users to ensure the privacy of their content from other users. However, access control cannot capture many of the privacy violations in real-world OSMs as pointed out by the Facebook News Feed example earlier.

We argue that the main reason is in the real world, OSM users often care more about who actually accesses (i.e., views) content, as opposed to just who *can* access the content. Thus, a user might feel their privacy is violated even when a *publicly* shared rant is viewed by her boss. We define the exposure set (or simply *exposure*) of a piece of content as the set of users who eventually view the piece of content.

Intuitively, mechanisms like the Facebook News Feed increase the exposure of content (more people actually viewing uploaded content) and hence users feel their privacy is violated. Thus, in this work we note that one should actually control exposure to manage privacy. Furthermore, we show that exposure control captures more facets of the theoretical definitions of privacy compared to access control—thus establishing that exposure control is an improvement over access control. Our model enables us to investigate scenarios in the real world where privacy management is important. However, exposure control mechanisms should be designed differently than access control. Particularly, the designs should be based on privacy requirements in different real-world scenarios. We individually investigate three such scenarios to design better exposure control mechanisms.

1.6.2 Understanding and controlling exposure using SACLs

Today, a common way for users to control the exposure of posted content is by specifying social access control lists (social ACLs, or SACLs) and allowing only a subset of their social contacts to access a content. The majority of OSM operators today provide fine-grained mechanisms for specifying SACLs, allowing users to restrict their sensitive content to a

select subset of their friends. However, it remains unclear how these SACL mechanisms are used today. In order to design better privacy management tools for users, we need to first understand the usage and complexity of SACLs specified by users. To do so, we conducted the first large-scale study of fine grained privacy preferences of over 1,000 users on Facebook, providing us with the first ground-truth information on how users specify SACLs on an OSM. Overall, we find that a surprisingly large fraction (17.6%) of content is shared with SACLs. However, we also find that the SACL membership (whether a social contact belongs to a SACL) shows little correlation with either profile information or social network structure; as a result, it is difficult to predict the subset of a user's friends likely to appear in a SACL and automate the SACL creation process. On the flip side, we find that SACLs are often reused, suggesting that simply making recent SACLs available to users is likely to significantly reduce the burden of privacy management on users.

1.6.3 Controlling exposure by limiting large-scale social data aggregators

We argue that when a user uploads her social content in OSM, she most likely does not intend her content to be viewed by certain users. A prime example of such unwanted views are the ones coming from large-scale third party data aggregators like Spokeo. In this case, the OSM operators might proactively want to control exposure of user generated content by limiting the large-scale third party data aggregators. In this work, we propose Genie, a system that can be deployed by OSMs operators to defend against large-scale social data aggregators (or *crawlers*) in OSMs.

Genie identifies crawlers by exploiting the fact that the browsing patterns of honest users (i.e., non-crawler users) and crawlers are very different: even a crawler with access to many accounts needs to make many more profile views per account than an honest user, and the crawler needs to view the profiles of users that are more distant in the social network. Genie works by deriving a credit network [50, 75] from the social network (created using

friendship relations in OSM). In brief, credit is associated with links in the social graph, and a viewer must "pay" credits to the viewee, along paths in the social graph, when viewing a profile. Experiments using real-world data gathered from Renren (a Chinese OSM) show that Genie frustrates large-scale crawling while rarely impacting honest users; the few honest users who are affected can recover easily by adding a few friend links.

1.6.4 Understanding and controlling longitudinal exposure in OSMs

We note that the exposure set of a piece of content, i.e., the users who actually view the content, increases with passing time. Thus, controlling longitudinal exposure becomes more challenging with the passage of time. In fact, users' privacy preferences for sharing content are known to evolve over time [25]. 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 (e.g., married) 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 content 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. 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 these residual activities one can recover significant information about the withdrawn tweets or even characteristics of the deleted

accounts. To make users more aware of the flaws in the existing exposure control mechanisms, we designed a Twitter app, deployed at `http://twitter-app.mpi-sws.org/footprint/`, where anyone can login with their Twitter account and check the residual activities around their posts.

We further found that the difficulty of improving longitudinal exposure control mechanisms arises from the fact that there are multiple parties (aside from the user who posts the content) who are affected by longitudinal exposure control mechanisms. For example, a user who reports hate speech posted by some other user might not like to see the hate speech withdrawn since in that way the user who posted the hate speech might evade accountability. We investigate this phenomenon in depth by performing a survey of deletion ecosystem of OSMs. We identify the longitudinal exposure control mechanisms, as well as the actors involved in this ecosystem. Laying the groundwork in this manner, we identify concrete research questions for assessing and improving longitudinal exposure control mechanisms. Finally, we propose and evaluate three longitudinal exposure control mechanisms to augment and improve Twitter's deletion ecosystem.

1.7 Organization

The rest of the thesis is organized as follows:

In Chapter 2, we present the model of exposure control and analyze the merit of exposure control relative to existing privacy theories.

In Chapter 3, we present how users today control exposure by specifying Social Access Control Lists (SACLs) and how we can help these users.

In Chapter 4, we present Genie, a system to control data exposure by limiting large-scale third party data aggregators in OSMs.

In Chapter 5, we present an investigation of longitudinal exposure, i.e, how users control exposure for their old content. We point out one very important challenge towards

improving longitudinal exposure control mechanisms is to understand the nuanced notion of "ownership of content in OSMs" in the deletion ecosystem of OSMs.

In Chapter 6, we present a systematic analysis of the deletion ecosystem, which involves actors other than just the content creator. This systematization lays the groundwork for assessing and improving longitudinal exposure control mechanisms in OSMs. Furthermore, we propose improved longitudinal exposure control mechanisms for Twitter and evaluate their merits and drawbacks.

In Chapter 7, we conclude by summarizing our approach for controlling privacy via exposure control in OSMs and describe directions for future work in this space.

CHAPTER 2

Exposure control: A new privacy model

In order to provide OSM users better privacy management mechanisms for protecting their data from other users, we first need to understand what does *privacy* mean in OSMs. We already hinted at answers to this question in Chapter 1.5.1 by briefly describing seven privacy theories. In this section we will further expand on different aspects of privacy as pointed out by these theories. However this theories present an understanding of "what" privacy is. A subsequent step is to actually build mechanisms for preserving privacy. Thus in this section we review the dominant model for privacy management in OSMs—the access control model. We investigate the aspects of privacy definitions captured by the access control model and point out that while effective, the access control model still does not capture certain crucial aspects of privacy. To that end, we introduce our model of exposure control and examine the usefulness of our model to manage privacy in OSMs. We start this chapter by exploring what we mean by privacy in OSMs.

2.1 What do we mean by privacy in OSM?

Privacy in general is acknowledged to be a complex issue which might vary in different scenarios or even from one individual to another. In fact, the word "privacy" is an example of an untranslatable lexeme [17] and many languages do not have a specific word for "privacy" (e.g., Russian). In spite of this complexity (or perhaps encouraged by the complexity), for a long time legal and philosophical scholars have tried to systematically understand and define

privacy. These privacy definitions date back long before the advent of OSMs—Warren and Brandeis were one of the first to systematically define privacy in their influential 1890 law review article "The Right to Privacy" [168] as the "right to be let alone". However, definitions of privacy have evolved over the years and even now researchers are still attempting to understand all the dimensions of privacy and are still presenting new definitions. Thus, we start with a chronological review of the prominent definitions in our context: understanding privacy in OSMs. Note that the goal of our review is not to compare different privacy definitions and theories, but rather to understand the different aspects of privacy captured by these theories.

2.1.1 Privacy as right to be let alone

The first publication in the United states to advocate a right to privacy appeared as an article in Havard law review [168] by Warren and Brandeis in 1890. They argue that the contemporary adaptation of instantaneous photography in society and the widespread circulation of newspapers violated a basic right which was not captured by the law at that time. They pointed out that this unaddressed right of an individual is his "right to privacy".

Warren and Brandeis reviewed the contemporary existing legal rights of individual and checked if they sufficiently captured the right to privacy. They found that the law of slander and libel (forms of defamation) did not capture the right to privacy since those laws "deal only with damage to reputation." In their view, even when some personal information of a person is widely circulated (e.g., via gossip columns in newspapers), defamation law only protects an individual when the individual suffers a direct effect in his or her interaction with other people. They mentioned that "however painful the mental effects upon another of an act, though purely wanton or even malicious, yet if the act itself is otherwise lawful, the suffering inflicted is damnum absque injuria."[1] They also reviewed intellectual property laws and concluded that "the protection afforded to thoughts, sentiments, and emotions, expressed

[1] a loss or harm from something other than a wrongful act and which occasions no legal remedy.

through the medium of writing or of the arts, so far as it consists in preventing publication, is merely an instance of the enforcement of the more general right of the individual to be let alone". Warren and Brandeis further elaborated their idea of the right to privacy as: "where protection has been afforded against wrongful publication, the jurisdiction has been asserted, not on the ground of property, or at least not wholly on that ground, but upon the ground of an alleged breach of an implied contract or of a trust or confidence".

In the case of OSM content sharing, the definition by Warren and Brandeis captures a basic aspect of privacy requirement of the OSM user. Namely, an OSM user should have the right (and ways) to enforce who can (or cannot) see their content (i.e., letting them be alone). This definition, albeit the first of its kind, is from a different era and does not capture many aspects of privacy which are very important in OSMs. For example, the authors explicitly mentioned that "the right to privacy does not prohibit any publication of matter which is of public or general interest". However, even when information in OSMs (e.g., a Facebook profile picture) is accessible to the public (and needs to be publicly available, e.g., for finding the Facebook profile of a specific user), an OSM user might feel that her privacy is violated if her public profile picture is posted and discussed in an online forum like Reddit. These feelings of privacy violation may specifically arise if the picture is discussed out of context and viewed by people she did not imagine would see/distribute her picture while uploading the picture. Furthermore, according to this definition, the right to privacy ceases upon the publication of the information by the individual. However, as the example above shows (public Facebook profile picture posted in Reddit), even when users post their content publicly, they have an expectation of privacy. Over the past century, scholars have thus subsequently refined the definition of privacy.

2.1.2 Westin's definition of privacy

Alan Westin's theory of privacy [169], proposed in 1967, captures how people protect their privacy by temporarily limiting access to themselves by others. Westin noted that privacy,

in addition to other needs, enables people to adjust emotionally to their daily interactions with others. For Westin privacy is (i) a dynamic process, i.e., people regulate privacy to fulfill their momentary needs and requirements and (ii) a non-monotonic function, since people can have too little, sufficient or too much privacy. Notably, according to Westin's theory, privacy is neither self-sufficient nor an end in itself, but it helps people achieve their overarching end of self-realization.

Specifically, Westin defined privacy as: "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others. [Moreover] ... privacy is the voluntary and temporary withdrawal of a person from the general society through physical or psychological means".

In his theory, Westin proposed four functions or purposes of privacy. These functions essentially try to address the *why* of privacy: *Personal autonomy* is the desire for an individual to avoid manipulation, domination, or being exposed by others. *Emotional release* is the release of an individual from the tensions of social life, e.g., role demands, the management of losses, etc. In other words, privacy (alone or with other means) provides "time out" to an individual from social demands and hence enables emotional release. *Self-evaluation* is the function which concerns integrating experience into meaningful patterns and exerting individuality on events. Examples of such situations are processing information and supporting the planning process (e.g., the timing of disclosures). The final function, *Limited and protected communication*, has two aspects: interpersonal boundaries are created by limiting communication; protected communication fulfillls the desire for sharing personal information with specific trusted individuals.

According to Westin, these four functions are achieved by four key states of privacy. These four states necessarily provide the *how* of privacy in Westin's theory.

1. Solitude: refers to the state of being free from observation by others.

2. Intimacy: refers to the state of limiting communication within a small group to achieve a close, relaxed and frank relationship.

3. Anonymity: refers to the freedom from identification and from surveillance in public.

4. Reserve: refers to the desire to limit information disclosure to others and others recognizing and respecting that desire.

Specifically, Westin's theory suggests that in OSMs, privacy management mechanisms should try to enable OSM users to achieve these key states. Westin further viewed privacy in individual, group, and organizational levels. The privacy states and functions presented in Westin's definition have inspired further research [110]. These investigations into Westin's theory support (to varying degrees) the validity of Westin's definition of privacy (i.e., states and functions) to the current day.

However, Westin's theory did not clearly identify the inter-dependency within his four functions and did not discuss whether those functions are fundamental or some key dimensions of privacy underlie these functions. To that end, some researchers have tried to understand privacy from a different angle—they explored interpersonal communication patterns to understand privacy.

2.1.3 Altman's privacy theory

Altman offered [12, 13, 14] another point of view on the characteristics of privacy, which is different from Westin's theory. Altman is a social and an environmental psychologist and social interaction is at the heart of his theory. According to Altman, privacy is neither static, nor it is a withdrawal process for the individual. Altman presented privacy as the "selective control of access to self" regulated as dialectic and dynamic processes. There are boundaries between privacy and publicity and these boundaries are constantly changing based on an individual's social interactions (thus dialectic processes). Due to these interactions, an individual's boundaries also change over time (thus dynamic processes).

More concretely, Altman pointed out five properties of privacy: (i) privacy involves a dynamic process of interpersonal boundary control. (ii) Altman pointed out that an

individual's desired and actual levels of privacy differ. (iii) privacy is a non-monotonic function; there is an optimal level of privacy (desired level = actual level). However, there can be too much privacy (desired $<$ actual) or too little privacy (desired $>$ actual). (iv) privacy is bi-directional, involving inputs from others (e.g., via a social conversation) and outputs to others (e.g., by sharing a privacy concern with friends). (v) finally, privacy operates at both the individual and group level.

Altman pointed out that there are multiple behavioral mechanisms for regulating privacy (e.g., cultural norms), and those mechanisms operate as a coherent system. Note that Altman viewed privacy as a social process. Thus, his theory pointed that properly understanding privacy will require understanding the interplay of people, the social interactions, the physical environment and also understanding the temporal nature of social phenomena. In fact, he pointed out, privacy has a cultural context, specifically, the psychological aspects of privacy are culture specific.

Altman's theory is one of the prominent theories of privacy today and it provides a fairly general understanding of privacy as social processes. However, a central issue with Altman's theory is that it is not clear if the interpersonal boundary in Altman's theory is a metaphor or a theoretical construct. This issue becomes even more important when researchers want to apply Altman's theory to concrete real-world systems like OSMs. Thus, some recent efforts to understand privacy explore the specific boundaries and their regulations in the real world in order to extend Altman's theory.

2.1.4 Petronio's Communication Privacy Management (CPM) theory

Petronio attempted to extend Altman's dialectical and dynamic conception of privacy (as tensions between opening and closing boundaries) to the realm of digital communications. To that end, she presented her Communication Privacy Management (CPM) theory [128]. CPM theory suggests that individuals regulate their privacy by regulating privacy boundaries that can range from complete openness to complete closeness. An open boundary means

an individual is willing to grant permission to access private information. On the other hand, a closed boundary signifies an individual's desire to conceal private information. The relationships between these boundaries are dialectical (as in Altman's theory), since individuals constantly adapt their level of privacy and disclosure to internal and external state. Moreover, Petronio's theory postulates that individuals achieve their desired levels of privacy via the use of privacy rules. That is, if an individual wishes to change her privacy boundaries, she users a rule-based privacy management system that regulates the degree of boundary permeability (how much information is shared). This rule-based management system allows CPM theory to consider how decisions are made about revealing and concealing private information.

Five propositions underpin the CPM theory. They are stated below:

1. Privacy in CPM theory is defined in terms of ownership; thus, when people believe some information belongs to them, they consider that information private.

2. Since individuals believe that they own their private information, they also believe that they also have the right to control the distribution of that information.

3. People develop and use privacy management rules to control the distribution of information. These privacy rules are influenced by people's cultural backgrounds, their gender, the context of the communication (e.g., telling personal details while asking for help after an accident), motivation (e.g., by reciprocity of information sharing) and based on the assessed risk-benefit trade-off of disclosing information.

4. Once an individual shares private information, a collective privacy boundary is formed and others receiving that information becomes co-owners of that information. From the point of view of the original owner, co-owners share fiduciary responsibilities to jointly control this private information in a way that is consistent with the original owner's rule. Privacy rule co-ordination between the original owner and co-owners revolves

around permeability, co-ownership responsibilities, and linkage rules. Furthermore, these privacy rules also change over time.

5. Finally, when the privacy rules are not coordinated between the original owner and co-owners, there is a possibility of boundary turbulence, which can lead to privacy violation. In other words, boundary turbulence occurs when co-owners fail to effectively manage the flow of information to honor the desire of the other owner.

To summarize, Petronio extends Altman's original proposal of privacy regulation via her CPM theory by concretely pointing out rule-based processes of boundary regulation. This extension can be realized in OSMs by giving users the ability to specify who should be able to access their data. However, other researchers have contested this rule-based depiction of privacy and pointed out that privacy as a concept is too complicated to capture via a set of rules. To that end, they proposed another extension of Altman's theory.

2.1.5 Palen and Dourish's extension of Altman's theory to the networked world

Palen and Dourish [127] contested the extension of Altman's theory by Petronio. They pointed out that "Privacy management is not about setting rules and enforcing them; rather, it is the continual management of boundaries between different spheres of action and degrees of disclosure within those spheres. Boundaries move dynamically as the context changes. These boundaries reflect tensions between conflicting goals; boundaries occur at points of balance and resolution". Palen and Dourish specifically extended Altman's theory to the modern day networked world, where new types of communication technology (e.g., OSMs) are revolutionizing the way people interact and consequently changing our understanding of privacy.

Specifically, Palen and Dourish pointed out that a world driven by information technology brings forth key changes in communication: "The significance of information

technology in this view lies in its ability to disrupt or destabilize the regulation of boundaries. Information technology plays multiple roles. It can form part of the context in which the process of boundary maintenance is conducted; transform the boundaries; be a means of managing boundaries; mediate representations of action across boundaries; and so forth". To that end Palen and Dourish pointed out three boundaries which lie at the core of their theory.

1. **The Disclosure Boundary: Privacy and Publicity** This boundary captures the tension of control of individuals over their information: individuals selectively disclose information to maintain a balance between openness and closeness. Thus, privacy does not only consist of making all information secret, but rather it requires the ability to control which information is disclosed to the public and which is not. For example, academics maintain their public web page to disclose information about their research. However, these same academics might like to keep the details of their personal life out of public view. However, the disclosure boundary is not totally under an individual's control. For example, a friend might post a picture of a recent wild party without the subject's knowledge, or a Google search might reveal an individual's past activities (e.g., Usenet posts), which do not represent how the individual has changed over the years. Thus, aside from disclosure boundary Palen and Dourish identified two more privacy boundaries to address these cases.

2. **The Identity Boundary: Self and Other** The identity boundary captures the social nature of privacy regulation and identifies tensions not in an individual's total control. In any information exchange, there are originators of information and recipients. Thus, the tension at the identity boundary arises from the fact that in a networked world, information can be easily interpreted without context (e.g., via viral sharing of a picture). Consequently, if the recipients interpret a piece of information quite differently than what the original uploader had in mind, the uploader's privacy might be violated.

3. **Temporal Boundaries: Past, Present and Future** Lastly the temporal boundary represents the dynamic nature of privacy. A piece of information disclosed in the past by an individual might wrongly represent the current state of the individual. Thus, Palen and Dourish presented temporal boundaries as—"while past and future interpretations of information disclosure are out of our control, the way in which current privacy management is oriented towards events in the past or future is a matter of active control and management. Our response to situations of disclosure, or our interpretation of information we encounter, is framed and interpreted according to these other events and expectations. We negotiate boundary locations as we act in the world." Hence temporal privacy control in a networked world involves revisiting old content and determining their disclosure boundaries.

Note that both Petronio's CPM theory and Palen and Dourish's theory are extensions of Altman's theory of privacy. However, with the advent of social media and in general novel communication platforms like the internet, other aspects of privacy are also starting to become important. One such example is how to maintain the balance between an individual's privacy and societal necessities like surveillance in public places. Another example would be the case of a Facebook user. Although she is ok with her Facebook profile picture being public, she might not be comfortable with publishing this picture on the first page of the New York times. To address these issues two more recent privacy theories have been proposed.

2.1.6 Solove's taxonomy of privacy

Solove [146, 147] pointed out the broadness and complexity of the idea of privacy and some of the struggles people face while defining and approaching privacy as a legal and social construct. Solove's idea of privacy is data-centric—it revolves around data collection, processing, dissemination and invasion. He uses Wittgenstein's concept of 'family resemblances' as a way to capture the notion of privacy people have in their mind. Solove pointed out that privacy has many meanings; however, just like family resemblances, they are all

related. Solove provided a rich taxonomy of privacy [146] as the core of his theory. He noted different disruptions in each phase (data collection, processing, dissemination and invasion), which are important when identifying privacy violations. The basic structure of his taxonomy is:

1. **Information collection:** surveillance (the watching, listening to, or recording of an individual's activities), interrogation (pressuring of individuals to divulge information).

2. **Information Processing:** aggregation (gathering together information about a person), identification (connecting information to an individual), insecurity (problems caused by the way information is handled and protected), secondary use (use of data for purposes unrelated to the purposes for which the data was initially collected without the data subject's consent), exclusion (failure to provide individuals with notice and input about their records).

3. **Information Dissemination:** breach of confidentiality (breaking a promise to keep a person's information confidential), disclosure (revealing true information about a person to others), exposure (revealing another's nudity, grief, or bodily functions), increased accessibility (amplifying the accessibility of information), blackmail (threat to disclose personal information), appropriation (use of the data subject's identity to serve the aims and interests of another), distortion (the dissemination of false information about a person).

4. **Invasion:** intrusion (invasive acts that disturb one's tranquility or solitude), decisional interference (the government's incursion into the data subject's decisions regarding her private affairs).

Solove uses his taxonomy to explain the reason for privacy violations in different scenarios, e.g., during surveillance. Note that Solove's taxonomy covers the privacy concerns

not only in the context of OSMs, but captures the broader privacy concerns in our data-driven world. However, Solove's taxonomy categories do not explicitly identify the dynamic nature of privacy and the social norms that control privacy.

2.1.7 Nissenbaum's "privacy as contextual integrity"

Contextual integrity (CI) is a theory of privacy developed by Nissenbaum in [24, 124]. CI provides a normative model of privacy, i.e., CI points out what are the normal behaviors that preserve privacy. CI comprises four descriptive claims:

1. Privacy is preserved by *appropriate* flow of information.
2. Appropriate flow of information conforms to contextual information norms.
3. Each contextual information norm consists of five independent parameters: data subject, sender, recipient, information type, and transmission principle.
4. Conceptions of privacy are based on dynamic ethical concerns that evolve over time.

CI provides a framework to argue about violation/preservation of privacy and allows for the evolution as well as alteration of informational norms, often due to novel socio-technical systems like OSMs. The CI framework states that for any given social context (offline or online), there are informational norms defining appropriate flows for various types of information. Privacy violations result when these norms are broken. A prime example of the effectiveness of CI is to explain the phenomenon of "privacy in public". OSM users might upload personal content which is accessible to everyone in the internet (e.g., a Facebook profile picture). However, those users feel their privacy is violated when a third party crawler collects this "public" information and puts it in a torrent [83]. According to CI, the contextual integrity is violated in this case, since it does not conform to the transmission principles (the user did not consent for it to be collected and put in a third party database).

Thus, even if a piece of information is shared with the public, there is some notion of privacy of that information.

2.1.8 Summary

In this chapter, we reviewed seven privacy definitions, proposed at different points in time and capturing different dimensions of privacy. We note that one key aspect of privacy captured in these definitions is the user's expectation/desire of who would see their content and how the content is further re-used. This expectation is personified by Palen and Dourish's "identity boundary" and Nissenbaum's "informational norms." However, all of these definitions and theories are designed to understand *what is privacy*; a subsequent step is to make these definitions actionable by incorporating privacy by design into the systems. Next, we investigate *access control*, the dominant privacy model in current OSMs, and we discuss the extent to which access control captures elements from different definitions of privacy.

2.2 Current model for controlling privacy in OSMs: Access control

In OSMs and in general in computing systems, privacy has typically been accomplished via *access control*, which requires enumerating the users, groups, or roles who are or are not able to access information. The popularity of OSMs has led to a renewed discussion about whether access control is a satisfactory model for user privacy. To that end, first let us rebiew what aspects of privacy are indeed captured by the access control model.

2.2.1 Capturing privacy definitions via the access control model

We reviewed definitions and theories in Chapter 2.1, which deals with understanding different aspects of privacy. Now, we check which elements of those theories, i.e., which dimensions of privacy, are captured by access control and which elements are not captured.

Our application scenario is simple: an OSM user (or *uploader*) is uploading her content to a social media site.[2] In our scenario, a user is concerned about privacy of her own content and consequently about her privacy. In the access control model, she manages privacy by allowing or denying access to others via including/excluding users in access control lists. She can also change the access control of her uploaded content at any point of time in the future. Note that the privacy theories we mentioned earlier are generic; not all of their characterizations of privacy conform to our scope (e.g., violating privacy by interrogating a user). We identify the relevant elements (in our application scenario) of privacy from each theory and review the effectiveness of access control in capturing those elements of privacy. We summarize our review in Table 2.1.

Aspects of Warren and Brandeis's theory [168] captured (and not captured) by access control: Warren and Brandeis defined privacy as the right to be let alone. Given that while uploading a piece of content, users can enumerate other users, groups, or roles who are or are not able to access this information, these uploaders can exercise their right to be let alone. For example, if an OSM user does not want her office colleagues to be able to see her vacation pictures, she can enforce privacy by simply denying her office colleagues access to her vacation pictures. However as we mentioned in Chapter 2.1.1, this theory is the first of many theories and it captures only a basic requirement of OSM privacy management mechanisms. Other more recent theories better capture the complex nature of privacy. Hence, we next focus on those theories.

[2]Note that in this thesis, we are interested in how users can better protect the privacy of their own uploaded content. Thus, investigating how content uploaded by user A violates the privacy of user B (e.g., when user A unwittingly publishes a picture of a homosexual party night and tags user B, who is a local school teacher) is out of the scope of this thesis.

Aspects of Westin's definition of privacy [169] captured (and not captured) by access control: Recall from Chapter 2.1.2 that Westin provided four key states of privacy: solitude (the state of being free from observation by others), intimacy (limiting communication within a small group to achieve a close, relaxed and frank relationship), anonymity (freedom from identification and from surveillance in public) and reserve (desire to limit information disclosure to others, and others recognizing and respecting that desire). An ideal privacy management model should enable users to achieve these key states. Out of these states, anonymity is out of our scope of discussion because anonymity deals with the relationship between user identity and content, whereas we are mainly interested in the privacy of the content itself. Out of the other three states, the access control model captures two: (i) solitude—a user can exclude specific people from accessing their content. (ii) intimacy—when the user wants an intimate interaction, she can allow access to her content for only a small group.

Unfortunately, access control does not enforce reserve. For example, even if a user uploads a photo to Facebook and makes it accessible to everybody in the internet, her desire might be that this content is only meant for identifying her Facebook profile. However, web crawlers who mechanically aggregate and mine public data can just collect this picture and add it to their database of profiles for future background checks by hiring managers. Thus the desire of users of how to use her content cannot be captured and enforced by access control.

Aspects of Altman's privacy theory [12, 13, 14] captured (and not captured) by access control: Altman defined privacy to be "the selective control of access to the self". Chapter 2.1.3 pointed out that according to Altman there are five properties of privacy. Out of these properties, access control captures three properties: (i) Access control enables OSM users to control the dynamic process of interpersonal boundary control, since a user can allow or deny access to her content and can change those access control lists over time. (ii) Users can also maintain the bi-directionality of privacy by changing the access

control settings based on her social interactions. (iii) Finally users can enforce privacy at the individual or group level by allowing or denying users based on individual preference (e.g., show birth date to only close friends) or as a group (e.g., share office party pictures with all office colleagues).

However, access control does not capture the frequent scenario [103] that the access-control settings a user chooses does not match the user's intentions (i.e., the desired privacy) due to the inherent usability difficulties of specifying access-control policies. Moreover, Altman stated that privacy is a 'dynamic' process, whereas access control is static. Because temporal changes in privacy have been observed frequently [25], the access control model does not sufficiently capture this dynamic nature.

Aspects of Petronio's Communication privacy management (CPM) theory [128] captured (and not captured) by access control: Petronio's theory is based on five propositions (Chapter 2.1.4). Out of these propositions the access control model addresses the first three: (i) the access control model acknowledges the ownership of the original uploader since only she (barring the OSM operator) can allow or deny access (ii) Users can specify who can access their content, thus capturing the right to control the distribution of content. (iii) The access control model enables users to specify access control lists for OSM content (i.e., lists of people who are allowed/denies to access the content), thus capturing privacy management rules.

The two remaining propositions, however are not captured by the access control model. First, the access control model does not properly address co-ownership of OSM content, i.e., the access control model enforces that *everybody* that has access to the content are co-owners of that information. However, intuitively, the user might only desire to have only a small subset as co-owners, specifically the people she intends the content for, e.g., a user might only want her Facebook public profile picture to be seen by a few friends; she may not want stalkers to collect her pictures. Second, this loose idea of co-ownership in the access control model can result in boundary turbulence (while not violating access control), when

some non-intended people view the user's content and consequently the flow of information deviates from the user's expectation. In other words, the access control model does not concretely capture the user's desired audience.

Aspects of Palen and Dourish's extension of Altman's theory to the networked world [127] captured (and not captured) by access control: Palen and Dourish discuss three boundaries related to privacy as part of their theory (Chapter 2.1.5). Access control captures two out of these three boundaries: (i) The access control model captures disclosure boundary, since access control enables users to pick and choose the people and disclose the content only to them. (ii) Furthermore access control partially enables users to control tension at the temporal boundary. Conceptually, diligent users could manage the tension at the temporal boundary by revisiting access control settings regularly, but this requires a huge time investment from users and is ripe for mistakes.

However, the access control model does not capture the identity boundary. Intuitively, tension at the identity boundary occurs when some of the audience of OSM content misinterpret the content (e.g., reading a post on political opinion out of context). We stress that it is extremely hard for a user to anticipate such misinterpretation. However, she might be relatively certain that her intended audience for a particular content are not going to misinterpret that content. Access control forces uploaders to specify the *right* audience a priori which might not always be possible.

Aspects of Solove's taxonomy of privacy [146, 147] captured (and not captured) by access control: Solove provided a taxonomy of privacy violations in data flow which can be used in real-world data-driven systems like OSMs. However, he considered a generic data-driven system, so not all privacy violations Solove mentioned are applicable to OSMs. Recall that there are four broad categories of data flow pointed out by Solove (Chapter 2.1.6): information collection, information processing, information dissemination and invasion. Out of them invasion of an individual in OSMs (e.g., by forcing a user to divulge her Facebook password) is again out of the scope of our discussion. Out of the other three categories the

access control model concretely captures only information collection; the access control model enforces that only the people a user explicitly allowed can access a piece of OSM content.

Information processing can result in privacy violations that are not captured by access control. For example, access control does not stop user data aggregation or inferences based on that data. Third-party crawlers like Spokeo cause this violation through their use of public OSM data. The access control model also does not address the dangers of information dissemination. In fact, it does not capture the scenarios when privacy violations are caused by increased accessibility of data. However, Facebook faced privacy violation accusations due to their timeline feature, which makes digging up a user's historical content extremely easy (thus increasing accessibility).

Aspects of Nissenbaum's "privacy as contextual integrity" [24, 124] captured (and not captured) by access control: Nissenbaum's contextual integrity (CI) model (Chapter 2.1.7) of privacy postulates that privacy is preserved by maintaining only appropriate flows of information that conform to contextual information norms. A privacy violation occurs when the contextual informational norms are violated. Access control captures a very basic idea of contextual integrity for a piece of content: access control ensures privacy only when the informational norms are very concretely specified by users via access control lists (list of users who are allowed/denied to access the OSM content).

However, contextual integrity takes much more nuanced forms than captured by access control. For example, earlier work [126] pointed out that many OSM users do not concretely understand what are the correct informational norms, often due to the novelty of these OSMs. However, the same work revealed that people still have an implicit idea about how contextual integrity should be preserved (e.g., a user's location, although publicly posted in OSM, should only be revealed to the people who are close to the user). One prime example of this scenario is the "privacy in public" scenario. CI pointed out that there are implicit informational norms and hence appropriate information flows associated with even

public information. However, the access control model assumes that public information is accessible to anybody and thus access control does not enforce any constraints on who should see the information or how they should use it.

Summary: Our review of access control in light of privacy theories revealed that access control does capture many aspects of privacy in OSMs. However, the access control model still misses some important dimensions of privacy. Specifically, an important aspect of privacy mentioned in the modern privacy theories is "user desire", which is not captured by access control. Users desire/expectation is captured by reserve in Westin's theory, identity boundaries in Palen and Dourish's theory, desired level of privacy in Petronio's theory or the informational norms in Nissenbaum's theory. However, access control does not sufficiently capture user desire for a number of reasons. First, with access control, users must *a priori* specify precisely who can or cannot access information by enumerating users, groups, or roles—a task that is difficult to get right. Second, access control fails to separate who *can* access information from who actually *does*, because it ignores the difficulty of *finding* information. Third, access control does not capture if and how a person who has access to some information redistributes that information. We summarized the dimensions of privacy encapsulated in the modern theories of privacy, distinguishing among those that are captured by the access control model and those that are not, in Table 2.1. However, since access control is quite useful in designing privacy preserving mechanisms in OSMs as well as generic online systems, we ask: are there cases of privacy violations in OSM in the real world that concretely point out the shortcomings of access control?

2.2.2 Case studies of real-world privacy violations not captured by the access control model

OSMs provide privacy controls based on *access control* and require users to allow or deny access to their content by specific users or groups. Recently, there have been a number of

Privacy definition / theory	Aspect(s) captured by access control	Aspect(s) NOT captured by access control
Warren and Brandeis's theory [168]—privacy as the right to be let alone.	The right to be let alone, given OSM users know a priori the exact set of people with whom they want to share content.	-
Westin's definition of privacy [169]	Solitude (being free from observation by others) and intimacy (small group seclusion for members to achieve a close, relaxed, frank relationship).	Reserve (desire to limit disclosures to others).
Altman's privacy theory ("the selective control of access to the self") [12, 13, 14]	Dynamic interpersonal boundary control (e.g., changing access control over time), bi-directionality, individual/group level.	The difference between desired and actual levels of privacy and consequently the notion of optimal privacy (i.e., when desired = actual).
Petronio's Communication privacy management (CPM) theory [128]	Ownership of information and privacy rules created by users.	The boundary turbulence due to mismatch of privacy rules by owner and co-owners (in effect, mismatch between desired privacy and actual audience).
Palen and Dourish's extension of Altman's theory to the networked world [127]	Tension at disclosure boundary (selective disclosure of personal information) and temporal boundary (via change of access control over time).	Tension at identity boundary (how would recipients interpret a piece of information).
Solove's taxonomy of privacy [146, 147]	Information collection methods like surveillance (the watching, listening to, or recording of an individual's activities).	Information processing (e.g., aggregation—gathering together of information about a person) and information dissemination (e.g., increased accessibility—amplifying the accessibility of information).
Nissenbaum's privacy as contextual integrity [24, 124]	Informational norms which are explicitly specified by users via access control lists.	Situations like "privacy in public" i.e., implicit informational norms.

Table 2.1: The aspects of different privacy definitions which are captured by access control and the aspects which are not captured by access control.

incidents that call into question whether access control is the right mechanism with which to implement privacy. We list a few here; this list is by no means exhaustive.

1. When Facebook introduced the *News Feed*—a feature that automatically presents updates from friends when a user logs in, as opposed to requiring the user to visit the friends' pages—users objected strongly and accused Facebook of privacy violations. Strictly speaking, News Feed did not change the access control policy; all users who viewed content through the News Feed had access to the content before. However the change from a pull mechanism to a push mechanism resulted in users feeling that their privacy had been violated.

2. There was a similar outcry of privacy violations when Facebook introduced *Timeline*, a feature that indexes a user's content by date of upload and allows users to quickly browse content by upload date. As with the News Feed, Timeline did not change the access control policy of any content. Instead, Timeline made accessing old (and potentially embarrassing) content significantly easier.

3. Google's *Street View* project—providing photos of houses and other property taken from public street—has also been accused of violating the privacy of users. In the U.S., there is no legal expectation of privacy on a public street (i.e., Street View photos can legally be posted publicly), but many users feel uncomfortable that Street View has made information easily and widely accessible that previously was visible only to those physically present.

4. Data aggregator *Spokeo* links together public information from different services (e.g., government databases, sites like LinkedIn, etc). While each individual piece of content that Spokeo aggregates is publicly available, users have complained that their privacy is violated when this information is linked together. For example, Spokeo cross-references users' addresses with property records, allowing others to quickly estimate someone's wealth using public information.

While perceptions of privacy and what constitutes a privacy violation are subjective, most people would likely agree that each of the incidents above affect someone's privacy.

However, the take-away from all of these incidents and others is that none of them involved a violation of *access control*. As a result, we argue that privacy is not adequately captured by access control alone, and the research community should re-consider how to improve the access control model and reason about user privacy.

2.2.3 Goal: A more inclusive privacy model

In this thesis, we carefully reconsider the issue of privacy in the age of the web and social media. We propose a model of privacy based on *exposure*, where the exposure of a piece of information is defined as the set of principals (people) who are expected to eventually know it.[3] Users implicitly reason about the exposure of various pieces of information; a violation of exposure occurs when the set of users who become aware of a piece of information is much different from what the user expected. In fact, recent work [27] by Facebook researchers have shown that such exposure violations are commonplace; e.g., many users significantly underestimate the number of users who actually view their content.

For example, consider the case of a user's public Facebook page being linked to from a high-profile web site such as the New York Times. Strictly speaking, there is no access control violation; the user's profile was previously publicly visible. However, a significant change of exposure occurs as the set of people expected to see the page increases from a small set of users likely to visit the user's page to the much larger set of New York Times readers. We argue that exposure naturally captures the privacy change of such an incident, and makes clear why access control alone is insufficient. Note that our model of exposure control (i.e., controlling the entities who actually view the content) can be interpreted as a simple improvement over access control (controlling the entities who can access the content). We will compare access control and exposure control in more detail later in this chapter.

[3]Note that our definition of the term *exposure* is quite different from the term exposure in Solove's taxonomy (Chapter 2.1.6). Solove defined exposure as revealing one's nudity, grief, or bodily functions. However, in this thesis, we define exposure to refer to the set of entities who actually view content.

We discuss mechanisms that could increase user's control over privacy by moving from *access control* towards *exposure control*, and describe how these mechanisms could be built into today's content sharing systems.

The remainder of this chapter is organized as follows. In Chapter 2.3, we provide a more formal definition of exposure and compare the exposure control model with more traditional access control. In Chapter 2.4, we describe approaches that could provide users with improved privacy via exposure controls. We expand upon our proposal to control exposure via prediction in Chapter 2.5 and explore the feasibility of controlling exposure via prediction. Finally, we summarize our findings in Chapter 2.6.

2.3 Defining exposure

In this chapter, we propose a simple model for exposure control.

Let I be an item of information (e.g., Alice's date of birth is Jan 1, 1980). Informally, I's exposure is the set of principals we expect to *eventually* learn I. The exposure set includes principals who learn I directly from Alice or indirectly from a third person with knowledge of I, and those who infer I from other knowledge available to them.

More precisely, we define the *prominence* $P_I(t)$ as the set of principals who are aware of I at time t.[4] I's exposure $E_I = \lim_{t\to\infty} P_I(t)$. Note that E_I is always finite, because the set of principals (i.e., the world's population) is finite. However, the exposure of most information items, even if they are publicly accessible, is much smaller than the world's population, because they are of interest to only a small community.

Normally, $P_I(t)$ is unknown for $t > currentTime$. Future values of $P_I(t)$ must be estimated using a probabilistic model, which captures how information spreads among principals; the exposure is given by the expected steady-state prominence predicted by the model. An example of such a model is discussed in Chapter 2.5.

[4]Prominence is assumed to be a monotonically non-decreasing function of time. That is, we ignore that people forget or misplace information.

2.3.1 Aspects of exposure

The exposure of an item I is influenced by two factors:

1. The set of principals N_I that meet the preconditions required to learn I. (Preconditions include the expertise, access to the tools, and knowledge of initial leads required to discover or infer I.)

2. The subset of principals in N_I that is sufficiently motivated to actually learn I.

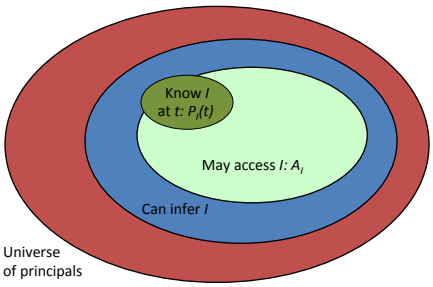
For instance, if learning I requires correlating several pieces of related information, traveling to a particular location or performing a measurement, then it is likely to be learned only by principals with the necessary resources and a strong interest. If, on the other hand, I is online and indexed by a search engine, then it can be learned by anyone with access to the Internet, the expertise to use a search engine, knowledge of appropriate keywords, and sufficient interest to actively issue a search query. Lastly, if I is posted on the front page of the New York Times, then all principals who visit the site on a daily basis will likely learn I serendipitously, even if they are only mildly interested.

The exposure of an item of information may change over time. For instance, when a little-known website is listed on Slashdot, the set of users likely to discover the information contained in it increases dramatically and unexpectedly. Such events cause a discontinuity in the prominence function $P_I(t)$, and thus a potential change of exposure.

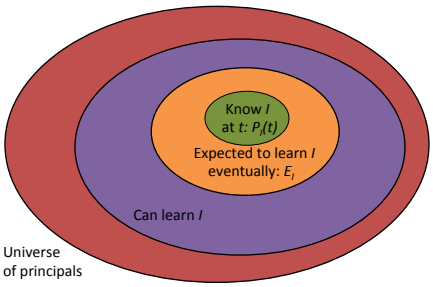
2.3.2 Privacy violations covered by exposure

Whether the release of information about a person is considered a privacy violation by that person is subjective and deeply rooted in the person's culture, history, situation, the nature of the information, and the specific set of people who learned the information.

In general, however, a person is more likely to feel violated if she is surprised by the fact that certain people have learned the information. There are two relevant cases. A person



(a) **Access Control:** A_I both over- and underapproximates $P_I(t)$.



(b) **Exposure:** E_I is the set expected to learn I eventually.

Figure 2.1: Access control and exposure of an information item I shown as a Venn diagram.

typically has some expectation about (a) the set of people who know or are likely to learn an item of information, and (b) a specific set of people they expect should not and will not learn the information. A person tends to feel their privacy is violated if the actual exposure of an item includes many more people than the expected exposure [46], or if people in the second set learn the information.

An example of the former case would be if Alice finds out that a picture showing her dancing wildly at a party has been seen by all of her friends, family and colleagues, when she expected that it would become known only to the people who had attended the party. An example of the latter case would be if Alice's work colleagues find out that she is gay (even though she shares this information freely with her friends and family, and she makes no attempt to hide it from people she encounters in her life outside of work).

Exposure captures these concerns because it reflects the set of people likely to find out the information. In general, the exposure control model captures the privacy violation scenarios where the expected exposure of users deviated from actual exposure.

2.3.3 Comparison with access control

Figure 2.1 contrasts access control and exposure using a Venn diagram. In the access control model, the set of principals is partitioned into those who are able to access I and those

who are not. Access control does not capture how many principals with access permissions actually access the information, nor does it account for principals without permission who nevertheless learn the information, either by inferring it from other information they can access, or from another principal with access.

We want to highlight two specific aspects of exposure control. First, although our general notion of exposure captures the privacy violation due to inference/re-sharing, measuring extent of such inferences is a hard problem and designing OSMs to protect against such attacks is even harder. So we will limit ourselves in this thesis to exploration of exposure control when there is no inference/re-sharing attack. Our assumption essentially makes the exposure set a subset of users who can access information I. Other researchers [117, 34] investigated inference/re-sharing attacks in OSMs and their work is complementary to ours.

Second, exposure control is an extension of access control (since exposure is a subset of who can access I by our simplified assumption above). Thus in practice, we envision exposure control augmenting access control in order to provide better privacy management mechanisms to OSM users. Specifically, a user can (ideally) control the exposure of a piece of content by including only the set of people who would actually view the content in the access control list of the content (e.g., very close friends). However, setting such perfect access control lists at the time of content creation might be difficult of even impossible for the user (exemplified by a plethora of OSM content shared publicly today). To that end, we review the additional aspects of privacy that our exposure control model captures which are not captured by access control alone.

2.3.3.1 Aspects of privacy definitions additionally captured by exposure control

Earlier, we reviewed which aspects of privacy are captured and not captured by access control (Table 2.1). Here we describe the additional notions of privacy that we can capture with exposure control. We summarize our findings in Table 2.2:

Aspects of Warren and Brandeis's theory [168] additionally captured by exposure control: Warren and Brandeis theory of privacy as the right to be let alone. As we mentioned, this early theory captures only a basic requirement of privacy, and the access control model already captures this basic requirement.

Aspects of Westin's definition of privacy [169] additionally captured by exposure control: Exposure control additionally captures the state of "reserve" (desire to limit information disclosure to others and others recognizing and respecting that desire), which is not captured by access control. This is because the exposure control model integrates users' intentions, not just the reality of their access control setting, as part of the expected exposure set. Thus, even if an OSM user uploads a photo to an OSM and sets the access control to the public setting (all of the internet), she might expect the exposure set of the photo to contain only her close friends and family. Her expected set likely does not include web crawlers.

Aspects of Altman's privacy theory [12, 13, 14] additionally captured by exposure control: Unlike access control, exposure control captures the difference between desired and actual levels of privacy using the notion that privacy violations might occur when expected exposure deviates from actual exposure. Thus exposure control indeed distinguishes between desired and actual levels of privacy. Moreover, just like Altman's theory, exposure control also strives to keep the actual exposure the same as the expected exposure, thus allowing a notion of "optimal level of privacy" (when actual exposure = expected exposure).

Aspects of Petronio's Communication privacy management (CPM) theory [128] additionally captured by exposure control: Our exposure control model captures two propositions of Petronio's theory which were not captured by the access control model. First, when an OSM user uploads her content, her expected exposure set consists of only the people who are expected to view the content. Thus, via this expectation, users are restricting the set of co-owners to the expected exposure set and not all people who just *can* access the content. Thus, the co-owners in the exposure control model are much more restricted and closer to

user expectations. Second, exposure control captures the idea of having an optimal set of co-owners (i.e., expected exposure set = actual exposure set in exposure control) helps to maintain less tension while ensuring the desired flow of information for the content owner.

Aspects of Palen and Dourish's extension of Altman's theory to the networked world [127] additionally captured by exposure control: Recall that the access control model does not capture tension at the identity boundary (when some of the audience of OSM content misinterpret the content) in Palen and Dourish's theory. Exposure control helps to capture this boundary because the aim of the exposure control model is to help users align their actual exposure with the exposure they expect. We argue that users originally expected their content to be actually viewed by people in their expected exposure set. Thus, there is lower risk of tensions at identity boundaries when exposure control is in place along with access control.

Aspects of Solove's taxonomy of privacy [146, 147] additionally captured by exposure control: Exposure control captures some privacy violations in the information processing and dissemination phase that are not captured by access control. For example, access control does not recognize the danger of aggregating and mining public data or even the privacy risks of easy accessibility (e.g., advanced search functionality) of data. However, exposure control argues that aggregation or increased accessibility potentially increases the actual exposure of content—since these mechanisms make content easier to learn. Hence, exposure control captures these additional privacy violations that are not captured by access control.

We note that even our exposure control model does not capture some privacy violations pointed out by Solove. Examples include insecurity, secondary use, exclusion, disclosure, appropriation, distortion and decisional interference. Capturing these privacy violations in OSMs is part of our future research directions.

Aspects of Nissenbaum's "privacy as contextual integrity" [24, 124] additionally captured by exposure control: Exposure control catches the more nuanced idea of contextual

informational norms compared to the access control model, thus better capturing the notion of privacy as contextual integrity (CI). The reason is that the exposure control model argues for controlling actual exposure, so that it is the same as user's expected exposure. Thus expected exposure partially captures the past experience of users (which plays a role in following the expectation). For example, exposure control captures the idea of "privacy in public": CI pointed out that there are implicit informational norms and hence appropriate information flows associated with even public information. The exposure control model points out that even for public content (anybody can access the content), the user will still have an expected exposure of how many people will actually view the content. Thus, she still has some notions of expected exposure even for public content.

However some aspects of CI are not captured by exposure control. Since the exposure control model relies on the user's expected exposure, if the user has no prior expectation of what exposure her content should have, then the exposure control model can not help them. However, CI captures possible violations even in this case, since the contextual informational norms can be set by society at large and ethical concerns even without that particular user's expectation.

Summary: From these discussions, we note that our exposure control model considerably extends access control by capturing a key aspect of privacy—a user's expectation or desire about who views her content. We explicitly capture the notion of this user intention mentioned in privacy theories (e.g., reserve in Westin's definition, identity boundary in Palen and Dourish's theory, desired level of privacy in Petronio's CPM theory or some implicit informational norms in Nissenbaum's theory). We summarize our findings on the additional aspects of privacy captured by exposure control in Table 2.2. These aspects were not captured by the access control model alone. This table clearly demonstrates the benefit of designing OSM systems with mechanisms for controlling exposure. Let us now revisit the real-world privacy violations from Chapter 2.2.2 to further check the effectiveness of exposure control.

Privacy definition / theory	Additional aspect(s) captured by exposure control
Warren and Brandeis's theory [168]—privacy as the right to be let alone.	-
Westin's definition of privacy [169]	Reserve (desire to limit disclosures to others), since exposure considers the user's expected exposure while controlling exposure.
Altman's privacy theory ("the selective control of access to the self") [12, 13, 14]	The difference between desired and actual levels of privacy and consequently the notion of optimal privacy (i.e., when desired = actual). This is because actually the people who saw a content is the actual privacy and expected exposure is the desired privacy.
Petronio's Communication privacy management (CPM) theory [128]	The boundary turbulence due to mismatch of privacy rules by owner and co-owners. Since exposure control model captures, in effect, the tension between desired exposure (from the owners) and actual exposure (affected by co-owners).
Palen and Dourish's extension of Altman's theory to the networked world [127]	Tension at identity boundary. Since the people in expected exposure set will possibly interpret a piece of information as the user intended.
Solove's taxonomy of privacy [146, 147]	Information processing (e.g., aggregation—gathering together of information about a person increases actual exposure of content and violate privacy without violating access control) and information dissemination (e.g., increased accessibility—amplifying the accessibility of information also increases actual exposure by making the content easier to find).
Nissenbaum's "privacy as contextual integrity" [24, 124]	Situations like "privacy in public",i.e., some implicit informational norms. Since the implicit informational norms can be partially captured by the expected exposure of a content.

Table 2.2: Additional aspects of different privacy definitions captured by exposure control. The aspects of privacy captured by access control (in Table 2.1) and exposure control in aggregate captures most aspects of existing privacy theories.

privacy violations we discussed in Chapter 2.2.2. Recall that these violations are not captured by the access control model alone.

1. Exposure captures the changes caused by the introduction of the Facebook News Feed: Prior to its introduction, the exposure of an item I on Alice's profile was the number of unique users who visit Alice's Facebook page during I's lifetime, which could be much smaller than the set of users N_I with permission to access I, particularly if Alice chose to make I public. With the News Feed, in contrast, I's exposure includes all of Alice's friends plus any user in N_I who Facebook deems potentially interested in I. I is pushed to these users, who will learn I serendipitously the next time they log into Facebook.

2. Similarly, the introduction of Facebook's Timeline pushes selected information about a person's history to a set of principles in N_I that Facebook deems interested. Previously, finding such information would have required a user in N_I to visit Alice's profile and scroll potentially deep down into her historic News Feed.

3. Google Street View has made available online, in an aggregated and searchable fashion, public information that was previously available only to principals who were physically present at a particular geographic location. Exposure reflects this fact.

4. Spokeo aggregates people's personal information like name, address, data of birth, income, property value and family tree, which is available from different online sources, and makes it available and searchable under the person's name and place of residence. By making it far easier to learn this information, its exposure is increased.

Thus, in short, exposure control captures real-world privacy violations that the access control model does not capture. However, exposure control still has its own set of limitations.

2.3.4 Limitations of exposure control

Our model of exposure has a number of limitations. First, it focuses on an individual item of information. When considering a collection of personal information about a person, for

instance, additional aspects of privacy like the credibility or authenticity of the information come into play, which are not captured in our model. Second, there are aspects of privacy theories which are still not captured by exposure control; these cases include some aspects of Solove's privacy taxonomy (insecurity, secondary use, exclusion etc.) as well as the implicit contextual norms that are deep-rooted in social processes. For example, suppose an OSM user is not aware that she is being stalked by ex girlfriends/boyfriends/strangers (and thus include the stalker in her expected exposure set). However stalking is still violating current social norms and thus constitutes a violation of contextual integrity.

2.4 Managing privacy via exposure

In this part, we discuss how the concept of exposure can be leveraged in practical systems to enable users to better manage their online privacy. We mention three strategies to control exposure below. In this section, we will expand upon one of these strategies. We will systematically investigate the rest of the strategies in details later in this thesis (Chapters 3, 4, 5 and 6).

2.4.1 Controlling exposure via setting conservative access control

OSM users can naturally control exposure by meticulously choosing a subset of their friends who they expected to view the content and only including that subset in their access control list. We refer to these subsets of users who are able to access a particular piece of content as social access control lists (social ACLs, or SACLs); by definition, a SACL is a proper subset of a userâĂŹs friends who are selected by the user to access a piece of content. Note that this solution is not perfect, since entities outside of SACL can often infer a particular piece of information from other available information. However, the ubiquity of access control mechanisms in OSMs makes this approach readily available to users. To that end, our work sheds light on how users are deploying conservative SACLs in the real world, allowing only

a subset of their social connections to access a particular piece of content. Furthermore, based on our understanding we propose mechanisms to help users specify their SACLs and control exposure. We present this work in Chapter 3.

2.4.2 Controlling longitudinal exposure

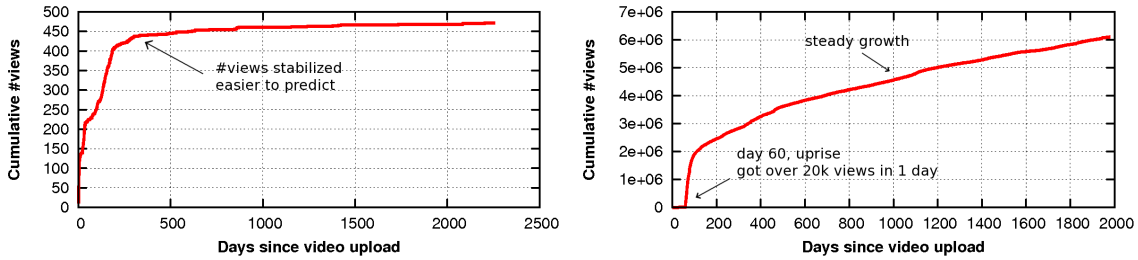
Earlier we mentioned the temporal dimension of exposure, in which a piece of content can be seen by more and more people as time passes. This longitudinal exposure may be quite concerning to OSM users, especially if they think that others may view and interpret some past content totally out of context. To that end, a natural way to control longitudinal exposure for users is to simply withdraw their past content. We further investigate the withdrawal of past content in OSMs and how we can improve longitudinal exposure in Chapter 5 and Chapter 6.

2.4.3 Controlling exposure via prediction

The former two strategies of exposure control are primarily user-driven, e.g., users themselves set SACLs or users control their longitudinal exposure by withdrawing historical content. However OSM operators can be more proactive in designing their systems to control exposure, e.g., by predicting the expected exposure of a content while creation using past history and minimize deviation of actual exposure from that prediction. In Section 2.5, we develop a general methodology of controlling exposure via prediction. In Chapter 4 we will thoroughly investigate one simple case of such exposure control—controlling exposure by limiting large-scale social data aggregators.

2.5 Generic framework to control exposure via prediction

There are two important notes to make before we discuss exposure control via prediction. First, our goal is to propose a general methodology that could be broadly applied to control



(a) **Niche video with total 471 views** (b) **Popular video with total 6,101,294 views**

Figure 2.2: Popularity growth patterns of two Youtube videos. The popularity of the niche video stabilize and becomes predictable within a year, but the popularity of the popular video exhibits an uprise and shows steady growth.

exposure of users' information in a variety of online systems. Thus, our discussion is not specific to any one system. Second, there are several interesting research questions that remain to be investigated through a concrete implementation and deployment of our proposal. However, such a full-blown deployment-based study is beyond the scope of this thesis and we view the deployment challenges of our generic framework as specific future research directions.

2.5.1 Predicting exposure

Modeling and predicting the growth in popularity of information on social Web sites like Facebook photos, Twitter posts or YouTube videos [100, 86, 154] has received significant research attention recently. These studies use empirical data of how information became popular in the past to build models for information propagation that can predict the future popularity of similar information. Popularity growth models are relevant because they can predict he cardinality of exposure. The prediction models vary from very simple models that extrapolate from the historical growth in popularity of a single piece of information; to more complex models that take into account various factors including attributes of the information (e.g., quality, type, and length of a video), historical data about the spread of other similar pieces of information, and the effectiveness of different information dissemination channels

(e.g., social links between users or personalized recommendations or search results). More sophisticated models with higher prediction accuracy have been developed over time. While a detailed discussion of these models is beyond the scope of this work, we make two general observations that are relevant to our discussion:

1. Prediction accuracies are higher for less popular (niche) information than they are for more popular information. For example, it is easier to predict the future popularity of YouTube videos with a few hundred views after 1 year than those with few million views after 1 month [154]. As shown in Figure 2.2(a), the dissemination of niche videos tends to stabilize to a predictable rate sooner than those of popular videos.

2. Most models cannot anticipate the occasional sharp, disruptive jumps in popularity that arise due to unanticipated events, such as when a piece of information goes viral or when it is featured on a prominent site [136]. Figure 2.2(b) shows an example YouTube video whose number of views experienced a sharp jump on day 60 due to coverage on popular media and blogs.

2.5.2 Making the predictions transparent

We argue that system operators (e.g., Facebook or YouTube administrators) should make both past popularity data and predictions for the (cardinality of) exposure of users' information transparent to the user. Currently, some systems provide users with a limited view of the popularity of the information they upload. For example, Facebook and YouTube allows users to check the number of views or "Likes" for their posts. However, no site today explicitly provides estimates of the future popularity of a piece of information. For example, neither Facebook nor YouTube offer guidance on how many and which people might see a photo over the next week. We see this as a missed opportunity because (i) the site operators are often in the best position to make such predictions as they have the best access to all the empirical data on how information disseminates through their sites and (ii) such

estimates would enable users to (re)calibrate their expectations for the future exposure of their information and check for potential privacy violations.

When providing estimates for the exposure of a piece of information, it would also be useful to estimate the likely exposure through different dissemination channels separately. For example, Facebook might choose to provide estimates of views a user's photo might achieve through updates on personalized news feeds versus graph search versus profile browsing [70]. Doing so would enable users to understand the predicted exposure of their content, and to modify the sharing settings if it does not match their expectations.

2.5.3 Controlling exposure

Providing users with more accurate exposure estimates for their information does not by itself eliminate the risk of privacy violations. System designs need to enable users to tune (i.e., increase or decrease) the exposure to the values they desire. Further, systems should be designed to also prevent the actual prominence from deviating significantly from the predictions (after they are tuned to desired values). Below we propose mechanisms to achieve the above two goals.

Tuning exposure: When a user finds that the predicted exposure of her information is different from what she desires, there need to be mechanisms that would allow her to tune the exposure. A user could do this in several ways: first, she could enable or disable one or more dissemination channels. For example, on Facebook, one could opt-in/-out of being part of "directory or graph search." Such opt-in/-outs from one or more dissemination channels could help users manipulate their exposure to desired levels.

Second, users can resort to more expansive or restrictive access controls (i.e., who is allowed to see or not see a piece of information) to change the exposure of a piece of information. For example, to increase exposure of a piece of information originally shared with her 1-hop friends, a Facebook user might choose to make it available to 2-hop friends (i.e., friends of friends). To decrease exposure, the user might choose to make it

available to only a subset of 1-hop friends (e.g., friends with whom the user shares a common university affiliation). By changing access controls, the user can expand or contract the list of potential viewers and thereby, change the list of predicted viewers. Thus, we envision access control being used in conjunction with exposure control to more closely match the user's expectations.

Limiting divergence from predictions: Even after a user tunes the exposure to match her expectations, unanticipated events (e.g., the information goes viral and is featured on the front page of a popular site) might cause the actual exposure to deviate significantly from the predictions and consequently, the desired exposure. As mentioned earlier in Chapter 2.5.1, no model can accurately predict such occasional disruptive changes to the prominence of a piece of information.

To contain privacy violations in such scenarios, we propose that systems adopt *tripwires* that automatically make a piece of information inaccessible whenever the actual exposure of a piece of information deviates significantly from the predicted exposure and notify the user of the unanticipated divergence. Upon notification, users can explicitly choose to keep the information inaccessible or re-enable access to the information (and readjust the tripwires). Alternately, systems can allow users to specify tripwires that upper-bound the views (e.g., no more than 10 views per day or 50 views in total) to a piece of information.

We believe that tripwire mechanisms can be easily enabled in current systems like YouTube or Facebook. In fact, YouTube already allows users to limit the total number of views to their videos to a preset value of 50 (effectively providing a limited form of exposure control) [176].

2.5.4 Feasibility of controlling exposure via prediction

There are a number of interesting research questions that need to be studied through a concrete implementation and deployment of our proposal of managing privacy via exposure. Such a detailed study is the subject of our future work and beyond the scope of this work.

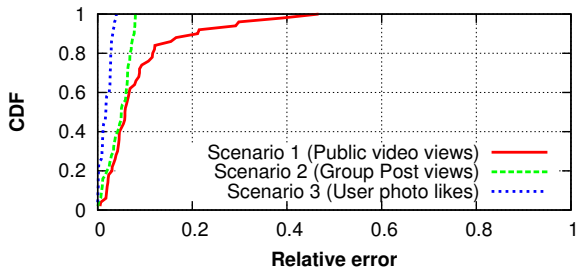


Figure 2.3: Prediction of number of people accessing a content in three scenarios: scenarios 1, 2 and 3 corresponds to prediction of public YouTube video views, Facebook group post views and photo likes on personal Facebook photos respectively. The relative error of prediction decreased from scenario 1 to 3.

However, we discuss two important concerns that one might have about the feasibility of a practical deployment of our proposal: (i) our ability to accurately predict the future exposure of a piece of information and (ii) the overheads associated with fine-tuning exposure.

2.5.4.1 Accuracy of exposure predictions

There is ongoing research on predicting information propagation and dissemination in online systems. These works leverage the ability of sites to gather and analyze detailed historical information about the exposure of billions of pieces of information posted by hundreds of millions of users to make accurate predictions. For example, in a recent study [27], Facebook researchers were able to predict the audience size of a new post by a user within an 8% error margin, using data such as the number of friends and the likes and comments the post received. To illustrate the ability to make such predictions in different scenarios, we conducted a small-scale study in three different real-world scenarios, each using different access control policies—(i) public posts on sites like YouTube, (ii) posts limited to members of Facebook groups and (iii) personal posts limited to one's Facebook friends.

In each of these scenarios, we used linear regression [102] for predicting the future popularity using past information and then measure the *relative error* of our prediction. Relative error is defined as $\left|1 - \frac{\text{Predicted value}}{\text{Actual value}}\right|$. The lower the relative error, the more accurate the predictions. In each of the scenarios, we show how system operators can use different

types of past information to predict the future popularity of a content with low error, i.e, high accuracy.

Scenario 1: Predicting future popularity of public YouTube posts. We analyzed the past number of daily view for publicly posted YouTube videos to predict their future popularity. Specifically, we collected data about the number of daily views that 50 randomly chosen YouTube videos obtained in their first 6 months. We used this historical data to predict how many views the videos will get the immediate next day.

Scenario 2: Predicting future popularity for new posts in Facebook groups. Next, we tried to predict the number of views for a new post in an open Facebook group [61] using the popularity of older posts in the same group [55]. We collected popularity data for the posts of 50 open Facebook groups and predicted the number of views for the latest post in each of these groups using the popularity of older posts.

Scenario 3: Predicting future popularity of personal posts limited to one's Facebook friends. We collected data about the pictures shared by 50 Facebook users (randomly selected users of a Facebook application [71] created by authors) with their friends along with the number of "Likes" on those pictures. Using this data, we predicted the number of "Likes" a user would get on a future photo shared with the same access control policy. We performed this prediction for the latest photo of each user.

We present the distribution of relative errors in predictions for each of these scenarios in Figure 2.3. Note that, intuitively the set of people who can learn about the content decreases from scenario 1 to scenario 3. Figure 2.3 shows that consequently the relative error decreases from scenario 1 to scenario 3. However, even in the case of scenario 1, where the videos are public and the information can spread through multiple possibly unknown channels, for 75% of the videos the relative error is less than 0.1 (i.e. actual value is within $\pm$10% of the predicted value).

Our study, while conducted at a very small scale, hints at the potential for accurately predicting the future popularity in different real-world scenarios. We plan to conduct larger-

scale studies in the future. Furthermore, there is significant ongoing research on predicting information propagation and dissemination in online systems and new techniques are being proposed [31, 70, 86, 27]. As the field advances, we expect the accuracy of predictions to improve as well.

2.5.4.2 Overheads associated with fine-tuning exposure

At first glance, it would appear that supporting exposure control would require users to check and fine-tune the exposure for every single piece of their information separately, which raises usability concerns. In practice, users might want to organize all their pieces of information into a small number of groups, each with a different level of desired exposure. So when a new piece of information is uploaded, they can easily set its exposure to the desired level by choosing the appropriate exposure group. The overheads involved here would be no greater, if not lower, than the overheads involved with configuring privacy settings of uploaded content in social media sites today.

2.5.5 A special case: Controlling exposure by limiting large-scale social data aggregators

We note that a OSM operators can deny access proactively to entities who they can predict to most likely not be in the expected exposure set of a user. Of course the prediction overhead is minimal in this special case for the OSM operators. A practical manifestation of this situation occurs when third party large scale data aggregators, i.e. crawlers (e.g., Spokeo) aims to collect public user data from OSMs. Although the data is public, the OSM operator can (with high confidence) predict that a crawler is not in the expected exposure set of a piece of public user generated content and take steps to limit access of the crawler and fine-tune the exposure. Today OSM operators control exposure to crawlers by putting rate limits on number of views from particular IP addresses or particular accounts. However their technique cannot stop crawlers who leverage fake (Sybil) or compromised accounts.

We will revisit this scenario in Chapter 4 to propose a system for OSM operators to control exposure from large scale third party data aggregators.

2.6 Discussion

We argue that access control, the traditional model for managing privacy, is inadequate in today's online world. In this work, we propose a model for information privacy based on exposure to fill the gaps left by access control. A key difference compared to access control is that exposure captures the principals who *actually learn* a piece of information rather than who *can directly access* a piece of information. We believe exposure is an intuitive measure that captures the privacy implications of publishing information and strengthens privacy management in OSMs.

We discussed how the concept of exposure can be leveraged in practical systems to enable users to manage their online privacy better. We propose three strategies from the real world—(i) controlling exposure via SACLs, i.e., social access control lists (ii) Controlling exposure via prediction, a special case of which is controlling exposure from large scale third part crawlers (ii) controlling longitudinal exposure. We will investigate each of these scenarios to demonstrate the effectiveness of exposure control in the next chapters, starting with exploring how exposure can be controlled using SACLs.

CHAPTER 3

Understanding and controlling exposure using SACLs

OSM users can naturally control exposure by meticulously choosing the subset of their friends who they expected to view the content and only including that subset in their access control list. Today, while much OSM content is shared with default settings (e.g., visible to all of a user's friends or even to everybody on internet), certain sensitive content is indeed often shared with subsets of friends. For example, on Facebook, users may explicitly enumerate friends to allow or deny the ability to view a photo, or create friendlists for the same purpose. We refer to these resulting sets of users who are able to access content as *social access control lists* (social ACLs, or SACLs); by definition, a SACL is a proper subset of a user's friends who are selected by the user to access a piece of content. Due to the privacy-sensitive nature of the content the SACLs protect, one of the hardest parts of using today's OSM privacy management tools is defining appropriate SACLs for different pieces of content. Thus for effective exposure control the OSM operators needs to simplify the SACL specifications for the OSM users.

Many prior studies have examined the privacy concerns that arise when users share content on Facebook, such as the problem of "over-sharing" content with default settings that make the content visible to everyone in the network [103]. As a solution, researchers have proposed grouping friends into subgroups based on their relationship type (e.g., high school friends, work colleagues, family) or community structure in the one-hop network of the user,

and sharing content with specific subgroups [59]. However, most of these approaches rely on small scale user studies where they conduct a survey to understand the privacy preferences of users to evaluate their technique. None of these approaches have been evaluated on how well they capture real privacy preferences specified using SACLs. Given that content shared with SACLs is likely to be the most privacy sensitive (and therefore, likely the most important), having an understanding of the SACLs in-use is crucial to designing improved privacy mechanisms for OSM users.

In this work, we make three contributions: *First*, we conduct the first large-scale measurement study of use of SACLs in OSMs. Using a popular Facebook application installed by over 1,000 users, we collect a total of 7,602 unique SACLs specified by users.[1] We find that over 67% of users are sharing at least some of their uploaded content using SACLs, and that 17.6% of all content is shared with a SACL; these observations underscore the important and unstudied role that SACLs play in users' privacy management.

Second, we focus on understanding the membership of SACLs (i.e., how are the friends who are allowed to view a piece of content similar to each other, but different from other friends?). Examining the in-use SACLs that we collected, we find that for less than 10% of SACLs all the members of the SACL share a common profile attribute. Moreover, we find that only 20% of SACLs show strong community structure in the links between their members. Taken together, these results suggest that SACLs are likely to be difficult to detect automatically. This result is surprising given the existing work on automatically grouping friends based on network structure or attributes for better privacy management [112, 59]; We suspect that this difference occurs because these prior studies did not evaluate their techniques against ground-truth data about fine-grained content sharing in OSMs.

Third, we explore the difficulty faced by users in specifying SACLs today. Overall, we find that the complexity of SACLs (as defined by the number of terms[2] a user must select

[1] Our study was conducted under Northeastern University Institutional Review Board protocol #14-01-09.

[2] When creating a SACL, a user can specify either individual friends or pre-created lists of friends; we refer to both of these as *terms*.

when creating a SACL) is quite high for a non negligible fraction of our users: over 18% of users specify more than 5 terms per SACL on average. We observe that there is significant room for improvement in reducing the burden of specifying SACLs, and we find that simply allowing users to re-use previously used SACLs reduces much of the user overhead: for the vast majority (> 80%) of users, 90% of their content is shared with fewer than 5 unique SACLs.

The remainder of the chapter is organized as follows. Chapter 3.1 presents background and related work on SACLs and OSM privacy. Chapter 3.2 describes how we obtained our SACL data set, and Chapter 3.3 provides some high-level statistics on SACLs. Chapter 3.4 explores the relationship between SACL members, while Chapter 3.5 investigates the user overhead in specifying SACLs today. Finally, Chapter 3.6 present a discussion of our findings and future directions.

3.1 Background and related work

In this chapter, we first provide some background on how OSMs have evolved in helping users to manage their data privacy today. Our focus is on the fine-grained privacy management tools that enable the sharing of privacy sensitive content on OSMs.

In this work, we focus on the largest OSM as of March 2014—Facebook. Up until 2005, Facebook split users on the site into different regional networks (based on geography, workplace or educational institution). By default, each user would share all of her content with everyone in the regional network and the service lacked any concrete privacy controls for sensitive data. By 2009, Facebook had 300 million [58] users and some regional networks grew too large (e.g., in India and China) to be used for privacy settings. There were widespread demands for better privacy management mechanisms for users [33], and by the end of 2009, Facebook rolled out more fine-grained privacy controls.

3.1.1 Mechanisms for privacy management

In December 2009, Facebook made an important change which allowed users to set access control policies for content they publish on a per-post basis [68]. For example, a user can share a particular photo with only family members and close friends. This change allowed users to customize their privacy settings on a per-content basis, instead of simply adopting the default privacy setting offered by Facebook, which allows access to "everyone" (all users on Facebook) [64].

Facebook introduced an additional mechanism called *friend lists* [73] to complement their existing fine-grained privacy controls. Users can create friend lists and add a subset of their friends to each of these lists. For example, a user can create a list called "co-workers" and manually add all of her friends who are co-workers into that list. This allows the user to group her friends into different lists that might be meaningful to her in terms of sharing content. Now, instead of handpicking individual friends for specifying a privacy setting for each content, users can use their pre-created friend lists for specifying privacy settings (e.g., share this photo with "soccer buddies" list). Friend lists are private to the user who creates them.

By October 2010, Facebook observed that only a small percentage (5%) of Facebook users had ever created friend lists [74]. This could be due to the manual effort required of the users to create and maintain friend lists. To help users further, Facebook started automatically creating friend lists for the user and populated the lists with a specific subset of the user's friends [57]; these lists are called *smart lists*. This automation is done by leveraging the profile attributes of the user and the user's friends, e.g., employer, location, family and education information provided by users. An example would be a list called "Family" that automatically groups all the friends of the user who have marked the user as a family member. In addition, Facebook also creates two empty smart lists for the user, "Close Friends" and "Acquaintances". However, instead of auto-populating these two lists,

Facebook only shows friend recommendations to the users based on the interaction between the user and her friends. In this work, we will refer to all of these Facebook-created smart lists as *Facebook lists*. Moreover, when using the term *lists* we are referring to both the user-created friend lists as well as Facebook lists.

So far, we observed that there are different ways in which a user can specify which friends have access to a piece of content on Facebook today. In the rest of the work, we will use the term *social access control lists* (social ACLs or SACLs) to refer to such privacy policy specifications. A more precise definition is below.

Social access control lists (SACLs): A SACL is a privacy policy specification attached to a piece of content containing a proper subset of the user's friends; friends specified in the SACL have access to view and perform other actions on the content (e.g, liking or commenting). SACLs can be specified using different mechanisms provided by Facebook: allowing or denying access to individual friends one by one, specifying friend lists, using Facebook lists, or using a combination of handpicked friends and lists.[3]

It is important to note that SACLs *only* encompass custom settings by users and do not include the Facebook pre-defined access permissions: "everyone" or "public" (share with all Facebook users), "regional network" (share with everyone in a regional network, deprecated in 2009), "all friends-of-friends" (share with all friends-of-friends), "all friends" (share with all friends), and "only me" (only visible to the user who uploaded the content).

3.1.2 Related work

Now that we understand the background of this work, we discuss related work along three directions.

[3]Facebook also allows users who are *tagged* in a specific post to see the content [67, 69]. However, since users did not specify tagged friends explicitly through the privacy management interface, we do not consider them to be the part of SACLs. We leave exploring privacy expressed through tags to future work.

Understanding privacy awareness of users Researchers have studied the privacy awareness of OSM users [48, 152, 92]. These studies examine the profile information sharing behavior of users over a long period of time (e.g., 7 years) to understand if users' attitude towards their data privacy changes over time. Dey et al. [48] and Stutzman et al. [152] have shown quantitative evidence that Facebook users are sharing fewer profile attributes (such as hometown, birthdate and contact information) publicly over time. Social media discussions about Facebook privacy [52, 141, 140, 153, 39] and Facebook regularly rolling out new fine grained privacy management features [66] for the last few years have caught users' attention and potentially increased their concerns about available privacy controls.

How effective are users in managing their privacy settings? Recent studies have explored how effective users are in managing their privacy settings. Studies have shown that there exists a mismatch between desired and actual privacy settings when users share content on Facebook [103, 85, 107, 29, 27]. Liu et al. [103] conducted a user survey about privacy preferences for photos uploaded by users on Facebook. They found that privacy settings match users' expectations only 37% of the time, and when wrong, users are exposing their content to a much wider audience (e.g., all friends, friends-of-friends or even everyone on Facebook) than they intended. While the exact reason for incorrect privacy settings is hard to infer, it could be due to poor privacy management user interfaces or the significant cognitive burden required to manage privacy of their sensitive content.

Techniques for better privacy management Several techniques have been proposed to reduce the burden on users when managing their privacy settings. We can organize work in this space into two high level categories: (1) The first approach is to assist in automatically pre-defining grouping of friends that might be meaningful to the user for sharing sensitive content later. Facebook allows the user to pre-define such friend groupings using the friend list feature. But friend lists on Facebook have to be manually specified by the user today and this user overhead could be reduced by these approaches. (2) The second approach is to help the user on the fly to specify SACLs while sharing content. They predict SACL

specifications with some input from the user. For example, if the user gives the name of a few friends that he wants to share content with, these approaches can automatically predict the remaining members of the SACL or provide recommendations of other possible SACL members. Now we will explain different proposals that fall in the above two categories.

PViz [112] is a proposal from first category that can automatically detect friend lists for a Facebook user and use it for better privacy policy visualization for the user. It leverages the network structure of the subgraph (where friendship relations are represented as edges and OSM users as nodes) induced by the user's friends (i.e., the user's "one-hop subgraph") to detect friend lists using a modularity-based community detection algorithm. Using information extracted from friends' profile, it can also automatically assign a label to each detected list. This helps the user to understand the composition of a list. Based on these predefined lists, PViz points out to users which of her friends from a list can view a particular content. PViz presents a user study based on 20 users, who find PViz useful for understanding their existing privacy settings better.

However, many previous works [15, 174, 59] fall in the second category. They focus on recommending friends on the fly to the users as the user starts sharing a piece of content and selects a few intended friends. Privacy Wizard [59] is one example of such a tool. Privacy Wizard leverages network structure and profile attributes (like gender, age, education, work, etc) to recommend friends for inclusion in a privacy setting. The process starts as the user tags a few of her friends as "allowed" or "denied" for a content. It then uses a machine learning algorithm to classify the remaining untagged friends into an allow or denied category. The authors designed a survey experiment with 45 Facebook users and 64 profile data items to evaluate the accuracy of their tool. They observed that on average if a user tags 25 of her friends, the wizard configures her privacy setting with high accuracy. However in their experiments they did not look into the ground truth data on how a user actually specifies SACLs while sharing sensitive content.

We conduct a large scale study comprising of 1,165 users and all their uploaded content to focus on the privacy settings used by users. Most prior work tries to approximate ground truth privacy preference data by asking user privacy preferences explicitly, most of the time via surveys or via a combination of surveys and profile data collection using apps [92]. However, none of these studies looked into the "ground truth" data on SACLs (i.e., how a normal user would share their content using SACLs without any external intervention). To the best of our knowledge, we are the first ones to look into ground-truth data on users' usage of SACLs to propose insights on how to assist users to reduce the overhead of specifying SACLs.

3.1.3 Key questions

While fine-grained privacy controls put users in better control of their data privacy, it is not clear how users are using these privacy mechanisms. In this work, we take a first look at SACLs specified by 790 Facebook users (users who created at least one SACL) for 212,753 pieces of uploaded content. Our analysis focuses on the following key questions:

- *How are users using SACLs today?* We analyze how often SACLs are specified by users and how different types of content are shared using SACLs.

- *Is SACL membership predictable?* We analyze characteristics of SACL members to understand if they have something in common that other friends of the user do not. Our analysis explores whether members have similar profile attributes, exhibit strong social network connectivity with each other, or share similar activity levels. If we are able to separate out SACL members from among all friends, we may be able to automatically create SACLs for users.

- *What is the user overhead in specifying SACLs?* We examine the overhead that users spend specifying SACLs today. The intuition is that, the more work required to create SACLs, the less usable the privacy mechanisms are.

- *What is the potential for reducing user overhead?* Based on our insights gained from analyzing SACL membership, we quantify the potential for further reducing the complexity of SACL specifications. Our findings serve as guidelines for designing better privacy management tools in the future.

3.2 Dataset

Now we describe the dataset we have collected about in-use SACLs on OSMs today.

3.2.1 Collecting data about SACLs in Facebook

Obtaining data at scale about user privacy specifications is quite challenging. In Chapter 3.1.2, we observed that most previous work used small-scale data about user privacy preferences. There are two main challenges in collecting large-scale data: First, we need permission to view the user SACLs. This is challenging as private settings on Facebook are private to the user; they cannot be obtained by crawling publicly visible user profiles. To address this challenge, we use the Facebook API [62], which offers methods to collect data about in-use privacy settings (provided the user gives us permission to do so). We therefore developed a Facebook application that helps users to better manage their privacy settings, and recruited users for the application. The Facebook application requests consent from the user to collect data about their SACL specifications for our research study. The data collection was performed under an approved Northeastern University Institutional Review Board protocol.

The second challenge is recruiting large number of users for the study who can provide consent to access their private SACL information. The traditional approach for recruiting users rely on personal communication (e.g., via email) or through an open call posted on a public bulletin board at a university or research lab. In such cases, the number of users that could potentially be recruited is usually limited to a few tens or hundreds. Another approach

is to use online crowdsourcing platforms like Amazon Mechanical Turk (AMT) [111]. However, using a Facebook application violates the AMT policy that workers should not be required to download and install an application [16].[4] Instead, to recruit a variety of users at scale, we leverage the Facebook advertising platform.

3.2.1.1 Facebook application: Friendlist Manager

We developed the Facebook app “Friendlist Manager" (or FLM) [72, 105] that helps the user to automatically create and update friend lists that could be used for specifying SACLs. This application reduces the user burden of manually creating friend lists. FLM automates list creation by leveraging the network structure in the “one-hop subgraph” of the user. It uses network community detection algorithms [30, 117] to find overlapping communities in the one-hop subgraph. We found that users found FLM to be helpful; 480 (41%) of our users allowed FLM to update at least one of their lists. It is important to note that for the analysis in this work, we only consider the content users shared and members of friend lists users had *before* installing FLM; this ensures that usage of our app does not impact the results.

When installing FLM, we request permission to access the following types of user data: basic user profile details including workplace, education, current city and family; privacy settings (including SACLs) used for all uploaded content (photos, videos, statuses, notes, music, questions, Shockwave Flash Player (SWF) movies, and checkins); and the friends, friend lists, and Facebook lists of the user. Should the user choose to not grant us access to their content, they are still allowed to use the application.

3.2.1.2 Recruiting users

To recruit users, we set up an advertisement for FLM on the Facebook advertising platform. The Facebook advertising platform allows us to reach out to the large Facebook population and target users with specific demographics. Our ad included the following text:

[4] Our data collection methodology requires users to install a Facebook application.

Need help to better organize your list of friends? Give FLM a try!

Starting from June 20th 2012, we ran five ad campaigns for 10 days targeting 10 countries where English is an official language: USA, UK, Australia, New Zealand, Canada, India, Pakistan, Singapore, South Africa, and Philippines. In total, 232 users installed our app during this period. After this initial push, our app received a steady stream of new users through August 2013; in total, we observed 1,007 additional users after the advertising campaign ended. We believe FLM also spread "virally", with users "liking" or recommending the app to their friends. While it is hard to trace the source of these new 1,007 users, we found that 59 of them were friends of users who installed FLM through ads. The remaining users likely found FLM through search tools (e.g., Google Search) or through word-of-mouth based propagation.

Overall, a total of 1,239 users installed our application. For the purpose of this study, we only focus on the 1,165 users (94%) who gave us permission to access all the data we required for our research study.

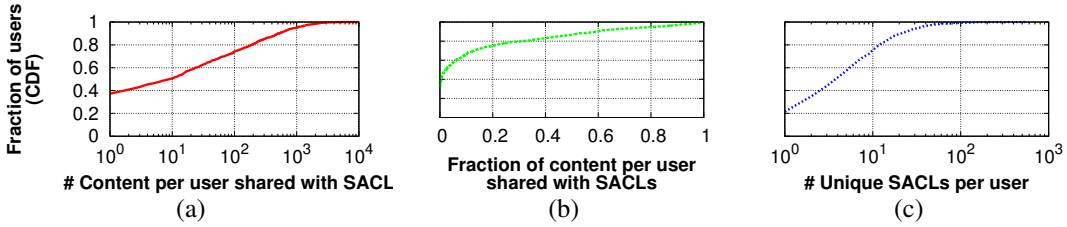


Figure 3.1: (a) Cumulative distribution of number of pieces of content uploaded by users using SACLs. A significant fraction (67.8%) of users in our dataset upload at least one content using SACLs. (b) Cumulative distribution of percentage of content per user shared using SACLs. More than 200 users in our dataset uploaded more than 30% of their content using SACLs. (c) Cumulative distribution of number of unique SACLs specified by each user who upload at least one content using SACLs.

3.2.1.3 User bias

One potential issue with user studies is a bias in the user population. In our case, it is challenging to obtain a random set of Facebook users. This is a fundamental issue with most

user studies, and the common methodology is to carefully characterize the users under study to understand how diverse the users are in terms of demographics. Our user population is by no means random, and we report the demographics and behavior of users below. We believe that the users who installed FLM are those who are interested in creating friend lists to better manage their privacy settings. It is known that Facebook has been promoting friend lists as a way to more efficiently specify privacy preferences [57]. Thus, our user sample is most likely biased towards privacy-aware Facebook users. Additionally, ads can only be shown to users when they are logged in to Facebook; we are therefore likely to get users who are active on Facebook.

3.2.1.4 Ethical considerations

The data we collected in FLM is highly privacy sensitive, and we took great care to respect users' privacy. First, we conducted our study under Institutional Review Board approval. Second, we only report aggregate statistics here, and in any future works. Third, we will never release any non-aggregated data to third parties. All of these steps were also included as part of FLM's Privacy Policy (provided to users when installing FLM).

3.2.2 FLM user demographics

We now examine the demographics of users who installed FLM and allowed us to collect their data. Users usually self-report their location, age, gender and education on their profile. We examine the "current location" attribute to estimate the location of users at a country level and find that 952 (82%) users have provided location information. According to this information, users are from 75 countries covering six continents. There were 19.2% users from North America, 18.1% from Europe and 35.5% from Asia. The top five countries were United States (20%), Pakistan (14%), India (7%), Brazil (7%), and Philippines (6%). Thus, we have users from a diverse set of geographic locations.

Next, we examine the age of users. Our users ranged between 18 and 65 and older, with the median age being 29; this distribution is in-line with the overall U.S. Facebook population [125]. Only 1.2% of users did not specify their gender and for the rest, we observed a strong male bias with 76% of our users being male; this differs from the overall U.S. Facebook breakdown of 47% male [125]. Finally, for the education level (reported by 73.8% of users), 67.8% of users have been to college, while 5.9% of them have been to graduate school. All these statistics demonstrate that we have a diverse set of users in our dataset.

As our users are recruited from a social network, one additional concern is that the users might be a “close-knit" group of friends (in terms of friendship), and not a more general sample of the user population. To evaluate whether this is the case, we check how closely related our users are by examining the number of users who are friends on Facebook, and the number of user pairs with at least one common friend. Out of the 678,030 possible pairs of users [$\binom{1,165}{2}$], 44 (0.01%) were direct friends and 1,266 (0.19%) were not direct friends but had at least one friend in common. Thus, while our population does show some correlation with the social network (unsurprising, given the viral spreading we observed before), the user population is not strongly biased towards one small region of the entire Facebook social network.

Finally, we examine the activity of users in terms of uploaded content. Overall, our 1,165 users have an average of 518 friends (median 332), and uploaded on average 1,040 pieces of content (median 506). 1,003 (84%) users have uploaded more than 100 pieces of content. Only 39 (3.3%) users have uploaded fewer than 10 pieces of content while 3 (0.2%) of them have uploaded none. When we look at activity of users over time, we observe that the activities of our users spanned over 8 years from 2005 to 2013. We further find that 90% of users have been active for more than 20% of weeks since they joined Facebook. Our analysis of users suggests that we have a fairly diverse population most of whom are actively uploading content on Facebook.

3.3 SACL usage

We begin by examining the usage of SACLs by OSM users. Specifically, we investigate how often and for what types of content users specify SACLs.

1. How widely are SACLs used? We first examine how often different users share content with SACLs, using the FLM user set described in the previous chapter. Figure 3.1(a) presents the cumulative distribution (CDF) of the number of content shared using SACLs per user. We observe that a majority of our users are using SACLs for content sharing: 790 (67.8%) users out of 1,165 shared at least one piece of content using a SACL. In total, these 790 users uploaded 212,753 pieces of content using SACLs; this content accounts for 17.6% of all content uploaded by all 1,165 users. In the remainder of the work, we focus only on these 790 users and the content they uploaded using SACLs. We note that the fraction of users using SACLs in our dataset is comparable to that reported for Google+ [94], where 74.8% of the users used SACLs. However, these Google+ users shared significantly more (67.8%) of their content with SACLs. This difference in the percentage of shared contents in Facebook and Google+ is likely due to the differences in user interface between the platforms. We leave a full exploration of the comparative use of SACLs across online social media sites to future work.

Next, we observe that users use SACLs to different extents. In particular, we examine the percentage of content that each user shares with SACLs in Figure 3.1(b) (i.e., for each user, what fraction of their content is shared with a SACL?). We observe a biased distribution across users, but a significant fraction of users select SACLs for much of their content: 20% of users share more than 30% of all their content using SACLs.

Thus, we observe that SACLs are widely used by our users for sharing content, which encourages us to further explore the composition of SACLs and complexity of SACL specification in the following chapters.

Content type	Total content items	Number shared with SACLs	Percent shared with SACLs
Status	786,800	139,112	17.7%
Photo	264,714	45,308	17.1%
Video	111,676	20,880	18.7%
Album	26,527	4,415	16.6%
SWF	9,794	1,554	15.9%
Note	8,500	883	10.4%
Checkin	3,224	548	17.0%
Question	374	25	6.7%
Music	355	27	7.6%
Offer	9	1	11.1%
Total	1,211,973	212,753	17.6%

Table 3.1: Distribution of the number and percentage of content shared with SACLs across different types of content.

2. How many SACLs do users need to create? Having observed that SACLs are widely used, we now investigate how many different SACLs users create from amongst their friends. Figure 3.1(c) shows the cumulative distribution of the number of unique SACLs specified by each user. A large fraction (75%) of the users use less than 10 SACLs, and 20% of the users use only a single SACL. However, there are 5 heavy SACL users, who have used more than 100 unique SACLs. We find that these are heavy users of privacy settings and use different combination of a small number of lists and a set of handpicked friends to specify multiple SACLs for multiple pieces of content. Overall, most users only require a limited number of SACLs to share sensitive content; we leverage this finding later in Chapter 3.5 to reduce the user overhead in specifying SACLs.

3. Does SACL usage vary with content types? Facebook allows users to upload a variety of content types. Table 3.1 presents a breakdown of the total number of content items of different types, and the fraction of those items shared with SACLs. We are interested in understanding if users are biased towards a few types of content when using SACLs. The third and fourth columns of Table 3.1 show the number and fraction of each type of content shared using SACLs. We observe that SACLs are used across all nine different types of

	SACL created using		
	only lists	only friends	both lists and friends
Number of Users	593	555	213
Percent of users using SACLs	75.9%	71.1%	27.3%
Percent of SACLs	33.5%	44.3%	22.2%
Percent of content shared with SACLs	61.4%	27%	11.6%

Table 3.2: Distribution of the number of users using different mechanisms to create SACLs while sharing contents.

content. In fact, 10-20% of almost all types of content are shared with SACLs. Questions and music are shared least often with SACLs; we suspect that these types of content tend to be more public and are usually not privacy sensitive. This widespread use of SACLs across all types of content further justifies looking deeper into SACL membership and complexity, with the goal of increasing the usability of SACLs.

4. How are SACLs created? Facebook users can construct their SACLs in different ways. As mentioned in Chapter 3.1.1, while creating a SACL the user may allow or deny access to individual friends, or lists, or use a combination of friends and lists. Table 3.2 shows the distribution of number of users using different mechanisms to create SACLs. We observe that more than 70% of users are creating at least one SACL by individually selecting their friends and more than 44% of SACLs fall in this category; this is surprising, as selecting friends individually is a somewhat tedious task. Interestingly, only 27% of SACL content is shared with such SACLs; users share the majority of their content with SACLs created using lists.

5. How many users are in SACLs? Next, we examine the size of SACLs (i.e., how many of a user's friends are in different SACLs). Figure 3.2 presents a CDF of fraction of the SACL owner's friends that the SACLs contain. We observe that the distribution exhibits three distinct regions, described below:

1. **Include only few friends:** The first region is highlighted in gray on the left side of the graph; this region contains SACLs with between 0% and 5% of the user's friends. This region contains 25% of all SACLs. In the remainder of the work, we refer to these as include few SACLs.

2. **Exclude only few friends:** The next region is highlighted in gray on the right side of the graph; this region contains SACLs with between 95% and 100% of the user's friends. This region contains 26% of all SACLs. In the remainder of the work, we refer to these as exclude few SACLs.

3. **Include subset of friends:** The final region is in the middle of the graph; this region contains between 5% and 95% of the user's friends. This region contains the plurality (49%) of the SACLs. In the remainder of the work, we refer to these as include subset SACLs.

As we suggest below, the distribution of SACL sizes is very likely influenced by the interface for SACL specification. We use our categorization in the rest of the chapter when we try to characterize the SACL members across different features. However, for exclude few SACLs we also want to see whether we can characterize the excluded friends; for these, we also examine the *excluded members of exclude few SACLs*. The plurality of include subset SACLs shows that our users are not simply including or excluding a handful of their

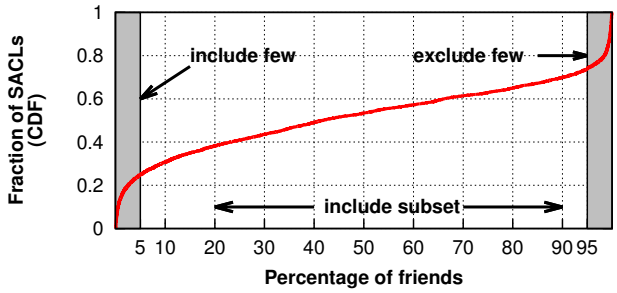


Figure 3.2: Percentage of friends included in SACLs. SACLs in the left gray region are the include few SACLs, SACLs in the right gray region are exclude few SACLs, and the SACLs in middle white region are the include subset SACLs.

friends, but are often including large subsets of their friends. This result further motivates the need to understand how these subsets are selected.

Overall, our observations suggest SACLs are being widely used today by a majority of our users to control access to a non-trivial fraction of their content. SACLs are used at different rates by different users, but they do appear to be used to share many different types of content. Finally, SACLs show wildly different sizes, with many SACLs containing few or almost all of a user's friends. With this understanding, we turn to examine the membership of SACLs in the following chapter.

3.4 SACL membership

We now take a closer look at the membership of SACLs. In other words, are the members of a given SACL distinguishable from the SACL creator's other friends? (e.g., do the members share a profile attribute?) This question is interesting to examine, as any automatic detection of SACL membership would only work if the SACL members were distinguishable. Moreover, existing work [59, 112, 174] hypothesizes that profile attributes, network structure, and user activity can help us to automatically detect clusters corresponding to SACLs; we aim to see if this is true using our dataset of real-world fine-grained privacy settings.

3.4.1 Methodology

Our analysis in this section explores the possibility of characterizing SACL members as a group across three features: (i) profile attributes, (ii) social network structure, and (iii) activity. In other words, we would like to see whether the SACL members form a distinct cluster among the friends of the user. To do so, we form clusters based on these three features and then examine how closely the SACLs of the user match our cluster (e.g., we form a cluster of all user's friends who attended the same high school and then look to see if

this cluster matches any SACL). To compare our automatically detected clusters and the user's SACLs, we address three separate questions:

1. Do the automatically detected clusters match SACLs? Once we have the clusters of friends for a given feature, for each SACL, we try to find the best matching cluster. To compute the "goodness" of a match, we use the F-score metric [109] which provides a measure of detection accuracy. It is computed as the harmonic mean of precision and recall; F-score varies from 0 to 1, with 1 representing a perfect match.

Unfortunately, a low F-score does not necessarily imply that SACLs are not correlated with automatically detected groups. For example, the members of a SACL could be split between two automatically detected groups; in this case, the F-score for both groups would be quite low, but the F-score of the union of the groups would be quite high. Looking deeper into this issue takes us to our next question.

2. How distributed are the SACLs across clusters? In order to check how widely the SACL members are spread across the clusters we use the metric *entropy*[43]. For a given SACL and a cluster c from a set of automatically detected clusters C, we can compute $p(c)$, the probability of a SACL member belonging to c. Then we measure the *entropy* of the SACL as

$$-\sum_{c \in C} p(c) \log_2 p(c)$$

A higher value of entropy signifies more diversity within the SACL members (i.e., they are spread across more clusters).

To be able to compare across the SACLs which belong to different users (with different numbers of clusters and friends per cluster), we normalize the entropy using maximum possible entropy per SACL. A SACL will have maximum entropy when its members are uniformly distributed across all clusters [43]; in this case the entropy will be $\log_2 |C|$, where

$|C|$ is the number of clusters in C. We therefore calculate *normalized entropy* as

$$-\frac{\sum_{c \in C} p(c) \log_2 p(c)}{\log_2 |C|}$$

Normalized entropy for a SACL ranges from 0 to 1. A normalized entropy close to 1 indicates that a SACL is uniformly spread across the maximum number of clusters and a normalized entropy close to 0 indicates that all or most of the SACL members are part of one cluster.

3. How are SACLs different from random groups? One outstanding issue remains: We are examining the entropy of SACLs, but we would really like to measure what's the likelihood of selecting the members of a SACL by pure chance. For example, suppose all of a user's friends attended the same high school; in this case, the "high school" cluster (all friends in a single cluster) would perfectly match any large SACL.

To measure the uniqueness of SACLs relative to random groups, we use the Adjusted Rand Index (ARI) [134] to determine the similarity between SACL and automatically detected clusters. ARI is a similarity metric normalized against chance and varies from -1 to 1. An ARI of 0 indicates no better similarity than a random group, a negative ARI implies worse similarity than a random group, and an ARI of 1 indicates exact similarity. For each SACL, we calculate at the ARI provided by the most similar cluster. If most of the SACLs have ARI close to 0, then the automatically detected clusters are no better in detecting the SACLs than simply using random groups.

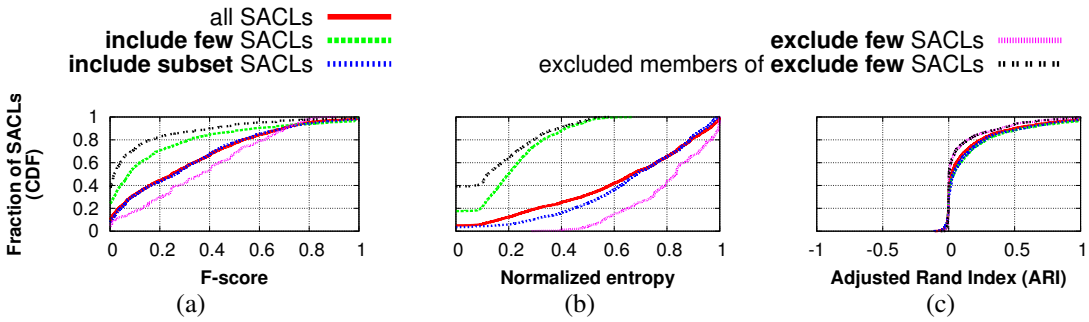


Figure 3.3: **Correspondence between the attribute-based clusters and SACLs, with the cumulative distributions of (a) F-score, (b) Normalized entropy, and (c) ARI. Figure 3.3(a) shows only 15% of the automatically generated attribute-based communities have a F-score of more than 0.6, indicating low number of SACLs showing high match. Figure 3.3(b) shows that larger SACLs are spread across multiple such clusters and have higher normalized entropy. The reverse is true for include few SACLs and excluded members of exclude few SACLs. However, Figure 3.3(c) confirms that more than 40% of these SACLs show better similarity with attribute-based clusters than random groups.**

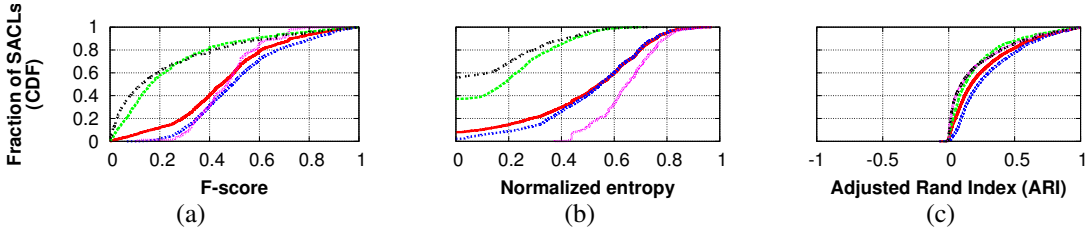


Figure 3.4: **Correspondence between the network communities and SACLs. Figure 3.4(a) shows 21% of network communities have a F-score of more than 0.6, indicating a relatively poor match between network communities and SACLs. Figure 3.4(b) and Figure 3.4(c) confirms that though the larger SACLs have higher entropy (i.e., they are distributed across multiple communities), more than 90% of these SACLs show better similarity with network-based clusters compare to random groups.**

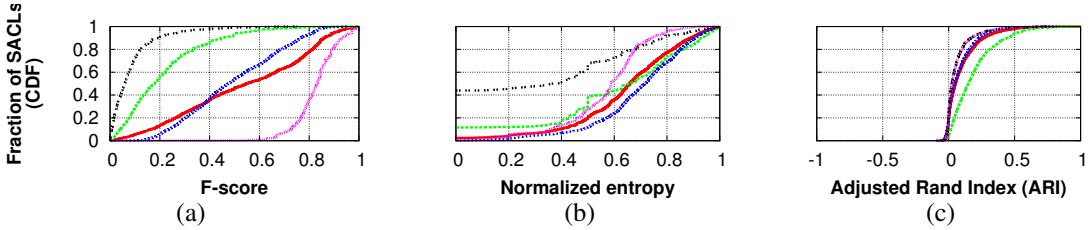


Figure 3.5: **Correspondence between the activity-based clusters and SACLs. Figure 3.5(a) shows 47% of the automatic attribute clusters shows a F-score of more than 0.6, indicating comparatively strong match between activity communities and SACLs. However, Figure 3.5(b) shows that the larger SACLs have higher entropy, and Figure 3.5(c) shows that only 4% have ARI more than 0.3. As a result, random groupings of friends the same size as SACLs would likely show a degree of similar matching.**

3.4.2 SACLs and user attributes

We first explore whether SACL membership is correlated with a common profile attribute. To do so, we leverage the profile attributes provided by Facebook users, focusing on four attributes: workplace, education, current city, and family. We choose these attributes as they have been shown be most strongly correlated with groupings of users in social networks [117]. Using these attributes, we group the friends of a user into clusters who share a common attribute. We report results for all attribute groups in aggregate for brevity; the results are similar when considering each attribute type alone.

We begin by using the F-score metric to check how many of the SACLs exactly match the attribute-based clusters. We present the cumulative distribution of F-scores across all SACLs in Figure 3.3(a). The figure shows that only 15% of all SACLs have a F-score of more than 0.6, indicating a good match for a small subset of SACLs. The result is even worse for very small SACLs (include few or the excluded members of exclude few), with only 10% of such SACLs having a F-score more than 0.6.

We explore the reason for the low F-scores by analyzing the normalized entropy of these SACLs in Figure 3.3(b). The figure shows that the small SACLs have a low entropy (with 20% of include few SACLs with entropy 0) indicating they are mostly part of single attribute-based clusters (this is unsurprising, given that these SACLs are small). On the other hand, the larger SACLs show a high entropy with 35% of exclude few SACLs having an entropy of more than 0.8. These results suggest that our attribute clusters are overestimating the smaller SACLs (indicated by low entropy and low F-score) and underestimating the larger SACLs (indicated by high entropy and low F-score).

Finally, we examine whether SACLs match attribute clusters better than random groups using ARI. As mentioned in Section 3.4.1, an ARI of 0 indicates similarity no better than random groups. Figure 3.3(c) presents the cumulative distribution of ARI across all SACLs;

we observe that 68% of all SACLs have ARI larger than 0, indicating they have more similarity with attribute-based clusters than a purely random set of friends.

Overall, our results suggest that only a small number of attribute clusters serve as a close match for SACLs. However, the SACLs do show some correlation with attribute groups when compared to random subsets of the user's friends. Next, we look into the correlation between SACLs and the social network to see if network-based clusters more closely approximate the SACLs.

3.4.3 SACLs and network structure

In order to explore whether the SACLs correspond to the network structure, we first identity clusters of the user's friends that are tightly connected in the social network (these clusters are often called *network communities*). Specifically, we extract all of the friendship connections between the user's friends, and then use a community detection algorithm that has been shown to work well in grouping a user's friends into a small set of clusters [105]. This algorithm is a combination of a global community detection algorithm [30] and a local community detection algorithm [117] to detect overlapping communities.[5]. Our results were similar with these algorithms, and so we omit the results for brevity.

We begin by examining how many of the SACLs exactly match one of the social network-based clusters. Figure 3.4(a) presents the cumulative distribution of F-scores across all SACLs. The figure shows that 21% of all SACLs have a F-score of more than 0.6, indicating a good match for 21% of SACLs. This is significantly higher than the attribute-based clusters in the prior section, but still does not show a strong correlation.

We next analyze the normalized entropy of these SACLs in Figure 3.4(b). Similar to the attribute-based clusters, the network communities tend to overestimate smaller SACLs and

[5]We note that there are a large variety of community detection techniques in the literature. To make sure our choice of algorithm did not bias the results of our analysis, we performed the same analysis with two additional community detection algorithms [40, 130] similar to ones used in earlier work on unsupervised detection of privacy settings [112, 59]

underestimate larger SACLs, but at a much lower rate. We verify these findings using ARI in Figure 3.4(c). We can observe that 98% of all SACLs have ARI more than 0, indicating almost all of the SACLs have more similarity with community based clusters than a purely random set of friends.

Overall, the network communities show better match with SACLs compared to the attribute clusters. However, still only a small fraction of SACLs have strong correlation with network communities, making it unlikely that network communications could be used to infer SACLs for sharing content. Next we will look into the correlation between activity-based clusters and SACLs.

3.4.4 SACLs and user activity

For our final user feature, we examine whether the membership of SACLs is correlated with the strength of the link between the user and their friends. As a proxy for link strength, we use *activity*; this is a common way to estimate how closely connected two users are [19]. For each user, we collected data about four different types of interaction between the user and their friends: (i) posting on the user's wall, (ii) liking the user's posts, (iii) commenting on the user's posts, and (iv) being tagged in the user's status and photos. We observe that 94% of the users who used SACLs have at least one such interaction with their friends.

Using this data, we cluster each user's friends by activity (i.e., frequency of interaction) and see whether the activity-based clusters matched the SACLs. We use the same algorithm as prior work [19] to find the activity clusters. The algorithm is essentially a k-means algorithm modified to automatically find the optimal number of clusters. As a result, all friends with a similar number of interactions will be put in one cluster. After running the algorithm, we find that the median number of clusters across all users is four.[6]

[6]Interestingly, this observation matches Dunbar's sociological study [51, 84] where the number of Dunbar's circles, the number of activity-based clusters in people's offline network is also four.

As before, we begin by examining the cumulative distribution of the F-score in Figure 3.5(a); we observe that 47% all SACLs have a F-score of greater than 0.6. This is even better than network-based communities. The larger SACLs (e.g., exclude few SACLs) show an even stronger match with high F-scores, but the match is considerably worse for smaller SACLs (e.g., include few SACLs), with only 3% of such SACLs having a F-score more than 0.6. Closely examining the activity-based clusters, we hypothesize that our method of creating activity communities often results in creating a large cluster containing all friends with low levels of interaction with the user. As a result, this single, large community alone overlaps with the large SACLs considerably, making their F-score quite high.

To confirm this hypothesis, we also calculate the cumulative distribution of normalized entropy (Figure 3.5(b)) and ARI (Figure 3.5(c)). We find a poor match between SACLs and activity-based clusters using both pieces of analysis; the ARI values for the SACLs are very close to 0 for almost all activity-based clusters (e.g., only 8% of the SACLs have ARI more than 0.3). This finding confirms that any random groups of the size of larger SACLs will show the same level of similarity with the activity-based clusters, thus making the clusters a poor mechanism for approximating SACL membership.

3.4.5 Summary

In this section, we examined the membership of SACLs by trying to correlate SACL members with attribute, network structure, and activity-based clusters. Our results show that very few of these clusters show a significant correlation with SACLs, suggesting that automatically detected SACLs-based on these features are unlikely to be very accurate. This finding is in opposition to the results from prior work [59, 112, 174], which suggest that it is possible to use automatically detected clusters to create SACLs. We believe this difference is due to the fact that these prior works were not able to evaluate their proposals against ground-truth SACLs. In fact, others have also found [97] that users are able to group their friends in meaningful groups, but find it difficult to choose the right group to share content

Figure 3.6: Facebook's interface for specifying SACLs.

with. Consequently, we explore alternative approaches to increase the usability of SACLs in the next section.

3.5 SACL specification

We have observed that SACLs appear to be quite difficult to infer automatically. We now examine the "overhead" (i.e., the amount of work that users must perform) in order to specify SACLs today. Then, we explore how we can increase the usability of these SACLs by reducing the user overhead and making SACLs easier to use. If successful, these approaches would make privacy easier for users to manage, thereby increasing the usability of OSMs in general.

3.5.1 SACL specification overhead

The act of specifying a SACL—choosing which friends to share content with—induces cognitive overhead on the user. While there may be multiple dimensions of this overhead (e.g., deciding whether to include a specific user, using the interface, etc), many of these are quite challenging to measure. As a first step, we define the *SACL specification overhead* to

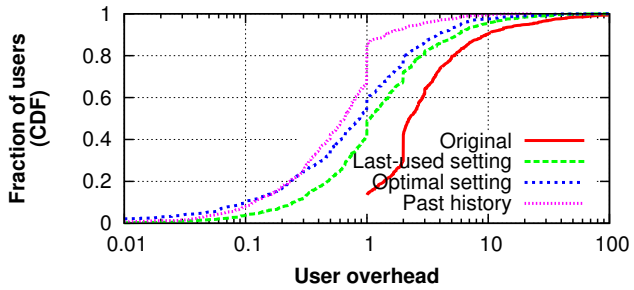


Figure 3.7: Cumulative distribution of overhead for specifying SACLs. Shown are the distributions of measured SACLs (Original), measured SACLs taking into account Facebook's last-used setting (Last-used setting), the optimal overhead for measured SACLs (Optimal), and the overhead of our proposed mechanism of presenting the user with the last 5 SACLs (Past history). Our proposed mechanism shows a substantial reduction in user overhead.

be the number of *terms* used to specify a SACL. Our reasons for doing so is that Facebook allows users to specify SACLs using an allow/deny interface, where users can select friends or lists to allow or deny access (a screenshot of Facebook's interface is shown in Figure 3.6). Thus, the amount of work the user has to do is proportional to the number of friends/lists that the user selects to allow or deny. Of course, we recognize there are dimensions of overhead that this measure fails to capture; we leave the task of characterizing those dimensions of user overhead to future work.

As an example, consider the screenshot shown in Figure 3.6. In this example, the user is choosing to allow friends Bob, Carol, the list Football Friends, and the Facebook list Family. The user is also choosing to deny the list Drinking Buddies. As a result, the SACL specification overhead for this SACL is five (A total of five terms appear in the allow and deny settings.) It is important to note that the size of the SACL is different from its specification overhead: Consider the case of a user only denying access to a single friend. In this case, the specification overhead is low, but the SACL has many users in it.

We define the *average user overhead* as the average of all SACL specification overheads for content shared by a given user. Formally, if a user specifies access to her content

$\{c_1, c_2, ...c_n\}$ with privacy settings $\{p_1, p_2, ...p_n\}$, the average user overhead for this user is

$$\frac{\sum_i |p_i|}{n}$$

where $|p_i|$ denotes the SACL specification overhead of setting p_i. Note that the p_i settings are not necessarily distinct, as multiple pieces of content may be shared with the same SACL. The best-possible average user overhead is 1, meaning the user only used the SACLs with a single term when sharing her content.

We present cumulative distribution of the average user overhead in Figure 3.7 with the line “Original". We observe that users do have significant overhead when specifying SACLs: 86% of users have an overhead more than 1, and there are more than 150 users with overhead more than 5. This suggests there is significant potential to reduce the SACL specification overhead for users, making SACLs more usable.

3.5.2 Last-used setting

Facebook’s default privacy setting for content is set to select the *last-used* privacy setting [65]. So, if a user selects a SACL for a newly uploaded piece of content, all future pieces of content will be shared with the same SACL until the user chooses a different privacy setting. We therefore modify our definition of average user overhead to capture this behavior; if a user selects the same SACL as the previous piece of content, we define this SACL specification overhead to be 0. As a result, a user’s average user overhead may be less than one.

We present the cumulative distribution of the average user overhead, taking into account the last-used setting, in Figure 3.7 with the line “Last-used setting". We immediately observe a significant reduction in the measured overhead, which we believe more accurately captures the work a user must do. The figure shows that this simple technique of using last setting as the default can significantly reduce the user overhead: this technique lowers the overhead by more than 50% for almost half (48%) of the users. Thus, Facebook’s choice to enable default

last-used settings is useful in reducing user overhead. For the remainder of our analysis, we use average user overhead, taking into account the last-used setting, as our baseline.

3.5.3 Optimal overhead

It is important to note that there may be multiple ways of specifying a given SACL: For example, a user could specify the SACL by only allowing the friends in the SACL. Or, the user could use an existing list, and exclude the users not allowed to access the content. Or, the user could allow all friends, and then deny only the friends who should not be able to access the content. We now examine how close the user's chosen specifications are to the *optimal specification*, in terms of the SACL specification with the minimum overhead.

To do so, for each SACL we observe, we determine the optimal specification overhead.[7] We then present the cumulative distribution of average user overhead in the optimal case in Figure 3.7 with the line "Optimal setting". We observe that there is still room for improvement from using the last-used setting alone; many users could express their SACLs in a manner than involves fewer terms.

3.5.4 Using past history

In this section, we explore a generalization of the last-used setting, with the goal of further reducing the average user overhead. The results in Section 3.3 suggested that there are certain SACLs that users select to share content with a significant fraction of the time. Figure 3.8 plots the cumulative distribution of the percentage of content shared with the top k most frequently used SACLs for each user. For example we can see that if we allow each user to use their top 5 SACLs, this would cover over 80% of the content for the vast majority (90%) of users.

[7]Note that computing the optimal overhead of a setting is a modified version of the NP-hard set cover problem [96] where the setting is the universe and lists and individual friends are subsets of the universe. We use a brute force solution to the problem, which is feasible due to small number of subsets in this case.

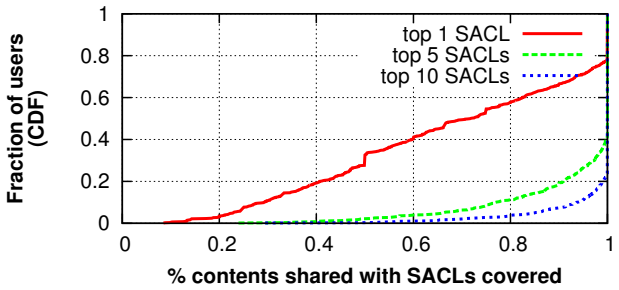


Figure 3.8: Cumulative distribution of percentage of content covered by the top k SACLs. Even if we set k=5, most of the content for the majority of users are covered.

This observation means that we may be able to significantly reduce the average user overhead by allowing users to choose from their *k most used* SACLs, rather than just the last-used SACL. To do so, we calculate the average user overhead, assuming one would have made it possible for the user to directly use the top 5 most frequently used settings while sharing content (Should the user re-use these settings, we calculate the overhead as 0). A cumulative distribution of the resulting overhead is shown in Figure 3.7 with the line "Past history". We immediately observe a dramatic reduction in user overhead (In fact, the overhead is lowered for 86% of users).

In summary, this approach of leveraging past history has the potential to significantly reduce the user overhead in specifying SACLs. An OSM operator can create these SACLs based on user's past history, and provide them as options to select from, when the user uploads a new piece of content. Should the user select one of the previously-used SACLs, it will reduce their overhead and make privacy specification more usable.

3.6 Discussion

OSMs are increasingly popular and users are sharing ever more content on these services. In this work, we focused on the most privacy-sensitive of these content: the content with hand-crafted privacy settings selected by the users to effectively control exposure of the content. We found that these SACLs are surprisingly common (over 17.6% of all content

is shared with SACLs), but that the membership of these SACLs shows relatively little correlation with the profile attributes, the social network structure, or the activity level of the members. As a result, there appears to be little hope of automatically detecting more than a few of these SACLs. We also found that the act of specifying SACLs is often complicated for users, but a simple technique like remembering a few of the most frequently used SACLs is likely to significantly reduce this burden in practice.

However, much work remains to be done. In the remainder of this section, we discuss a few of the limitations of our study, as well as future directions for exploration.

Understanding motivation for SACLs Our study explores the use of SACLs, but does not reveal *why* users create SACLs or how they choose the content to share with SACLs. Possible reasons include dissatisfaction with the default privacy settings, the sharing of highly privacy sensitive content, or using SACLs as a mechanism to choose the audience for a particular content.

Moving target We quantified the way users create in-use SACLs today, but Facebook is known for changing their privacy interface over time [63]; these changes are likely to impact the usage of SACLs for individual users. We aim to repeat our analysis as Facebook makes these changes, hoping to capture resulting changes in user behavior.

SACL accuracy It remains an unexplored question as to which of the friends users would *ideally* want to share their content with (i.e., who does the user want to be in a SACL, regardless of who is actually in the SACL). Prior work has shown that users often misunderstand other Facebook privacy settings [103], and we suspect that this would likely hold true for SACLs as well.

SACL overhead In our calculation of overhead, we took into consideration the number of terms specified by users explicitly while specifying SACLs, where a term can be a friend list or an individual friend. However, this quantification does not directly account for the mental effort required for a user when creating SACLs (e.g., certain SACLs may be easier or harder to create, even if they have the same number of terms). We leave a full exploration of this

effect (possibly via detailed debriefing interviews of a small sample of Facebook users) to future work.

After, exploring how users can control exposure of their content using SACLs and devising a way to simplify the SACL specifications using simple caching, we now move into another mechanism to control exposure. In the next chapter, we check how OSM designers can help OSM users to control exposure by limiting large scale social data aggregators like Spokeo.

CHAPTER 4

Controlling exposure by limiting large-scale social data aggregators

We described a generic framework (in Chapter 2.5) for controlling exposure in OSMs—predicting future exposure of a content and limiting actual exposure to the expected value. A simple application of this generic framework is limiting the privacy violation due to third party crawlers. OSMs allow users to browse the (public) profiles of other users in the OSM, making it easy for users to connect, communicate, and share content. Unfortunately, this functionality can be exploited by third-parties to aggregate and extract data about millions of OSM users. Once collected, the data can be re-published [83], monetized, and mined in ways that may violate users' privacy. For instance, it has been shown that private user attributes like sexual orientation can be inferred from a user's set of friends and their profile attributes [83, 117]; a third party with access to aggregated user data could easily apply these techniques.

These third-party aggregators, which we refer to as *crawlers*, represent a significant problem for OSM operators as well. User data provides OSM operators with a revenue stream (e.g., via targeted advertisements); stopping crawlers is therefore in the OSM operators' business interests. Additionally, OSM operators cannot ensure that data collected by a third party is used according to the operator's privacy policy. Yet, OSM operators are likely to be held responsible if crawled data is used in ways that violate the policy, at least

in the court of public opinion. For example, Facebook was widely blamed in the popular press [47] when a single crawler gathered public profiles of over 100 million users [83].

To that end, OSM operators can predict with high confidence that these third party crawlers are not in the user's expected set. Thus, OSM operators need to design effective mechanisms to thwart large-scale crawling of OSMs [148] in order to help users control exposure of their content.[1] Today, OSM operators typically limit the rate at which a single user account or IP address can view user profiles [150], in order to discourage large-scale data collection. Unfortunately, crawlers can circumvent these schemes by creating a large number of fake user accounts [7], by employing botnets [88] or cloud services[36] to gain access to many IP addresses, or by using the compromised accounts of a large number of real users [4].

In this work, we propose Genie, a system that OSM operators can deploy to limit large-scale crawlers and control exposure. Genie leverages the differences in the browsing patterns of honest users and crawlers to effectively thwart large-scale crawls of user profiles. Genie's design is based on the insight that honest users tend to view the profiles of others who are well connected and close in the social network (i.e. network induced in OSMs via friendship relations). A crawler, on the other hand, is limited in his ability to form or control enough links to be close to all users whose profiles he wishes to view. Genie exploits this fact by enforcing rate limits in a way that is sensitive to the distance and degree of connectivity between viewer and viewee.

Using profile view data from RenRen, a Facebook-like OSM that is popular in China [138], we observe that average social network distance between honest users and the profiles they view tends to be low (1.62 social network hops). We demonstrate that a crawler, on the other hand, would require a very large number of well-distributed accounts (e.g., controlling 3%

[1]Not all large-scale crawls are for nefarious purposes. Researchers, for instance, already tend to obtain the consent of the OSM operator before doing their crawls for research purposes (e.g., [91, 37]); such authorized crawlers can be whitelisted even if a defense is in place.

of all existing accounts in OSM) to be able to view the profiles of all users while maintaining the same low average distance.

Genie works by deriving a credit network [44, 50, 75] from the social network. In brief, credit is associated with links in the social network, and a viewer must “pay" credits to the viewee, along a path in the social network, when viewing a profile. Compared to conventional per-account or per-IP address rate-limiting, credit networks offers two key advantages. First, in a credit network, the rate limits are associated with social network links rather than user accounts or addresses. As a result, a crawler gains little by creating many Sybil accounts or using many IP addresses [91]. Second, the greater the distance between viewer and viewee in the social network, the stricter the rate limit is imposed by the credit network on profile views. Consequently, even crawlers with access to a relatively large number of compromised user accounts are unable to crawl the network quickly.

The contributions of this work are as follows:

- We analyze profile viewing data from the Renren social network and show that the average distance in the social network between a honest viewer and a viewee is significantly smaller than that of a crawler. Moreover, a crawler interested in crawling the entire social network is fundamentally unable to blend in with honest viewers unless he controls a very large proportion of strategically positioned user accounts.
- We present the design of Genie, which leverages credit networks derived from the already-existing social network to block large-scale crawling activity, while allowing honest users’ browsing unhindered.
- We demonstrate the feasibility of deploying Genie with an evaluation using large partial social network graphs obtained from Renren, Facebook, YouTube, and Flickr, and a mix of real and synthetically generated profile viewing traces. We demonstrate that Genie effectively blocks crawlers while the impact on honest users is minimal.
- We show that Genie can scale to networks having millions of nodes by scaling up credit network operations. Thus, Genie is practical on the large OSMs of today.

4.1 Related work

In this chapter, we describe relevant prior work on limiting large-scale crawls of OSMs and leveraging social networks to defend against Sybil attacks.

Limiting large-scale crawlers There exist a number of techniques that aim to prevent crawls of web services. Two techniques commonly used in practice are `robots.txt` [8], and IP-address- or account-based rate limiting [135, 80, 161].

`robots.txt` is a file stored at the web server that indicates a set of pages that should not be crawled. Compliance with this policy is voluntary; `robots.txt` consequently provides little defense against malicious crawlers.

Large websites like Yahoo! often rely on a simple per-IP-address rate limit to control access to their web services [135]. Each IP address is allocated a maximum number of requests, which are replenished in 24 hour intervals. Once a user exceeds this limit, the operator either stops serving the user or may require that the user solve a CAPTCHA [9]. This approach limits the number of views a crawler can perform from an individual IP address, but is not effective against botnets that control many IPs. Additionally, dedicated crawlers can bypass defenses like CAPTCHA using available CAPTCHA-solving service providers [120], and other schemes exist that can bypass IP-address-based rate-limiting approaches [36, 77].

Online social media sites like Facebook, Google Plus or Twitter [150, 80, 161] often use account-based rate limits on requests to view profile pages. Similar to IP-based rate limits, this approach works well if crawlers control at most a small number of accounts; in the face of Sybils or compromised accounts, it is not effective.

Wilson et al. proposed SpikeStrip [173], a system designed to discourage OSM crawlers. SpikeStrip uses cryptography to make information aggregation from OSM websites inefficient. SpikeStrip rate limits the number of profile views allowed per browsing session and

prevents different browsing sessions from sharing data. Thus, crawlers cannot aggregate or correlate data gathered by different sessions.

Despite its elegant design, SpikeStrip restricts the functionality of the OSM. For example, SpikeStrip does not allow two OSM users to share website links of a common friend. Moreover, SpikeStrip would require OSMs like Facebook to change the way they use content distribution networks like Akamai to serve users' content. Unlike SpikeStrip, Genie does not affect the OSM functionality or content distribution. As we will show later, Genie can be deployed with minimal disruption to the browsing activities of honest users.

Social network-based Sybil defenses Recently there have been proposals to leverage social networks to defend against Sybil attacks [157, 178, 177, 45, 156, 133, 131, 116].

Sybil defense proposals such as SybilLimit [177] or SybilInfer [45] try to detect Sybils in the network and then block them from the service. However these Sybil detection mechanisms are not designed to address compromised accounts. Additionally, recent work has shown that these Sybil detection schemes suffer from limitations due to assumptions they make concerning the structure of the social network [167, 118].

The design of Genie borrows credit network [44, 50, 75] techniques from Sybil-tolerant [165] systems like Ostra [116] and Bazaar [131], and uses Canal [166] to manage credit network operations. In contrast to Sybil detection schemes, Sybil tolerant systems do not aim to detect Sybil users; instead they minimize the impact of Sybils on honest users.

Genie differs from existing Sybil tolerant systems in two fundamental ways: First, Genie considers crawlers that have access to both Sybil and compromised accounts; previous work considered only Sybil attacks. Second, unlike Ostra and Bazaar (which rely on users to provide feedback on whether a communication is spam or whether a transaction is fraudulent), Genie infers whether or not activity is malicious by exploiting differences in the browsing patterns of crawlers and honest users.

4.2 System and attack model

4.2.1 System model

OSMs such as Facebook, Renren [138], Google+, and Orkut share a common system model: Users create accounts, establish friend links with other users, and post content (often of a personal nature). Users have a "home page" on the OSM that links to all of the user's content; we refer to this as the user's *profile*. The graph formed by the entire set of friend links forms a social network. In the OSMs of interest to us, forming a friend link requires the consent of both users.

Users can typically choose to make their data *private* (i.e., visible only to the user and the site operator), *public* (i.e., visible to every user of the social network), or *semi-public* (i.e., visible to subsets of the user's friends or to friends of user's friends). In practice, many users choose to make their profile information public [103], despite the private nature of some of the information posted. Contributing to this choice may be that sites encourage public sharing [60], that the default privacy setting is "public" [132], that other privacy choices are not always intuitive [151], and that many users are not fully aware of the privacy risks [103]. It is these public profiles (that typically represent a large fraction of all user profiles [98]) that Genie is concerned with protecting.

4.2.2 Attack model

Today, social networking sites tend to impose a rate limit on the profile views a single user can request, in order to slow down crawlers. However, there are two ways in which a crawler can overcome these limits.

A crawler can conduct a *Sybil attack* by creating multiple user accounts, thereby overcoming the per-user rate limit. It is important to note that while the crawler can create an arbitrary number of links between Sybil accounts he controls, we assume that his ability to

form links between his Sybil accounts and honest users is limited by his ability to convince honest users to accept a friend link, regardless of how many user accounts he controls. The significance of this point will become clear in the following chapter, where we describe how Genie leverages social links to limit crawling activity.

A crawler can also conduct a *compromised account attack* by taking control (e.g., by obtaining the password) of existing accounts in the system. The crawler can gain access to such accounts via phishing attacks, by guessing the user's password, or by purchasing the credentials of already compromised accounts on the black market. A compromised account attack is more powerful than a Sybil attack, because every additional compromised account increases the number of links to honest users that the crawler has access to. Again, the significance of having access to such links will become clear in the following chapter.

We assume that a crawler with access to compromised accounts cannot compromise the accounts of strategically positioned users of his choosing. Defending against a crawler who can access any account of his choosing would require preventing social engineering attacks, which are outside Genie's attack model. Instead, we assume that compromised accounts are randomly distributed throughout the network. Additionally, we assume that the crawler does not actively form new links involving compromised accounts, as such activity would likely alert the actual owner of the account that their account has been compromised.

We are concerned about attacks where the crawler greedily attempts to gather as many distinct user profiles as possible. We assume that the crawler is agnostic to *which* users he crawls. Our crawler model captures both third-party crawlers [6] as well as research-oriented crawlers (e.g., used in studies of Facebook [91], Flickr [115], and Twitter [37]). However, our model excludes some crawlers that may be interested in repeatedly crawling the accounts of a small subset of users over an extended period of time, perhaps to gather their changing profile information. We make no assumptions about the crawler's strategy, i.e., whether the crawler employs random walks or BFS or DFS to fetch user profiles. Consequently, we

simulate attacks employing the strategy that optimizes for the crawler's goal of fetching as many distinct user profiles as possible.

4.3 Workload analysis

In this chapter, we compare profile viewing workloads of honest users and crawlers. In later chapters, we show how Genie can exploit the differences in the browsing behavior of honest users and crawlers to rate-limit crawlers, while rarely affecting honest users.

4.3.1 Honest users' profile viewing workload

We obtained anonymized user profile browsing data [91] from the RenRen social network [138], a Facebook-like social network that is popular in China. The data covers users in RenRen's Peking University (RR-PKU) network, and includes the links between PKU users and all other RenRen users. We pre-processed the social network and browsing trace to only include the subgraph of the PKU users, and then extracted the largest connected component (LCC) from the social network. Similar to prior work [116, 166], our analysis only examines users in the LCC (representing 91.2% of the users and 94.3% of the links from the pre-processed network). The LCC of the RR-PKU network has 33,294 users and 705,248 undirected links.

The data set also includes a trace of all profiles (both friends and non-friends) that each user browsed during a two-week period during September, 2009. Unfortunately, the RenRen trace does not provide timestamps or an ordering for profile views. Therefore, in experiments where we need the profile views to be ordered, we generate a time series by assigning each profile view a timestamp chosen uniform randomly within the two-week period covered by the trace. This time series was used in all analyses conducted in the work. We highlight our key findings below.

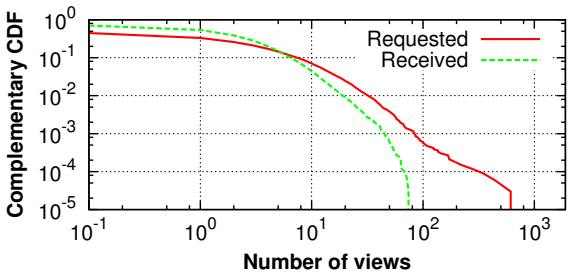


Figure 4.1: Complementary cumulative distribution of the number of profile view requests made and received by RR-PKU users during a two-week period.

1. Most users make (receive) few profile views, but a small number of users make (receive) a large number of views. Figure 4.1 shows the distribution of profile views made and received by individual users in the RR-PKU network. The plots show a considerable skew in the distributions: Most ($> 90\%$) users make or receive fewer than 10 views, while a handful of users ($< 0.4\%$) view 50 or more profiles. In particular, there are three users who viewed 1,827, 612 and 272 profiles (respectively) over a period of two weeks. These users show significant crawler-like behavior, and we return to discuss these users in Chapter 4.5.4. Thus, most users in the social network tend to make or receive views from a small number of other users in the network.

2. The number of profile views made (received) by users is significantly correlated with their degree The Pearson correlation coefficients between the rankings of users ordered based on the number of views they make (receive) and their degree is 0.67 (0.5). The high correlation coefficient affirms the intuitive hypothesis that users who are more active in the social network would also have more friends in the network. This finding is consistent with the profile browsing behavior previously studied in Facebook [22] and Orkut [26].

This observation suggests that the *imbalance* in user activity, measured as the magnitude of the difference between the number of profile requests made and received by individual users, divided by their degree would be tightly bounded. Intuitively, this can be explained as follows: both the number of profile requests made and received by a user tend to rise or fall with the user's degree. So, the difference in requests made and received,

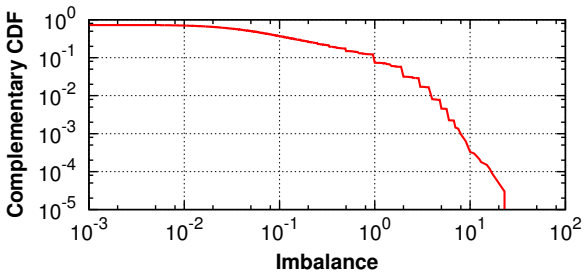


Figure 4.2: Complementary cumulative distribution of the imbalance of RR-PKU users.

divided by the user's degree would be expected to be small. Figure 4.2 confirms this by plotting the complementary cumulative distribution of the number of *unbalanced views* (i.e., $|Req\_made - Req\_received|$) per friend link for different users. Almost 99.9% users have an imbalance in viewing activity per friend link lower than 10.

3. Not all profiles views are unique; a small but non-trivial number of views are repeated. Users tend to repeatedly visit the profiles of others to track updates. In our two-week trace, we found 17,307 (17.8%) of the profile views to be such repeat views. Our estimate of the fraction that repeat views represent is likely to be conservative, as we are restricted to a two-week trace. (One would expect the percent of repeat views to increase with the length of the workload trace.) However, the implication of the presence of repeat views is that repeat views decrease the number of distinct profiles viewed by users. For crawlers, repeat views represent sub-optimal use of resources, as their goal is to view the profiles of as many distinct users as possible.

Network	Users	Links	Average degree
RR-PKU [91]	33,294	705,262	21.2
Facebook [164]	63,392	816,886	25.7
Youtube [115]	1,134,889	2,987,624	5.2
Flickr [114]	1,624,991	15,476,835	19.0

Table 4.1: Number of users, links, and average user degree in the large-scale social networks used to evaluate Genie.

Figure 4.3: Cumulative distribution of hop distance separating viewers and viewees in RR-PKU network. The profile viewing activities are highly local.

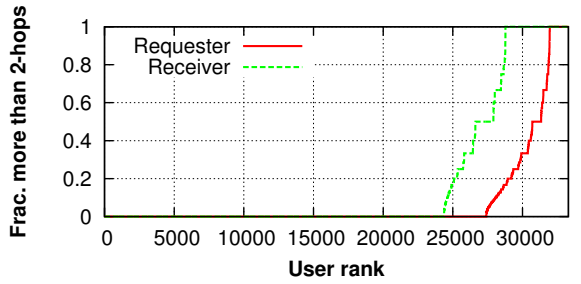


Figure 4.4: Fraction of requested and received views that lie beyond 2 hop-distances for individual users. For most users, only a small fraction of their views requested or received lie beyond 2-hop neighborhood.

4. Users tend to make (receive) profile views of others who are within their immediate (1 or 2-hop) network neighborhood. Figure 4.3 shows the distribution of network distance, measured in terms of hops, between the viewers and viewees in our RR-PKU trace. We observe that over 80% of all profile views are between users who are separated by two hops or less. This observation is consistent with prior studies of the Orkut social network [26], as well as studies of friendship formation in the Flickr social network [114]. Figure 4.4 shows the distribution of network distance for profile views made by individual users. For most users, only a small fraction of profiles they request, as well as a small fraction of users who request their profiles, lie beyond their 1-hop (direct friends) and 2-hop (friends-of-friends) neighborhood.

Our analysis considers only a single network, due to the difficulty in obtaining detailed profile viewing data. However, we note that many of our findings are consistent with prior

Network	Number of compromised accounts			
	1	10	100	1000
RR-PKU [91]	26	415	3,638	14,938
Facebook [164]	13	237	2,123	15,970
Youtube [115]	6	26	592	4,129
Flickr [114]	5	613	2,242	16,015

Table 4.2: Strength of crawlers in different social graphs. Each column corresponds to specific number of random compromised accounts, and the corresponding number of attack links in different graphs.

studies of other social networks [26, 22, 114, 172], suggesting that our RenRen data is likely to be representative of other social networks.

4.3.2 Crawlers' profile viewing workload

We now turn to examining the workload that a crawler would generate when run on the RenRen network from the previous chapter and three others: the Facebook New Orleans regional network [164], YouTube [115], and Flickr [114]. Table 4.1 provides more details on the number of users, links, and average degree in these networks.

We simulate crawlers of varying strength by allowing crawlers to "compromise" 1, 10, 100, and 1,000 randomly chosen users within the network. Table 4.2 shows the number of links that connect user accounts under the crawlers' control to honest users in the different graphs. Note that while a crawler with 1,000 compromised accounts might not seem particularly strong, it is important to consider the size of the networks. For example, the crawler controlling 1,000 accounts controls around 0.1% of all users in the YouTube network. As a point of reference, this would be equivalent to controlling 1,000,000 compromised accounts in the real-world Facebook network.

To generate the crawling workload, we implement the following strategy for the crawler: We assume that all crawler's accounts collude to view profiles of all honest users in the network. Each honest user is viewed only once, and the crawler's account nearest to an honest user will be assigned the task of viewing that user's profile. This strategy maximizes

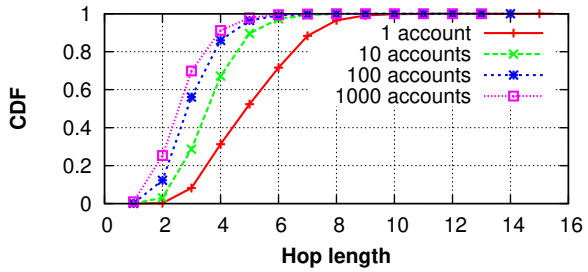


Figure 4.5: Cumulative distribution of the distance of crawler profile views in Flickr with different numbers of compromised accounts. To fully mimic the high locality in honest user views the crawlers have to control more than 1,000 compromised honest user accounts.

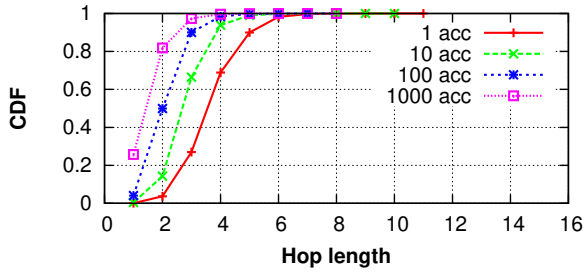


Figure 4.6: Cumulative distribution of the distance of crawler profile views in Facebook with different numbers of compromised accounts.

locality in profile views, making the crawler's workload look as "close" to the honest users' workload as possible.

We now compare the resulting crawler workload with that of honest users, noting two important differences.

1. In contrast to honest users' workload, profile views by crawlers are highly non-local. Figure 4.5 and Figure 4.6 shows the locality in profile visits by crawlers with different attacking capacities for the Flickr and Facebook graphs (the other graphs show similar behavior and are removed for brevity). For large network graphs like the Flickr and YouTube samples, we observe that even a powerful crawler with 1,000 accounts has only a small fraction (less than 30%) of requested profiles within the 2-hop neighborhood of the users under his control. For small network graphs like the Facebook and RenRen samples, the crawler does have a majority of users (around 80-90%) within a 2-hop neighborhood. However, in

these networks, 1,000 compromised accounts represent 3% of all users; considering that controlling 3% would require controlling 29 million user accounts in the current complete Facebook network, this is a very powerful attack indeed.

Overall, the results indicate that to mimic the high locality in profile views for honest users, crawlers would fundamentally have to control a very large fraction of all accounts.

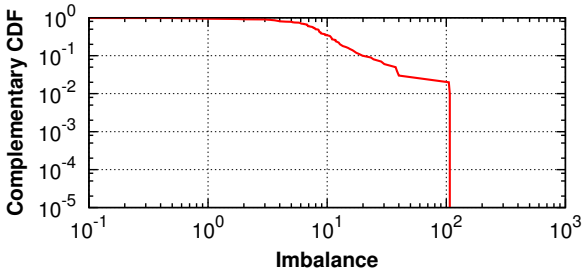


Figure 4.7: Complementary cumulative distribution of the imbalance of RR-PKU users views per link for crawlers in RR-PKU.

2. In contrast to honest users, crawlers request many more profile views than they receive. We observe that the median imbalance per link in profile views (shown in Figure 4.7) is 8.3 for a crawler with 100 user accounts, compared to honest users' median imbalance of 0.05. Such an imbalance is necessary, as even with 100 accounts, the crawler makes significantly more profile views that an honest user. In the next chapter, we present a system design that exploits these observed differences in browsing patterns of honest users and crawlers.

4.4 Genie design

In this chapter, we present the design of Genie, analyze its security properties, and discuss how Genie can be used to maliciously deny service to honest users.

4.4.1 Strawman

To motivate the need for Genie's more elaborate approach, we briefly consider a simple distance-based rate-limiting technique as a strawman design, and show that it is ineffective against Sybil crawlers. We know from Chapter 4.3.1 that honest users rarely view profiles outside their neighborhood social graph, whereas the crawlers have to view distant profiles. Our strawman uses distance based rate limiting to leverage this finding.

In the strawman design each user account is allowed to view user profiles at a maximal rate r, where viewing a profile K hops away counts as viewing $K - 1$ profiles. (Thus, viewing a friend's profile is not subject to rate-limiting.) The scheme discriminates heavily against crawlers, who tend to view distant profiles. However, just like the existing rate-limiting schemes discussed in Chapter 4.1, this design is vulnerable to Sybil attacks. A crawler can simply create more Sybil accounts to overcome the per-account rate limit. The same would be true for any per-account or per-IP-address rate-limiting approach, no matter how much it discriminates against workloads typical of crawlers.

To summarize, if the profile viewing privileges are assigned based only on the viewing user, then a crawler can view more profiles simply by creating additional Sybils. Instead, Genie attaches the profile viewing privileges with *paths* in the network, rather than users, as we describe in the following chapter.

4.4.2 Genie design overview

Now, we present the design of Genie. Genie relies on a credit network [50, 75] to make sure Genie's rate limits cannot be circumvented by using more Sybil accounts. Moreover, Genie uses the credit network to impose rate limits that discriminate against a crawlers' workload, in order to slow down powerful crawlers that use many compromised user accounts.

A credit network is a directed graph $G = (V, E)$, where each edge $(u, v) \in E$ is labeled with a scalar credit value $c_{uv} \geq 0$. Each node in Genie's credit network corresponds to an

OSM user and there is a pair of directed edges $(u, v), (v, u)$ in the credit network iff the users u, v are friends in the OSM. Genie allows user s to view user t's profile information iff the max-flow between s and t in the credit network is at least $f(d_{st})$, where f is a non-decreasing cost function, and d_{st} is the length of the shortest path between s and t in the social network.

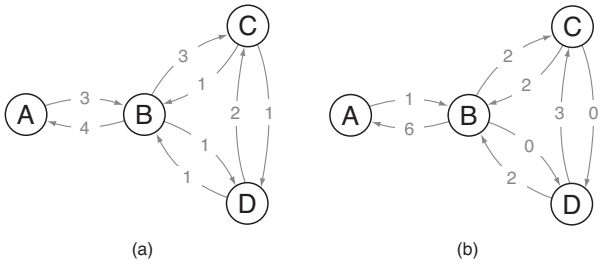


Figure 4.8: Example of Genie on a small credit network. (a) User A wishes to view D's profile; let us assume this costs two credits. No single path has two credits available, so both $A \rightarrow B \rightarrow C \rightarrow D$ and $A \rightarrow B \rightarrow D$ are debited 1 credit. (b) The state of the credit network afterwards; note that intermediate nodes B and C maintain the same total credit available, as debiting from one link automatically adds credit to the opposite link.

Thus Genie computes the *amount of credit charged* for a profile view based on the shortest path length between the viewer and viewee users. However, the credits can actually be exchanged over any set of paths between the viewer and viewee. For example, Figure 4.8 (a) shows an example of Genie in a small network, where user A wishes to view D's profile. A is charged based on the shortest-path distance (two hops) to D, but the credits may be exchanged over a set of longer, different paths.

If the view is allowed, then the credit value on each link (u, v) on each path p_i that comprises the flow is reduced by c_{p_i}; the credit on each link (v, u) is correspondingly increased by c_{p_i}, where $c_{p_i} \leq f(d_{st})$. An example of this credit adjustment is presented in Figure 4.8 (b). It is worth noting that Genie rejecting a profile view request does not necessarily mean the operator must block the user making the request; Genie merely flags individual views as suspicious. How the provider responds is a matter of policy—normally, the OSM operator provider would deny or delay the view, effectively slowing down the suspected crawler's activity, but not block the user permanently.

The net effect of this transaction should be that the viewer has $f(d_{st})$ fewer credits available while the viewee has $f(d_{st})$ more credits available for future activity. If j is an intermediate user on one of the paths p_i then the total number of credits available to intermediate user j is unchanged, though the distribution of credit among links adjacent to j is different. As long as user j is well connected and can reach other users through any adjacent link, the change in credit distribution is unlikely to impact the user [44]; if j is not well connected, then attempted views may be flagged due to a lack of *liquidity*.[2]

Social networks exhibit a very high degree of connectivity, ensuring good liquidity in the credit network for most users. New users, inactive users, or small fringe communities that have not (yet) established strong connections to the rest of the network may have issues with liquidity. We will consider this issue in more detail in Chapter 4.5.

Genie leverages three key characteristics of honest user activity identified in Chapter 4.3 to thwart large-scale crawlers.

Leveraging unbalanced view ratios We observed in Chapter 4.3 that honest users have a balanced ratio of views requested to views received, while crawlers issue many more views than they receive. Genie allows a user to use credits obtained by being viewed to perform views in the future. This ensures liquidity amongst honest users with balanced activity ratios while draining credits from crawlers.

Because honest users do not have a perfect balance of views, even honest users run the risk of eventually exhausting all of their credit. To address this concern, Genie rebalances the credits on a pair of links $(u, v), (v, u)$ at a fixed rate r_b. For example, this can be implemented by dividing time into intervals, and at the beginning of each interval

$$c_{uv} \leftarrow c_{uv} - \tfrac{r_b}{2}(c_{uv} - c_{vu})$$
$$c_{vu} \leftarrow c_{vu} + \tfrac{r_b}{2}(c_{uv} - c_{vu})$$

where $0 < r_b \leq 1$.

[2]Liquidity in the context of credit networks is defined as the capacity to route credit payments [44].

Leveraging different path lengths Honest users tend to view profiles of other users that are nearby in the social network; crawling a significant fraction of the social network requires crawlers to view users who are disproportionately far from the crawler in the social network. Genie discriminates against crawlers by charging more credits to view distant users. The cost, in credits, to view a user that is distance d_{st} away in the social network is defined by the simple cost function $f(d_{st}) = d_{st} - 1$.

Leveraging repeated views Honest users repeatedly view the same subset of profiles in the network; crawlers eschew repeat views unless they are re-crawling the network to track changes. Genie does not charge for repeat views that occur within a given time period. If user s views the profile of user t within T days of when s was last charged for viewing t then s is not charged for the profile view; if more than T days have elapsed then s is charged as normal. A typical value of T would be on the order of months.

4.4.3 Security properties

We now describe the security properties provided by Genie.

Let $C \in V$ be the set of user accounts controlled by the crawler and $H = V \setminus C$ be the set of user accounts not controlled by the crawler (i.e., the honest users). The crawler's goal is to view the profiles of all users in H as quickly as possible. (He can trivially obtain the profiles of users in C.)

We call a link (c, h) in the social network an *attack link* if $c \in C$ and $h \in H$. The cut separating C and H (i.e., the set of attack links) is called the *attack cut*.

To determine the rate r_c at which a crawler can view profiles in H, we need only consider the attack cut, because all of the crawler's views have to cross this cut. Profile views within H or within C are irrelevant, because they must cross the attack cut an even number of times, and do not change the credit available along the cut.

The rate r_c is determined by the following factors:

- A, the size of the attack cut (number of attack links): a powerful crawler has a large attack cut.
- d_h, the average OSM distance for honest profile views.
- d_c, the average OSM distance between users in H and the corresponding closest user in C.
- r_c, the expected rate of profile views received by users in C from users in H: per our threat model, the crawler has little control over this rate, and we can conservatively assume that it is the same as the expected rate of views received by a user in H.
- r_b, the rebalancing rate.
- f, the view cost function.

Using $f(d) = d - 1$, the maximal steady-state crawling rate

$$r_c = A\frac{r_b + r_c(d_h - 1)}{d_c - 1}$$

The numerator is the crawler's "income", the rate at which he can acquire credits. The denominator is the crawler's "cost", in credits, per profile view. As we can see, the maximal crawling rate increases linearly with the number of attack links, at a slope defined by the second term. A larger r_b, r_c or d_h increases, a larger d_c decreases the slope.

The credit network effectively makes the power of a corrupted account attack proportional to the number of acquired attack links.

A crawler can increase the crawling rate by obtaining additional attack links, which are difficult to obtain in large quantities. Obtaining an attack link requires forming a social link with a user not already controlled by the crawler, or compromising a user account that has social links with users not already controlled by the crawler. Creating more user accounts by itself is ineffective, because it does not yield new attack links. As a result, the credit network renders Sybil attacks as such ineffective. Additionally, purchasing many compromised accounts is unlikely to provide the attacker with much additionally crawling

ability: many accounts available in underground marketplaces have been observed to not be well connected users, but rather poorly connected users on the fringes of the OSM [121].

4.4.4 Potential for denial-of-service attacks

A key concern with any credit-based network is a credit exhaustion attack, where an attacker seeks to prevent legitimate transactions among innocent users. In the case of Genie, there are two questions to consider. First, can an attacker's attempt to crawl the network prevent good users from viewing each other's profiles? Second, can an attacker target specific users and prevent them from viewing other users' profiles, or from having their own profile viewed by other users?

Due to the properties of the credit network, the severity of the attack (i.e., the rate of failed profile views the attacker can cause) is identical in each case. However, in a targeted attack, the attacker can choose to selectively inflict pain on a subset of users. In the following, we consider both types of attacks simultaneously.

First, we note that the cost function f has a cost of zero credits for a 1-hop profile view. As a result, profile views among friends are never denied, no matter what the state of the credit network. Next, we consider legitimate profile views among non-friends.

Let us consider any cut of the social network other than the attack cut. This $\alpha\beta$ cut partitions V into V_α and V_β, C into C_α and C_β, and N into N_α and N_β, respectively. We call the cut between C_α and N_α the α attack cut, and the cut between C_β and N_β the β attack cut, respectively. Likewise, we call the cut between N_α and N_β the N-$\alpha\beta$ cut, and the cut between C_α and C_β the C-$\alpha\beta$ cut, respectively. See Figure 4.9.

Now, we consider the question to what extent the activity of nodes in C can impair nodes in N in their ability to view the profiles of users on the other side of the $\alpha\beta$ cut.

Observation 0: Without loss of generality, we can ignore views from users in V_α to users in V_α, and from users in V_β to users in V_β. The paths associated with such views must

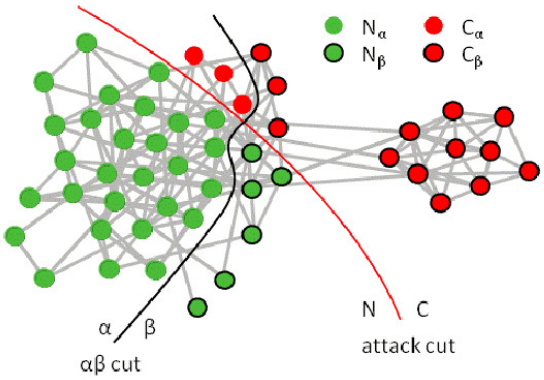


Figure 4.9: Illustration of attack and $\alpha\beta$ cuts.

cross the $\alpha\beta$ cut an even number of times, which means that they do not change the total amount of credit available on this cut.

Observation 1: Without loss of generality, we can focus our attention on the case where the nodes in C_α crawl nodes in N_β, but nodes in C_β do not crawl nodes in N_α. This is because the credit imbalance caused by the crawler's activities along the $\alpha\beta$ cut is maximized in this case.

Observation 2: The nodes in N_β cannot be impaired, because the crawler's activity increases the amount of credit available to them along the $\alpha\beta$ cut. However, nodes in N_α may be impaired, because the crawling of nodes in N_β by nodes in C_α reduces the credit available to N_α along the $\alpha\beta$ cut.

Observation 3: The crawler's activities can reduce the maximal viewing rate available to nodes in N_α by the ratio of the sizes of the α attack cut and the N-$\alpha\beta$ cut. If the α attack cut is larger than the N-$\alpha\beta$ cut, then the crawler can render the nodes in N_α unable to view the profiles of non-friends in V_β.

Observation 4: Social networks are known to have a very densely connected core, with smaller communities connected at the edges [115]. This means that cuts through the core of the network are very large. An attacker would have to be very powerful (i.e., have a very large attack cut) to be able to significantly impair such cuts, and thus a large number of users.

Observation 5: If the attacker is well connected to a small fringe community, it can exhaust credit along the cut that separates the small community from the core of the network. However, by definition this would only affect a relatively small number of users in that community. Moreover, these users can rectify the problem by adding more links to the core of the network.

To summarize, an attacker would have to be very powerful to be able to have a noticeable effect on cuts through the core of the network, which would impact larger numbers of users. A modestly strong attacker can impair users in small fringe communities. However, such users can respond by forming additional links to the core of the network. (Recall our assumption that an attacker cannot compromise the accounts of specific users of his choosing. If an attacker were able to do this, he could target weakly connected users or small communities very effectively.)

We will further explore the impact of crawlers on legitimate users empirically in Chapter 4.5.

4.5 Evaluation

In this chapter, we evaluate the performance of Genie over several different social networks. When evaluating Genie's performance, we focus on the two primary metrics of interest: (1) the time required for a crawler to crawl a Genie-protected network and (2) the amount of honest users' activity flagged by Genie.

4.5.1 Datasets used

To evaluate the performance of Genie, we need four datasets: (i) a social network graph, (ii) a time-stamped trace of honest users' profile views in the form of (X, Y, t) where user X views user Y's profile at time t, (iii) a crawler's topology (i.e., how the user accounts

controlled by the crawler are embedded in the network), and (iv) the crawler's profile crawling trace.

Social network graphs We evaluate the performance of Genie on social network graphs taken from four different online social networks: RenRen [91] (RR-PKU), Facebook [164], YouTube [115] and Flickr [114]), which were introduced in Chapter 4.3.2. Table 4.1 shows their high-level characteristics.

Gathering and generating workload traces Gathering profile viewing traces for large social networks is rather difficult, as it requires explicit cooperation from social network site operators. Unfortunately, many OSM operators are reluctant to share such traces due to competitive and privacy concerns [123]. Thus, we were able to obtain a profile viewing trace for the RR-PKU [91] network only.

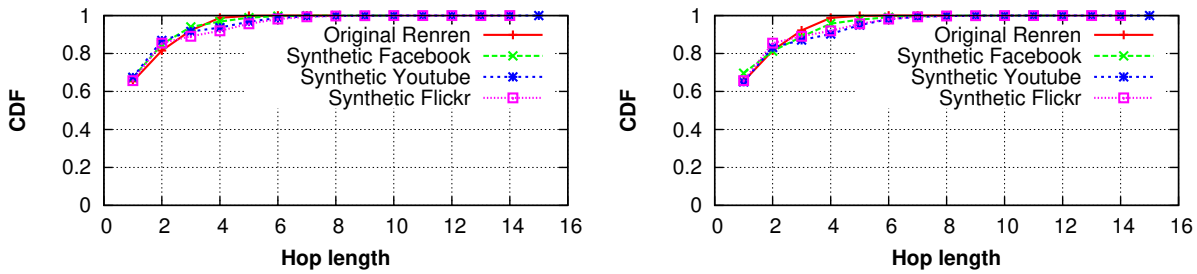
As a result, we design a workload generator that reproduces the key features of the original RR-PKU profile viewing trace that we observed in Chapter 4.3. We focus on two features that capture the correlation between profile view request/receiver user degree, the number of views per user, and the locality of profile views.

It is difficult to preserve the correlation between both requester and receiver user degrees and number of interactions while ensuring the locality of interactions because of varying degree distribution and path length distribution across different networks. Instead, we generated two synthetic traces: a *receiver trace* that preserves the correlation between the receiver user degree and number of received views while ensuring the locality of interaction, and a *requester trace* that preserves the correlation corresponding to requester user degree and number of requests made while ensuring locality of interaction. The high level statistics of one of the receiver trace workloads is shown in Table 4.3.[3]

[3]We note that while the larger graphs have higher number of profile request activities in the trace, the number of activities do not scale linearly with the number of nodes in the trace. This is to be expected because larger graphs have a greater fraction of low degree nodes in their network that generate few profile requests, if any. Our generator faithfully captures the correlation between node degrees and their activity. Other studies of social networks, such as [164], have observed similar trends.

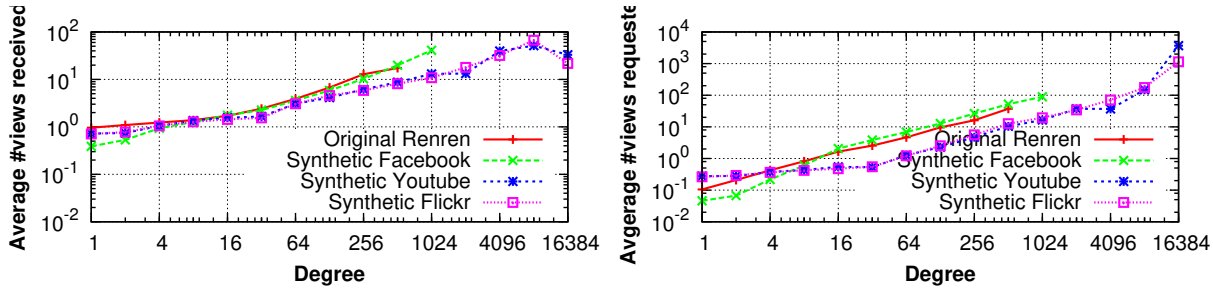
Network	Users	Profile views
RR-PKU [91]	33,294	77,501
Facebook [164]	63,392	98,960
Youtube [115]	1,134,889	984,425
Flickr [114]	1,624,991	1,703,831

Table 4.3: Statistics of synthetic profile view workloads generated for different networks.



(a) **Receiver trace** (b) **Requester trace**

Figure 4.10: Comparison of cumulative distribution of hop distances between different synthetic workloads and the original workload.



(a) **Receiver trace** (b) **Requester trace**

Figure 4.11: Comparison of degree to #views (requested or received) correlation between different synthetic workloads and the original workload.

One concern with our trace generation is that if the social networks are sparse and have very high average path length, the generated trace may not ensure locality of interaction while preserving the correlation between user degree and number of interactions. We cross check whether our traces preserve the intended key features. We do this by testing whether the newly generated traces preserve the two key features we aim to reproduce: (i) locality of interactions and (ii) the correlation between user degree and number of profile views received (or requested). We plot the locality of interactions in Figure 4.10 and the correlation

between node degree and number of profile views in Figure 4.11 for both the original and the synthetic workload traces. The curves match quite well for all three social networks (Facebook, YouTube and Flickr) indicating that the synthetic workload generator retains the key properties of the original RR-PKU workload. We used 5 synthetic traces generated using different random seeds for each of the Facebook, Youtube and Flickr networks.

Crawler's attack topology We model crawlers by simulating the crawler has compromised the accounts of random users in the social network. We simulate 1, 10, 100 and 1,000 corrupted user accounts in the RR-PKU, Facebook, YouTube and Flickr networks. As the crawler obtains access to more corrupted accounts in the network, he also acquires many more attack links to honest users. The varying strength of crawlers on different networks is discussed in Chapter 4.3.2 and Table 4.2.

Crawler's profile crawling trace To generate the crawler crawling workload, we follow the same crawler model discussed in Chapter 4.3.2. This models a crawler that achieves the lowest average path distance to the crawled profiles, and is the optimal attacker strategy.

Unless otherwise noted, all results are the average across 25 different runs of our simulator (5 synthetic honest user profile viewing traces, each paired with 5 synthetic crawler traces).

4.5.2 Trace-driven simulation methodology

To evaluate the performance of Genie, first we built a max-flow path based trace driven simulator. We use the social graph connecting the users to simulate a credit network. For each profile view in the workload trace, our simulator checks if there exists a set of paths in the credit network that allow $p - 1$ units of credits to flow between the viewer and the viewee, where p denotes the shortest path length separating the viewer and the viewee. To this end, our simulator computes the max-flow paths [131] between the viewer and the viewee. If the max-flow is larger than $p - 1$, then the profile view is allowed and if it is not,

then the view is flagged. If the profile view is allowed, the credits along the links of the max-flow paths are updated as described in Chapter 4.4.

A key input to our simulator is the credit refreshment rate, which denotes the rate at which exhausted credits on the links are replenished. We set the credit refreshment rate in our simulator by tuning the following two parameters described in Chapter 4.4.2: (i) the initial credit value, i, assigned to each link in the network at the beginning of the simulation, and (ii) the credit rebalance rate, r_b, which restores some of the exhausted credits on the links after each time step, say of duration t. We set the parameter r_b to 1, which has the effect of restoring the credit values on all links to i after every refresh time period (2 weeks in our experiments). So $\frac{i}{t}$ represents the effective credit refreshment rate, which determines the number of profile views accepted both for crawlers and honest users. As the value of credit refreshment rate increases, more profile views will be accepted from both crawlers and honest users. Thus, the key evaluation challenge that we address using our simulator is: *Does there exist a credit replenishment rate that significantly slows down crawlers, while flagging few views by honest users?*

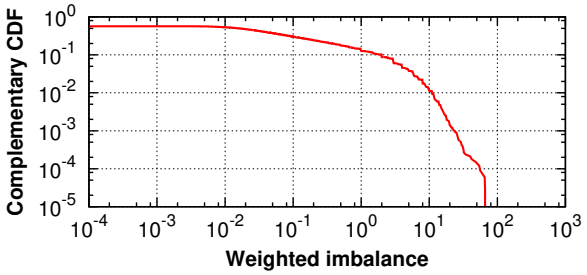


Figure 4.12: Complementary cumulative distribution of the imbalance of RR-PKU users (views are weighted by cost function mentioned in Chapter 4.4). Only 226 (0.7%) users have an imbalance greater than 12.

For a real-life deployment the OSM operator can estimate initial credits by using past browsing activity of users. The operator can build a distribution of user activity based on the imbalance between average number of profile views requested and received (weighted by the cost function mentioned in Chapter 4.4) made per outgoing link per user. Then the

operator can pick an initial credit value/link from this distribution so that most (e.g. 99.9%) users' activity is allowed, but a few profile views are flagged (probably from super active users or crawlers). We check this methodology with our current RR-PKU dataset. Our implicit assumption here is that honest user behavior shows constant trends over time. We show the weighted imbalance distribution in Figure 4.12. From this figure we can estimate the number of users affected at a given credit value. For example with a credit value of 12, our estimate shows that views from 33,068 (99.3%) users will be allowed and 226 users will have some views flagged. We will see in Chapter 4.5.4 that our estimate is quite good and indeed 275 users are flagged with this particular credit setting.

Scaling simulations to large graphs While we were able to run our max-flow path based simulator over the smaller RR-PKU network with 33,000 users, we found it computationally expensive to scale our simulations to the much larger YouTube and Flickr social networks with millions of users, links, and profile views. The computational complexity arises out of three reasons: (i) even a single max-flow computation over a large graph is expensive, the most efficient algorithms for the maximum flow problem run in $O(V^3)$ [79] or $O(V^2 \log(E))$ [49] time; (ii) we have to perform millions of such computations, one for each profile view in the trace, and (iii) worse, the computations cannot be parallelized and have to be done online and in sequence, as the max-flow computation for a profile view has to account for credit changes on links in the network due to all prior profile views in the workload trace.

To allow Genie to be deployed on much larger networks, we leverage the recently proposed *Canal*[166] framework. Canal speeds up computations over credit networks and enables credit operations on very large-scale credit networks (of the order of millions of users) with low latency (of the order of a few milliseconds or less). Canal uses a novel landmark routing based technique to pre-compute paths with available credit (between different users in the network) continuously in the background as new credit operations are processed. Canal trades off accuracy for speed to achieve this goal. It explores only a subset

Network	Avg. time (ms)	95^{th} percentile time (ms)
RR-PKU [91]	0.16	0.78
Facebook [164]	0.21	0.86
Youtube [115]	0.46	1.45
Flickr [114]	0.65	1.41

Table 4.4: Average and 95^{th} percentile time taken by Canal implementation to process one view request.

of all possible paths between two users to complete a credit network operation between them. Thus, Canal may not always find sufficient paths with available credit between two users, even if such credit exists. However, Canal can achieve over 94% accuracy on various large-scale social networks [166]. Using Canal we were able to use Genie on data from online social networks including YouTube and Flickr that contain millions of users and links. We show the latency for processing one profile view with the Canal implementation of Genie in Table 4.4. Our current Canal implementation runs on a single machine. Results in [166] shows that, using a single machine with 48 GB RAM and 24 CPU cores, Canal can support operations on graphs with over 220 million links. Canal could likely be scaled to networks with a billion links using multiple machines and graph parallel processing frameworks [81, 108].

Simulating Genie with Canal We implemented Genie using the Canal library. While processing a profile view, Genie asks Canal for the shortest path length between the viewer and viewee. It should be noted that Canal can only provide an approximate shortest path length because it uses a subset of all possible paths between two users. Genie uses this path length as the basis of charging for the view and then queries Canal again for a set of paths from viewer to viewee with sufficient credit values. If Canal returns a set of paths then the view is allowed and Genie deducts credit along the returned set of paths. The view is flagged otherwise. Canal requires two parameter settings to configure the amount of path data to

be pre-computed. We use the same settings[4] used in the original Canal work [166] (these settings provided over 94% accuracy when applied to the Bazaar [131] system).

As Canal may fail to find sufficient credit when it exists, but will not find credit that does not actually exist, deploying Genie with Canal provides an upper bound for the number of flagged profile views. We now examine how close the Canal estimates are to the true level of flagged user activity.

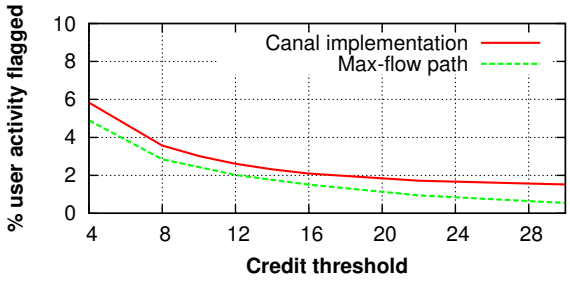


Figure 4.13: Variation of the fraction of user activities flagged with different credit values in RR-PKU in the presence of a crawler with 10 compromised accounts. We compare the results obtained using Canal implementation (red solid line) with those obtained using max-flow based technique (green dotted line). The red solid line provides a close upper bound for the green line.

To understand the effect of approximation error introduced by Canal in terms of flagged honest user activities, we compare the Canal implementation output with the output from the max-flow path based technique. We use the RR-PKU network for this part of the evaluation as the relatively smaller size of RR-PKU enabled us to use max-flow path based technique. Figure 4.13 shows the percentage of flagged honest user activity for the two techniques in the presence of a crawler with 10 compromised accounts. As the amount of credit available per refresh period increases, the percentage of flagged activity decreases for both techniques. On average, the absolute difference in flagged activities between the two techniques is only 0.7% of all user activities, suggesting that Canal provides good accuracy. In the rest of our evaluation we will present results using our Canal implementation.

[4] 20 level-3 landmark universes

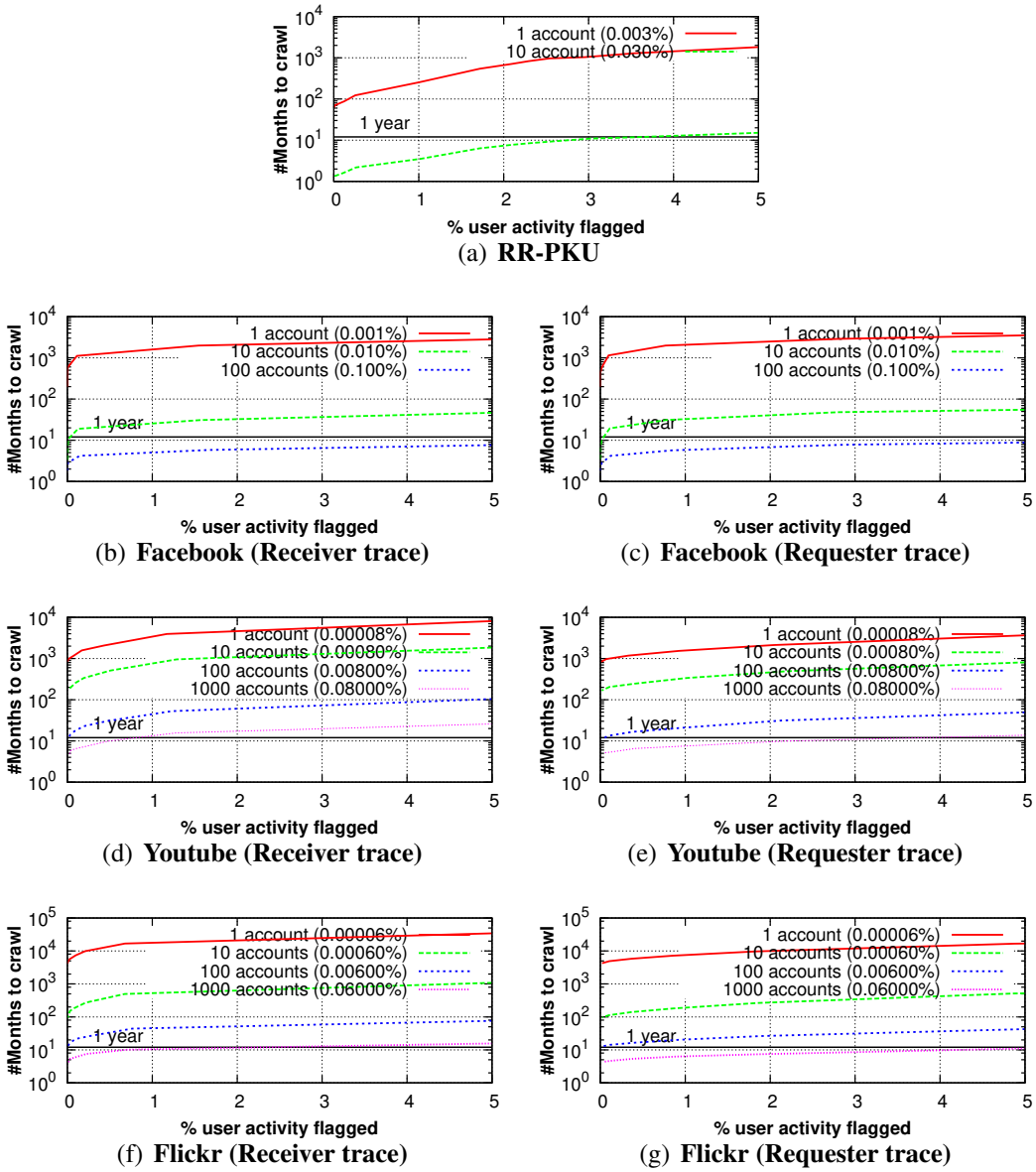
4.5.3 Limiting crawlers vs. flagging activity

We now switch our attention to the core tradeoff that is being made as we select the appropriate credit refreshment rate: namely, the amount of time it takes the crawler to crawl the entire graph and the fraction of honest users' activity that is flagged. We have already observed that to slow down crawlers effectively we need to replenish credits at a slow rate. However, a limited rate of credit replenishment opens up the possibility of honest users' views getting flagged. In this chapter, we explore the extent to which honest users' activity is flagged by Genie as it tries to limit crawlers.

We ran our Canal implementation for various different values of available credit per refresh period. For each replenishment rate, we compute two metrics: (i) the time it would take for a crawler to finish it's crawl and (ii) the percentage of honest users' activity that is flagged. We compare these two metrics looking for a good tradeoff, where crawlers are effectively slowed down, while good user activity is rarely flagged. We present the basic tradeoff for our different social networks in Figure 4.14.

For each social network, we show the results for crawlers of different strengths. On YouTube and Flickr graphs with more than 1 million users, we considered a crawler controlling up to 1,000 user accounts, while for RR-PKU with only 30,000 user accounts, we limited the crawler strength to 10 users. While the absolute number of compromised accounts controlled by the crawler might seem small ($< 0.1\%$), it is worth noting that the percentage of compromised users in these networks is still substantial as discussed earlier in Chapter 4.3.2.

The plots show that it is possible to slow-down crawls sufficiently to force a crawler to spend several months to tens of months to complete a single crawl. At the same time, the percentage of flagged user activity can be held to less than 5%. In many instances, the flagged activity can be held lower than 1%. Thus, there are two important take-away from these results: first, when Genie is employed a certain amount of honest users' activity will



(a) **RR-PKU**

(b) **Facebook (Receiver trace)**

(c) **Facebook (Requester trace)**

(d) **Youtube (Receiver trace)**

(e) **Youtube (Requester trace)**

(f) **Flickr (Receiver trace)**

(g) **Flickr (Requester trace)**

Figure 4.14: Trade-off between fraction of user activity flagged and time taken to finish a complete crawl in with crawlers of varying strengths over different social network graphs. As shown in figure we measure crawler strength by the total number as well as the percentage (shown in parenthesis) of compromised accounts in the network under control of crawler. Further, we conservatively allowed crawlers to exhaust the credits on links before allowing any honest users' activity.

unavoidably be flagged. Second, unless the crawler is quite powerful and possesses over 0.1% of the accounts the impact of the crawler on honest users is minimal.

4.5.4 Alternate strategies for flagged users

We observed in the previous chapter that a certain amount of blockage of honest users' activity is unavoidable. In chapter 4.4.2, we discussed that the OSM operator can make a choice based on some policy, once the profile view is flagged (or blocked) by Genie. Normally, the OSM operator would deny or delay the view to slow down the crawler's activity.

We now pose a simple question: can users do anything to minimize the amount of their flagged activity? To answer this question, we first investigate the flagged views in more detail. We then propose some recourses available to users with flagged activity.

We analyze the set of flagged activities in our extensive RR-PKU simulation, where we compute max-flow paths to verify if a profile view has to be allowed. We intentionally focused on max-flow based simulations because of the certainty that profile views flagged during such simulations are flagged due to lack of credit in the network.

For the analysis in this chapter, we focus on one particular simulation experiment with credit value 12, where 2.6% (or 2,574 activities) of the user activities are flagged and the crawler controlling 10 compromised accounts needs 8 months to complete the crawl.

A profile view can be flagged due to one of the three reasons: (i) the profile viewer runs out of credit on all links connected to itself (i.e., source blocked), (ii) the credit on links connecting the profile viewee is exhausted (i.e., destination blocked), or (iii) the view is flagged due to credit exhaustion somewhere in the middle of the network. Strikingly, only 190 out of 2,574 (7%) flagged views (i.e., 0.18% of all views) are flagged due to lack of credit on links in the middle of the network. The remaining 2,384 (out of which 1,961 views were blocked at the source) views which includes 93% of the flagged activities is due to credit exhaustion on links directly next to the viewers or the viewees. On examining the

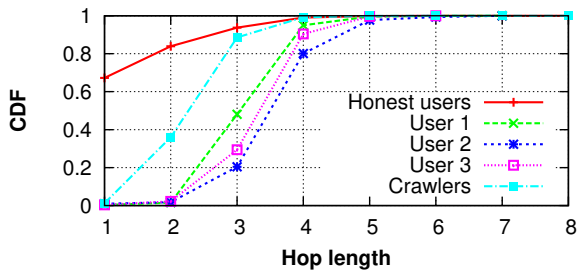


Figure 4.15: Comparison of the hop distance distribution of 3 users with too many views in RR-PKU. Their profile view characteristics deviates considerably from the honest user activities.

degrees of these viewers and viewees who are flagged, we find that 96% of them have degree 1 and 99% of them have degrees 5 or less. That is, most activities flagged near the source or destination, is due to source or destination users having way too few number of friends and lying on the fringes of the network graph. Our results support our observation in Chapter 4.4 that social network graphs are sufficiently well connected in their core that most flagged activity (and credit exhaustion) occurs close to the fringes.

Next, we investigated the amount of flagged activities for individual users. We found that a small number of users are bearing the brunt of the flagged activities. 1,808 of the 2,574 (or 70%) of the flagged profile views are made by 3 users in the network. These are the same super-active users mentioned in chapter 4.3.1. Interestingly, all three users issue two orders of magnitude more views than an average RR-PKU user and they are all flagged near the source. Further investigation suggests that these top three users exhibit crawler-like characteristics. Figure 4.15 shows that the three most blocked users issue considerably more long distance views than normal users. In fact, network locality of their profile requests resembles that of a crawler than that of a honest user with more than 60% viewed profiles lying beyond their 2-hop neighborhood. Ignoring these three users, whose view trace bears strong resemblance to crawlers, the percentage of total flagged activity falls to less than third of its original value, which is already a low percentage (2.6%) of all activity.

For the remaining users who contribute to only 30% of flagged views, we have already observed that most (99%) of the users have degree less than five. These are users who got

flagged because their low number of friend links are insufficient to support the reasonable number (on average 6 views) of views they issued. However, we argue that there is a simple and natural recourse available to them: they can simply form more links in the online social network.

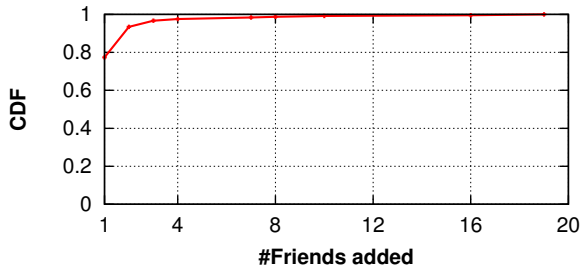


Figure 4.16: Cumulative distribution of how many extra links flagged RR-PKU users needed for completing their activities. Evidently majority of them needed just a few more links.

In order to test this hypothesis, we perform a simple experiment. We re-run the Genie simulation where each flagged user which consists of 275 users (falling in the low degree category), establishes a friend link to the destination of the flagged view (i.e requester sends a link request and the receiver approves it). This immediately leads to the acceptance of that view. At the end of the simulation, we look at the number of friend links established by each flagged user so that all the earlier flagged views could now be accepted.

Figure 4.16 shows the distribution of number of friend links established versus the ranked set of users. A significant majority (269 out of 275 or 97%) of the flagged users can get their views accepted by only establishing very few links (less than 4). Thus, most of the activity flagged by Genie would be accepted if the users of the flagged views spent a minimal effort to establish a small number of friends. In fact, if Genie were to be deployed, it would naturally incentivize users to form a few more friend links. Given that many OSMs already explicitly encourage their users to form more friend links, we believe that the overhead from Genie would be acceptable for a majority of users.

4.6 Discussion

In this chapter, we address the problem of preventing large-scale crawls in online social networks, and present Genie, a system that can be deployed by OSM operators to thwart crawlers and proactively control the exposure of the user generated content. Based on trace data from the RenRen OSM, we show that the browsing patterns of honest users and crawlers are very different. While most honest users view the profiles of a modest number of users who tend to be nearby in the social network, even a strong, strategic crawler must view profiles of users who are further away. Genie exploits this fact by limiting the rate of profile views based on the connectivity and social network distance between a profile viewer and viewee. An experimental evaluation on multiple OSMs shows that Genie frustrates large-scale crawling while rarely impacting browsing of honest users; the few honest users who are affected can recover easily by adding a few additional friend links.

In the next two chapters we investigate our last scenario of exposure control: controlling longitudinal exposure. In other words, we check how users are controlling exposure of their old content today and what we can do to help them.

CHAPTER 5

Understanding longitudinal exposure in OSMs

In this chapter, we focus on understanding longitudinal exposure management of OSM content—a dimension of user privacy that becomes more challenging to manage with the passage of time. Users' privacy preferences for sharing content are known to evolve over time [21, 25]. 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 chapter 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.

Our analysis of Twitter messages[1] reveals striking evidence of a significant fraction ($\sim$35%) of all Twitter users changing their privacy preferences over time. Only a minority ($\sim$8%) of all Twitter users selectively withdraw (i.e., delete or make them private) a small ($\sim$10%) fraction of all their public posts. On the other hand, a sizeable fraction ($\sim$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 [21, 25], 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

[1]This study was conducted respecting the guidelines set by our institute's ethics board and with their explicit knowledge and permission.

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 identify that, as a first step we need to understand the deletion ecosystem in OSMs in general and need to understand the ownership on OSM content in particular.

5.1 Related work

In this chapter 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 [48, 152, 92, 103, 85, 107, 29, 27, 112, 21, 25, 129, 180]. 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 [21, 25] 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 [21] 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 [25] 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 [106] on 802 teenagers found that 59% of respondents edited or deleted their content in OSMs. Almuhimedi *et al.* [11] 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 [11] missed a large part of deleted tweets posted in far past (e.g., 6 years back).

A few other studies [104, 90] explored the changing behavior of Twitter users over time. Out of them, Liu *et al.* [104] 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 OSMs. Bauer *et al.* [25] 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 popularity of systems like Snapchat [145] which deletes all users' posts after a predefined expiry time. Aylan and Toch [21] 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 explore smart policies for content withdrawal.

5.2 Understanding longitudinal exposure

In this chapter, 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?*

5.2.1 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 [11].

2. Withdrawing tweets via deleting account: All tweets posted by a user can be withdrawn by deleting her whole account.

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 5.1: 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.

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 chapter 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 5.1, 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).

As shown in Table 5.1, 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 work, 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*

[2] `https://dev.twitter.com/overview/api/response-codes`

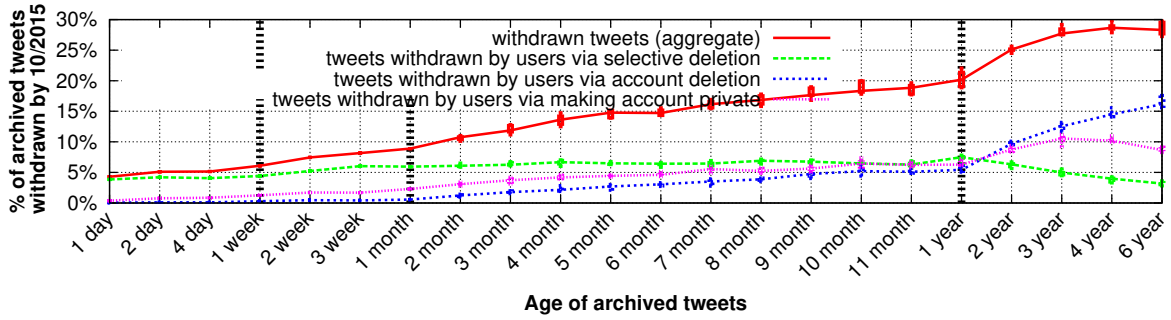


Figure 5.1: Percentage of tweets in our sample of archived tweets that have been withdrawn as of October 2015. The age of a tweet is the difference between the time when the tweet was posted and the time of querying the Twitter API with the tweet-ids (October 2015). The amount of withdrawn tweets is 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).

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 observing 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 chapters. 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.

5.2.2 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 [37], 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 5.1). 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 chapter 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 5.1 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 5.1, 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

[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.

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 5.1 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.* [11] *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.

5.2.3 Understanding user behaviors

In this chapter, 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 [37], 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).

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.

5.2.3.1 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 5.2 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.

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 5.2: 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.

[5]We obtained the country of our users by leveraging location data of Twitter users gathered by Kulshrestha *et al.* [99]. They used the location and timezone field of the Twitter profile for inferring location of users.

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 chapters) 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.

5.2.3.2 Correlating privacy preferences with demographics

Having identified users with different privacy preferences, we now check who these users are, by correlating the longitudinal 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 [122, 144, 99]. 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.* [158] noted a correlation between gender and privacy preferences of users in online social media.

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 5.3: 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.

We infer the gender from the self-reported first names specified in the user profiles using the methodology developed in [122]. Table 5.3 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 [158] 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?

5.3 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 chapter 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?

5.3.1 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?

5.3.1.1 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 Chapter 5.2.3). 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 Chapter 5.3.2 but not for the analysis of withdrawn tweets in this chapter. Thus, the data used in this chapter 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 activities 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 5.2 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.

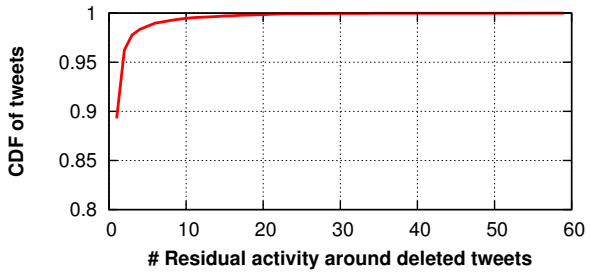


Figure 5.2: Cumulative distribution function (CDF) for number of residual activities per selectively withdrawn tweet. Each of the withdrawn tweets have non-zero residual activity around it.

5.3.1.2 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 5.3 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 chapter, and the boxes in Figure 5.3 indicate the 25th and 75th percentiles. Note that we could recover 16.7% of the keywords when the withdrawn tweets received two or more

[6]We use a list of English stopwords and a list of Twitter-specific stopwords from [179].

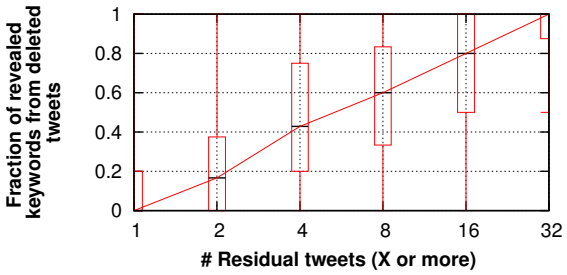


Figure 5.3: Fraction of keywords that could be extracted for each of the withdrawn tweets (with at least one residual tweet) with varying number of residual tweets. The boxes indicate the 25th and 75th percentiles in the fraction, and the whiskers indicate the minimum and maximum values. The recovered keywords from withdrawn tweets increase with the number of residual tweets.

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 5.4: 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.

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 5.4 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 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.

5.3.1.3 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 5.5 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

[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	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 5.5: 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.

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 who wish to control exposure of their old post through selective withdrawal.

5.3.2 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?

5.3.2.1 Residual activities around withdrawn accounts

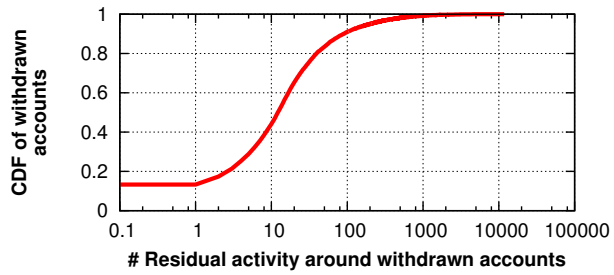


Figure 5.4: CDF of number of residual activities per withdrawn account. More than 55% of withdrawn accounts have more than 10 residual tweets.

We collected residual tweets around withdrawn accounts using a similar methodology as described in Chapter 5.3.1.1. We considered the withdrawn accounts from our random sample of 97,998 users from 2009 (same dataset from Chapter 5.2.3), 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.

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 5.4 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?

5.3.2.2 Recovering social connections

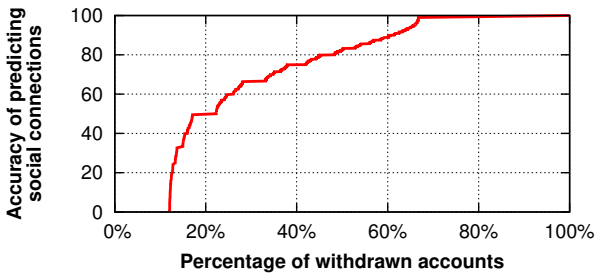


Figure 5.5: The accuracy of our social connection inference with the percentage of withdrawn accounts for which we get this accuracy. For more than 30% of withdrawn accounts, all of their residual tweets came from their social connections.

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.* [37] 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 5.5 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.

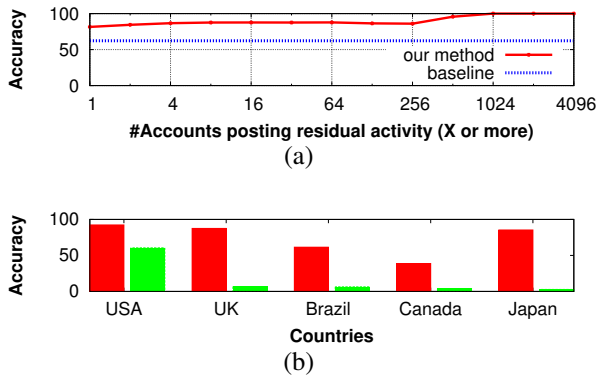


Figure 5.6: 5.6(a) Accuracy of our location inference leveraging residual activities. We can infer location with high accuracy and the inference is consistently better than baseline. 5.6(b) the accuracy for withdrawn accounts from different countries. First bar for each country is accuracy of our method and second bar is percentage chance that a random user will belong to that country.

A large number of existing studies pointed out that connected users in online platforms show homophily, i.e., have similar characteristics [155, 10]. 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.

5.3.2.3 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 [99]. 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).

Demographics prediction accuracy: Figure 5.6(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 Chapter 5.2.3). Figure 5.6(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.

5.3.2.4 Recovering topics of interest

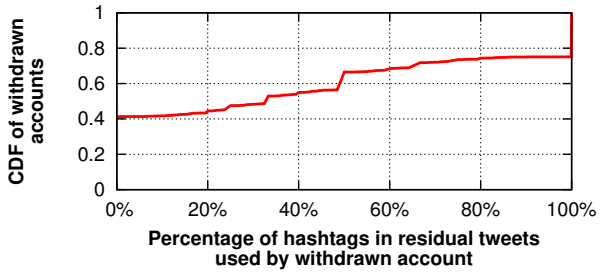


Figure 5.7: The percentage of hashtags revealed by residual tweets that were originally also used by a withdrawn account. 25% of the withdrawn accounts, who ever used any hashtag in their tweets, used all of the hashtags revealed from their residual activities.

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 [37], 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), 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 5.7 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 5.6: 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 5.6 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 Twitter.

5.4 Research challenge: How to better control longitudinal exposure?

Naturally, our findings call for improvements of longitudinal exposure control mechanisms, which will directly increase the usability of such systems from a privacy perspective. We note 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. Thus there is little chance of having a silver bullet to solve the problem to controlling longitudinal exposure. Furthermore, we observe that the complexity arises partially due to the complex issue of the "ownership of content" in the context of OSM content deletion. This issue is complicated since, in OSMs, multiple users interact with the same content (while uploading, sharing, commenting etc.); thus, good longitudinal exposure control mechanisms should be acceptable to majority of these users. To that end,

the concrete research challenge in this space is: we need to better understand and improve the deletion ecosystem in OSMs today. We take a first step to address this question in the next chapter of this thesis. Particularly, we take a deeper look into the longitudinal exposure control mechanisms today; We point out the key challenges in this space and discuss the relative merits and demerits of a few exposure control mechanisms in the next chapter.

5.5 Discussion

In this work, we explored longitudinal exposure control, i.e., controlling exposure of past content. 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 (and might result in privacy violation). We identify the need to better understand the deletion ecosystem in OSMs in order to improve longitudinal exposure control mechanisms. To that end, as a first step, in the next chapter, we survey the longitudinal exposure control mechanisms that are being deployed in different online social sites today and present a structured analysis of the deletion ecosystem, which involves actors other than just the content creator.

CHAPTER 6

Towards better longitudinal exposure control mechanisms

Our analyses in the earlier chapter show that a large number of users withdraw their past social content, but nonetheless often a significant amount of residual information is left behind, possibly leading to significant information leakage about withdrawn social content. To prevent such privacy violations, we must design improved longitudinal exposure-control mechanisms.

As we noted before, improving longitudinal exposure-control mechanisms is a complex problem, which requires consideration of multiple (and sometimes conflicting) factors, such as the desire to retain some old content while allowing other content to be completely removed without a trace [25]. In fact, analyzing the effectiveness of such a mechanism might require a far richer understanding of many dimensions like incorrectly (not) limiting the exposure of (non-)desirable content, the potential privacy impact of such false flags, the ownership of residual activities, and the ease of use and in general understanding the intent of all the actors involved. Hence, it is very unlikely that there is a silver bullet to solve all the problems with longitudinal exposure control.

Any longitudinal exposure control method involves modifying the past—via deleting or editing old uploaded content. However, in the online world and specially in OSMs, the modification of historical content comes with ethical concerns. These concerns are complex and often very contextual. On one hand, if an OSM user removes her historical content

containing hate speech, then she is evading any type of accountability for her hate speech. On the other hand, if an OSM user wants to correct grammatical errors in her earlier post, she is simply aiming to better express her thoughts. The complexity of ethical concerns with longitudinal exposure-control mechanisms arises from the multiple ownership of OSM content—users other than the uploader also interact with OSM content (via replies, shares or likes) and ethical concerns become more complicated since removing or editing past content concern multiple parties. For example, people who commented on an opinionated post might not want the content to be removed, since removing it will affect the context of their comments. In other words, modifying or deleting past content might cause a boundary turbulence among the boundaries set by owners (i.e., content creators) and co-owners (i.e, other users) of the content. Since general boundary turbulence has already been investigated in past privacy theories, we start with a brief review of the relevant theories.

6.1 Related work

The ownership of OSM content may appear to be a trivial or even inconsequential question, since whoever uploads OSM content should intuitively be the owner of the content. However, in OSMs or in general in online platforms the ownership of a piece of content has multiple dimensions—any user (other than the owner) that interacts with a piece of content (replying to the content, re-sharing the content etc.) should also have some control over how the longitudinal exposure of the content is to be handled. In fact, online forums like Reddit enable moderators to delete content contributed by other users. The dilemma with ownership is true even for other online platforms as well: For example, on a platform like ArXiv [1] if a research study is uploaded, cited in subsequent studies, and then at some point in the future removed for some reason, that might have tremendous negative impact on the studies that used findings of that paper. To that end, the communication privacy management (CPM)

theory by Petronio [128] and the theory of privacy in the networked world by Palen and Dourish [127] generally touched upon the idea of ownership for online content.

Petronio's CPM theory: Petronio pointed out in her CPM theory (mentioned in detail in Chapter 2.1.4) that, any user who uploaded the OSM content is the owner of the content. Furthermore, according to her theory, any other entities who have access to the content after uploading are the co-owners of the content. Both the owners and co-owners use rule–based privacy management to control the privacy of the content. CPM theory suggests that when the rules set by owners do not match with those set by co-owners, there are chances for privacy violations. However, in the context of longitudinal privacy management in OSMs, CPM theory neither points out the specific actors (exact owners and co-owners) nor the expectation of these actors from longitudinal exposure-control mechanisms.

Palen and Dourish's theory: Similar to CPM theory, Palen and Dourish (mentioned in detail in Chapter 2.1.5) argued that one of the key privacy boundaries in the online world is the "identity boundary", i.e. "self vs. other". In short, the privacy expectations of the original owner of content might not match with the privacy expectations of other users who can access the content. This mismatch creates tension at the identity boundary. Palen and Dourish further pointed out that the identity boundary is also affected by the content sharing platform which mediates between the content creator and the content recipients. However, they did not clearly mention who the "other" is—for example, if a user uploads a public Facebook profile picture, should the whole population of the internet be counted as "other"?

Thus, we need to investigate the general notion of "others" or "co-owners" in the specific context of understanding the deletion ecosystem in OSMs. Specifically, in our investigation we need to (i) identify the roles of *specific actors*, i.e. co-owners and (ii) identify the key longitudinal exposure-control mechanisms in OSMs today. We will investigate these two questions in this chapter. We will further point out the concrete research questions that needs to be answered for a comparative assessment of longitudinal exposure-control mechanisms. Finally we will conclude this chapter by presenting three improved exposure

control mechanisms in the context of Twitter and evaluating the merits and drawbacks of each of these mechanisms.

6.2 Identifying key actors in the deletion ecosystem

Currently, OSM sites like Twitter enable users to control the exposure of their own content. However, due to the social nature of OSMs, content from multiple people are often effectively intertwined. The social nature of OSM content and the resultant effect on longitudinal exposure control is embodied by "residual activities", which we investigated in Chapter 5.3. One easy way to deal with such residual activities is to consider OSM content and residual activities around them (i.e. replies, likes etc.) as a group and apply longitudinal exposure control to the group of content as a whole. However, the issue is that each group (a piece of content and all of its residual activities) are contributed by multiple entities and can be edited/deleted by multiple parties (e.g., a moderator in Reddit can delete content), including the original uploader and other OSM users/moderators who interacted with the content. Note that, in this thesis we try to provide users better privacy management tools for their own uploaded content. Specifically, in the scope of this thesis, we simply assume that if an OSM user uploads some content, she owns that content. Below we systematically explore different actors in the deletion ecosystem and their roles in this content.

Owner: The uploader of the OSM content is the "owner" of a piece of content. The uploader owns the OSM content and consequently has the direct right to control longitudinal exposure. Today, in OSMs this right is acknowledged by enabling the uploader to control the longitudinal exposure of their content (e.g., via deletion).

Direct exposure set: The users who interact with the content are the "co-owners" of the content. We note that, in a very liberal sense, all the people who are in the access control list can be considered as co-owning the content. However, as we explained in our exposure control model, in the context of privacy, the relevant set of users are those who actually

viewed the content, not all of those who are simply in the access control list. Thus, the users who actually viewed the content (i.e., the users in the exposure set of the content) or, more specifically, interacted with the content (via commenting, sharing etc.) should be the "co-owners". To that end, we take the human users[1] in the exposure set of a content as co-owners. Collectively we call these users as the "exposure set". Specifically we denote the users whom the owner explicitly points out as co-owners (e.g., via tagging them in the post) are part of *direct exposure set*. Note that this notation does not apply transitively. For example, even if a user in the direct exposure set tags another user in the same content, then the tagged user is still not in the direct exposure set as the owner herself did not tag him.

Indirect exposure set: We denote the users who were not assigned co-ownership by the owner, but rather gained co-ownership by interacting with the content (e.g., via commenting on, replying to or liking the post) as the members of the *indirect exposure set*. Note that, in the aforementioned case, where user A in the direct exposure set tagged user B, user B will be in the indirect exposure set. Intuitively, the members of the direct exposure set should have more knowledge about any longitudinal exposure mechanism applied to a piece of content compared to the members of the indirect exposure set. In this thesis, by the generic term "exposure set" we mean the union of users in the direct and the indirect exposure set.

Redactor: We denote the user who controls the longitudinal exposure of a content (via deletion or edit) as the "Redactor". The redactor might be the owner of the content. However, in certain cases, other users (e.g., a moderator or system administrator) can delete or edit a particular piece of previously posted content.

Suggester: Someone who flags content due to violation of some standards, potentially causing a redactor to delete it. A suggester might also act passively, e.g., by writing "You shouldn't have said that" in response to a post.

[1]OSM can discount bots or crawlers who viewed the content, as ideally only human users should not have ownership.

Key actors in the deletion ecosystem	Description
Owner	The user who uploads the content.
Direct Exposure set	Users whom the owner explicitly ask to view and co-own the content, e.g., by tagging.
Indirect Exposure set	Users who proactively co-own the content, by liking, commenting or sharing.
Redactor	Users who can edit or delete a content, for example the owner or a moderator of the OSM or even the OSM operator.
Suggester	Users who suggest editing or deleting a content, e.g., by flagging the content or by posting a negative comment about the content.
Impact set	Users who can detect an edited or deleted content.
Mediator	The OSM operator.

Table 6.1: The key actors in the deletion ecosystem of OSMs.

Impact set: The impact set is the set of users who have the ability to view any longitudinal mechanism applied to a content. Intuitively, if a site like Facebook maintains a public "edit history" (i.e., a historical edit log made by an owner) alongside any publicly uploaded content, then the whole population of Facebook is in the impact set of that content. However, intuitively, the impact set normally should only include the members in the exposure set of original content.

Mediator: In the online world, the mediator (i.e., in our case the OSM operator) hosts the content and delivers the content from the owner to the recipients. Note that mediator is not in the exposure set, since ideally the content delivery in OSM platform is automatic. However, the mediators are also in charge of providing users longitudinal exposure-control mechanisms as well as preserving accountability in OSMs (e.g., by detecting and stopping hate mongers, cyber bullies etc.). Thus the longitudinal exposure control used in the OSMs should also concern the mediators as they partially own the content by virtue of hosting it.

We will address these above mentioned entities as *actors* in this chapter. We point out the actors and their brief descriptions in Table 6.1. Observe that a single user can take multiple roles. For example, a moderator in Reddit can be both redactor and be in the

exposure set. Naturally, in order to avoid ethical concerns, longitudinal exposure-control mechanisms or in general past content modification mechanisms should respect the intent of as many actors as possible. Currently, different OSM platforms provide the owner a number of different content modification mechanisms to control the longitudinal exposure of their old content which has various side effects on the actors. Next, we review the available modification mechanisms in current OSMs.

6.3 Taxonomy of historical content modification mechanisms in OSMs

In this section we present a taxonomy of content modification methods in OSMs. We divide the historical content modification methods from the point of longitudinal exposure into two broad dimensions—1) the removal of content; and 2) the editing of content. We choose the top ten OSMs by the number of active users (as of April 2017) whose primary language is English [119]. These sites are Facebook, Whatsapp, YouTube, Facebook Messenger, Instagram, Tumblr, Twitter, Snapchat, Skype and Viber. Note that, five of these sites, Whatsapp, Facebook Messenger, Snapchat, Skype and Viber are primarily instant messaging services. However, they come under the broad definition of OSM [95]—systems that leverage Web 2.0 technologies to allow the creation and the exchange of user generated content (both from one user to another and in groups).

In addition to these ten most popular OSMs, we also choose three other popular OSMs that exemplify other important retrospective content-management mechanisms: Reddit, 4chan and Whisper. Reddit is a popular forum to post and discuss social content. It had 243 million monthly active users as of April 2016 [149]; interestingly Reddit employs two interesting longitudinal exposure-control mechanisms (i) not letting users delete their account (ii) anonymizing posts. 4chan [28] and Whisper [41] are two popular anonymous OSMs—4chan is an anonymous image board with multiple topic based discussion forums

OSMs	# active users
Facebook [53]	1,968 million
Whatsapp [170]	1,200 million
YouTube [175]	1,000 million
Facebook Messenger [56]	1,000 million
Instagram [89]	600 million
Tumblr [159]	550 million
Twitter [160]	319 million
Snapchat [145]	300 million
Skype [143]	300 million
Viber [163]	260 million
Reddit [137]	243 million
Whisper [171]	30 million
4chan [5]	22 million

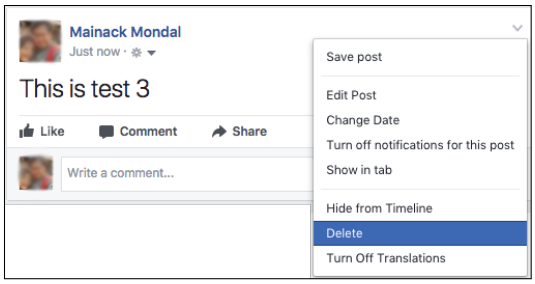
Table 6.2: List of most popular OSMs that we surveyed for identifying key longitudinal exposure-control mechanisms.

(e.g, for politics, music etc.) and Whisper is an anonymous micro blogging service, where users anonymously post short texts. Both 4chan and Whisper are quite popular; 4chan had 22 million monthly visitors [3] (as of December 2016) and Whisper had 30 million monthly active users [18] (as of September 2016). We choose them in order to examine how longitudinal exposure-control mechanisms work in anonymous OSMs. We present the list of the OSMs with their monthly active users in Table 6.2.

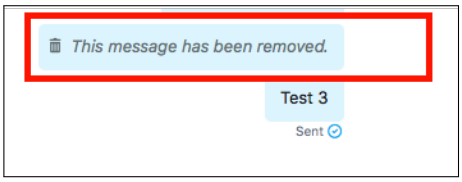
We surveyed these 13 OSMs and identify the key longitudinal exposure-control mechanisms deployed in these systems. Next, we will investigate the content removal and editing mechanisms separately.

6.3.1 Historical content removal mechanisms

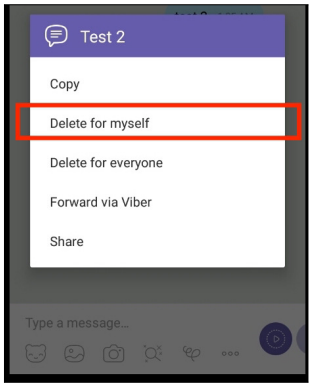
In this thesis, we consider mechanisms that deal with the removal of historical content *directly* (e.g., by choosing a single option in the interface). This conservative definition of removal allows us to differentiate removal mechanisms from editing mechanisms. For example, in a hypothetical scenario, if an OSM allows a user to overwrite her historical content without leaving any edit history, even then overwriting historical content completely



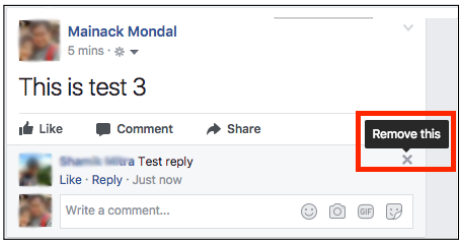
(a) **Deleting individual post (no trace) in Facebook.**



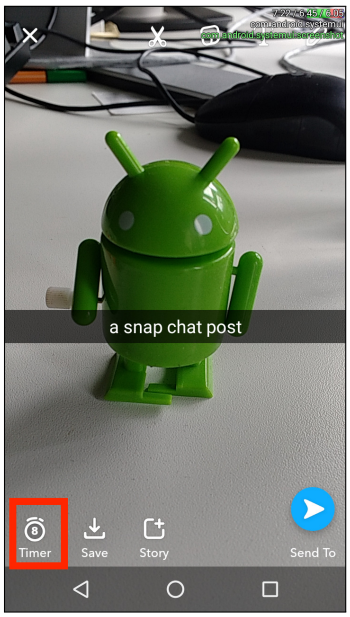
(b) **Deleting individual post (with marker) in Skype.**



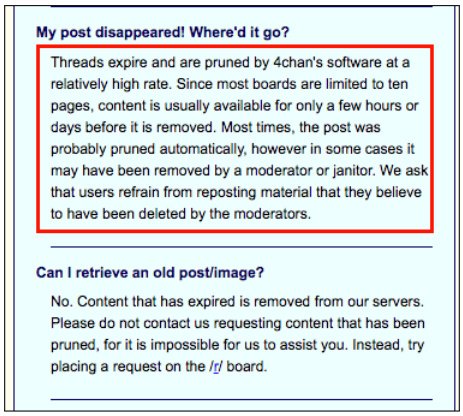
(c) **Deleting only owner's copy of content in Viber.**



(d) **Deleting content from the indirect exposure set in Facebook.**



(e) **Age-based deletion in Snapchat.**



(f) **Inactivity-based deletion in 4chan.**

Figure 6.1: Screenshots of different historical content removal mechanisms in different OSMs. Red boxes in some of the figures highlight the relevant feature in OSM interfaces.

will be considered by us as an editing mechanism, not as a deletion mechanism. To that end, we observe a wide spectrum of deletion mechanisms in these OSMs, ranging from completely deleting individual posts to not letting users delete posts. We summarize the removal strategies in Table 6.3.

- **Deleting individual post:** Many OSMs allow the owners to delete their individual posts from the system without any trace. In our survey we observed that Facebook, YouTube, Instagram, Twitter, Snapchat, some forums of 4chan and Whisper allow this type of deletion. In the case of 4chan if the post is not made by moderators then the post can be deleted. Figure 6.1(a) provides an example of such individual post deletion in Facebook.

- **Deleting individual post from the system (with a marker):** OSMs like Skype and Viber allow the owners to delete their individual post from the system (i.e., delete their posts from the senders as well as the receivers). However, they keep a "marker", i.e., a symbol or a piece of text to indicate that the post is removed. For example, Figure 6.1(b) shows that in Skype, the system replaces deleted content with a "this message has been removed" marker.

- **Deleting only the owner's copy:** Unlike Skype and Viber, some other OSMs only allow to delete the copy of the owners. This mechanism is similar to the deletion of emails. If Alice sends an email to Bob, then even if Alice deletes the email from her inbox, Bob still retains his copy of the received email. For example Figure 6.1(c) points out the delete options provided by Viber; The highlighted option (marked with a red box) clearly mentions that this option will only delete the content for the owner. Interestingly Viber also provides option for deleting an individual post from the system with a marker. However, OSMs like Whatsapp, Facebook Messanger or Tumblr only provide the option of deleting the owner's copy, and not the recipients' copies.

- **Deleting whole account:** All OSMs we surveyed other than Reddit provide users an option to delete their entire account and remove any content posted by that account. We note that out of the OSMs we checked only Reddit does not offer deletion of post or account out of the box. In Reddit, when a user attempts to remove her content, Reddit simply removes the user's identity from the content, but the content itself is not deleted.

- **Deleting activities around content:** Some OSMs like Facebook, YouTube, Instagram, Tumblr, Snapchat, and 4chan allow the owners to delete activities around their content (e.g., comments by others), i.e., content from the indirect exposure set. For example Figure 6.1(d) shows how Facebook allow an owner to delete the comments on their posts.

- **Age-based deletion:** Ephemeral OSMs like Snapchat allow the users to set a timer on their posts and remove the posts after the time is over. In Figure 6.1(e) we show an example of a Snapchat post; the timer (bottom left corner) allows the owners to auto-delete their posts after the preset time.

- **Inactivity-based deletion:** 4chan employs a type of inactivity-based deletion, where a discussion thread which does not have any new replies will be eventually deleted. Figure 6.1(f) shows the 4chan faq which explains this feature. Essentially, at any given time, 4chan shows only the most active threads, discarding the rest. Thus, a thread that is inactive for an extended period of time (hours or days as indicated by 4chan administrators in Figure) gets deleted.

- **Complete removal of post by redactors other than owners:** Many OSMs like Facebook, YouTube, Instagram, Tumblr, Twitter, Snapchat, Reddit, Whisper, 4chan allow moderators or system administrators to delete content which violates some community standards (e.g., hate speeches or spam). An example of this feature is suspension of user accounts and all of their posts by Twitter.

- **No deletion:** We observe that in some cases OSMs do not allow the users to at all remove any of their historical content. For example, in 4chan, sticky threads (the threads which are shown at the most prominent places irrespective of activity) which are made by OSM administrators or moderators cannot be deleted.

Having reviewed content-removal mechanisms, we next investigate content-editing mechanisms.

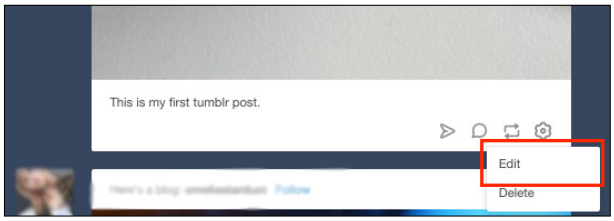
6.3.2 Historical content editing mechanisms

In addition to allowing users to delete content, OSMs also allow their users to edit their historical content, i.e. owners can go back to their content and overwrite their post or modify data surrounding a post (e.g., picture captions or tags). However, the specific editing options are specific to different OSMs. We summarize the editing options in Table 6.3.

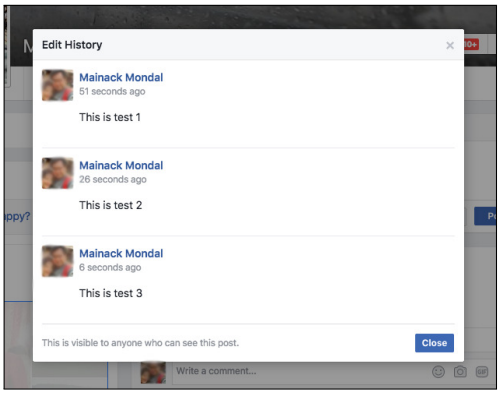
- **Editing post (no trace):** OSMs like Tumblr and Snapchat allow the owners to completely overwrite their old content without an editing log; i.e., in OSMs like Snapchat, others cannot potentially tell if the content is edited after upload. Figure 6.2(a) shows an example of such an edit feature for a Tumblr post.

- **Editing post (with edit history):** OSMs like Facebook store a full edit history of edited past content, which can be viewed by anybody who has access to the original content. Figure 6.2(b) shows the example of the Facebook interface for showing the edit history of a particular post.

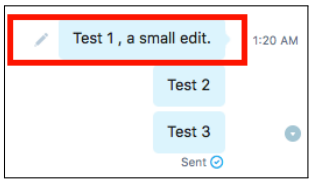
- **Editing post (with marker)**: OSMs like Skype or Reddit (specifically for comments to Reddit posts) do not show the edit history of an edited content. However, they display a marker (i.e, a symbol) to point out that a particular content is edited. For example Figure 6.2(c) shows that an edited Skype message has a small pencil icon next to it as marker.



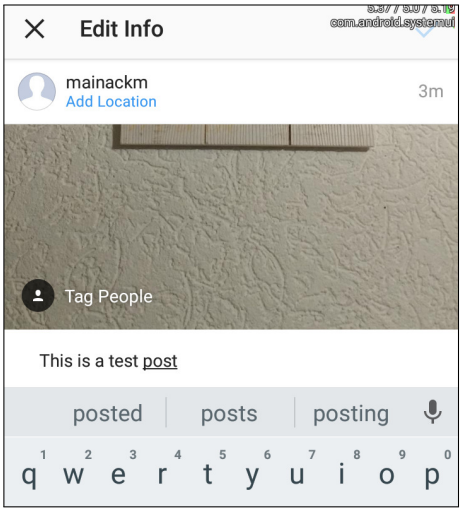
(a) **Edit post (no trace) in Tumblr.**



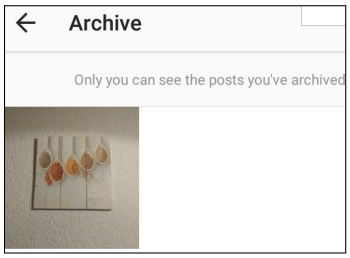
(b) **Edited post (with edit history) in Facebook.**



(c) **Edited post (with the small pencil icon as marker) in Skype.**



(d) **Editing only captions/tags in Instagram.**



(e) **Archive in Instagram.**

(f) **Anonymize in Reddit.**

Figure 6.2: Screenshots of different historical content edit mechanisms in different OSMs. Red boxes in some of the figures highlight the relevant feature in OSM interfaces.

content (photos for Instagram, videos for YouTube). Figure 6.2(d) shows this feature in the case of Instagram.

- **Archive:** OSM platforms like Instagram also allow the owner to "archive" their post. Once a post is archived, only the owner can view the post. They can revive the post at a later point of time if they wish. Figure 6.2(e) shows Instagram's archive feature by showing an archived post.

- **Anonymize:** Reddit also allows another editing mechanism—anonymizing the post. Reddit allows uploaders to remove their identity from the post, or in other words edit the post to detach the content owner's identity from their posts. Figure 6.2(f) shows an anonymized post, where the user identity is replaced by the string "deleted". We note that there are other possible anonymization mechanisms that OSMs can employ, e.g., deleting personally identifiable information (PII) from the historical post or just including a short summary of the user's bio (but not the precise user identity) automatically in order to preserve the context of the anonymized post.

- **No editing permitted:** Finally, we note that multiple OSMs, including Whatsapp, Facebook Messenger, Twitter, Viber, 4chan, Whisper do not allow any editing of past posts. An owner can remove their past post, but they cannot edit it. On Reddit, too, original posts cannot be edited once posted. However, comments on these posts can be edited as mentioned before.

6.4 Research challenge: Exploring usability of longitudinal exposure-control mechanisms

So far our investigation provides the foundation we need to explore the usability of different longitudinal exposure-control mechanisms. Our idea is simple: A good longitudinal

Historical content *removal* mechanisms		
Mechanism	**Description**	**Example OSMs that support this mechanism**
Deleting individual post	Completely deleting individual post from the system without a trace	Facebook, YouTube, Instagram, Twitter, Snapchat, some forums of 4chan, Whisper.
Deleting individual post from the system (with a marker)	Remove individual post from the system but keep a marker (e.g, "the message is deleted")	Skype, Viber
Deleting only the owner's copy	Remove content only from the owner's (e.g, in Whatspp a sender cannot delete receiver's copy of message)	Whatsapp, Facebook Messanger, Tumblr, Viber
Deleting whole account	Complete removal of individual user account information as well as all the associated post	Facebook, Whatsapp, YouTube, Instagram, Tumblr, Twitter, Snapchat, Skype, Viber, Whisper
Deleting activities around content	Removal of activities (e.g., comments by others) by owner	Facebook, YouTube, Instagram, Tumblr, Snapchat, 4chan
Age-based deletion	Automatic deletion of post (e.g., after preset time).	Snapchat
Inactivity-based deletion	Remove posts after a period of inactivity (e.g., no replies in 4chan)	4chan
Complete removal of post by redactors (other than owner)	A moderator or the OSM operator removing a particular content (e.g, spam)	Facebook, YouTube, Instagram, Tumblr, Twitter, Snapchat, Reddit, Whisper, 4chan
No deletion	The owner is not allowed to remove a post	Some forums of 4chan
Historical content *editing* mechanisms		
Mechanism	**Description**	**Example OSMs that support this mechanism**
Editing post (no trace)	Completely overwrite the old post	Tumblr, Snapchat
Editing post (with edit history)	Overwrite the old post, but all previous versions of the post are available on demand.	Facebook
Editing post (with marker)	Overwrite the old post but a marker (e.g., a pencil icon) is added to the post	Skype, comments on Reddit posts
Editing only captions/tags	Cannot edit the main content, but can only edit some surrounding data, e.g, caption of a picture or video	YouTube, Instagram
Archive	Deny access to the post for everybody except owner	Instagram
Anonymize	Anonymize a post (e.g., by removing user identity)	Reddit
No editing permitted	Neither owner nor any other redactors are allowed to edit a historical post	Whatsapp, Facebook Messenger, Twitter, Viber, 4chan, Whisper, Reddit

Table 6.3: Summary of the historical content modification techniques in OSMs.

exposure-control mechanisms should respect the *intent* of as many actors as possible in a given context. The intent of an actor is how she thinks a removal/editing operation should be carried out. We leave the investigation of user intent in different contexts to future work. However, in this section we provide concrete research questions for our future investigation.

Identifying intent of actors in different context: Different OSMs are created to share different types of content. For example, in the personal-content-sharing OSMs (like Facebook and Instagram) the owner's intent might be considered most important. However, in other platforms the OSM designer might also decide to put more emphasis on the intent of mediators and might forbid content removal in order to enforce accountability of owners. A bulletin-board social-news-sharing system like Reddit might want to emphasize community discussion on social news (i.e., interaction from users in the exposure set). In practice, Reddit indeed puts more emphasis on the intent of the users in the exposure set, consequently choosing not to allow the owner to remove content.

Investigating if different longitudinal exposure-control mechanisms respect the intent of different actors: Another important research challenge is to compare existing longitudinal exposure-control mechanisms in light of the actors' intent. We need to identify in which context a longitudinal exposure-control mechanisms respects the intent of the majority of actors and where these mechanisms fail to satisfy the intent of most of the actors. Furthermore, for a given longitudinal access-control mechanism, we need to determine *why* the mechanism fails to satisfy certain actor's intent.

Designing novel mechanisms like "notifications" to augment current mechanisms: Mechanisms like "archiving" might respect the intent of both the owners and the OSM operator. Nonetheless, archiving still affects all the interactions (from others) around the post. To that end, proactively sending deletion/edit notifications to the users in the exposure set might improve such mechanisms to respect the intent of some of the actors (e.g., users in the exposure set). In fact, notifications about removal/deletion can be proactively pushed to

the actors in some scenarios and in other scenarios it might be more rational to let actors pull the notifications (so that an actor will know about the removal/editing only the actor is curious). For example, in the case of archiving, the users who interacted with the post might receive a push notification that the post is archived and from then onwards, the actors can only access a read-only, archived version of the post.

Systematically exploring these research questions is part of our concrete future work. However, next we propose three exposure control mechanisms to show that we can indeed respect multiple actor's intents in specific contexts with different proposals. We take Twitter as our experimental platform and explore some improved longitudinal exposure control methods for Twitter from the deletion ecosystem in general.

6.5 Proposal for better longitudinal exposure-control mechanisms in Twitter

Twitter primarily provides a simple deletion mechanism to the owners for controlling longitudinal exposure-control mechanisms for old content and does not allow users to edit their past content. We start with exploring the existing longitudinal exposure-control mechanisms in Twitter.

Putting users in charge of controlling their longitudinal exposure: This mechanism is used in most of the popular online OSMs, including Twitter and Facebook, where the content owners 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 Chapter 5 demonstrated, even when owners withdraw their posts or accounts, the residual activity surrounding the withdrawn posts (authored by other users) could leak significant information about the withdrawn content. Thus, simply deleting old posts still might not satisfy the intent of an owner due to the residual activities in Twitter

(mentioned in Chapter 5). To that end, we propose several possible improvements and evaluate those proposals.

6.5.1 Proposal 1: Age-based withdrawal

Ephemeral OSMs such as Snapchat [145] and Cyber Dust [2] offer a potential way out of the residual activity problem which we can apply in OSMs like Twitter. On ephemeral OSMs, every message is associated with an expiry time after which the post is automatically withdrawn and becomes inaccessible to the users. Ayalon *et al.* [21] have suggested that the system operators of non-ephemeral OSMs (like Twitter) 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.

1. First, the default expiry time used in such mechanisms is generally too small (e.g., a 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 by mediators in sites like Twitter which promote social discussions. One might argue that this limitation can be overcome by simply setting the expiry time according to user's preference. Unfortunately, as noted by Bauer et al. in [25], users are generally bad at anticipating when a post should be deleted and thus there is little chance that this simple solution would work.

2. Second, age-based withdrawal or any related longitudinal exposure control methods have a common shortcoming. They solve the problem of residual activities by deleting *all* posts, effectively removing the history of any social activities and destroying the archive of historical social media posts. However, there is a big problem with such removal of old posts—no archive of historical social media posts. In other words

age-based withdrawal completely ignore the intent of actors like users in the exposure set or the mediators.

Some OSMs enable researchers, as well as social media analytics companies (part of the indirect exposure set of content), to study a huge volume of user generated data. A prime example of such an OSM is Twitter. Twitter makes a portion of user-generated data (i.e. tweets) available to researchers and businesses via their streaming or search API [162]. Research studies use this Twitter data to solve problems ranging from understanding human behavior to detecting spam. Many of these studies use historical Twitter data, i.e., data posted recently or in the past (weeks, months or even earlier) in their research (e.g., while investigating user sentiment for last five US elections). Unfortunately, age-based withdrawal (at least the current implementation in platforms like Snapchat) or similar mechanisms makes any such research efforts impossible.

To that end, next we present two more proposals to improve the longitudinal exposure-control mechanisms.

6.5.2 Proposal 2: 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. 4chan offers similar ephemerality by deleting inactive discussion threads over time.

Note that 'interaction' is a general term that can involve several tasks based on the OSM; 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 log all of these interactions.[2] 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

[2] `https://support.twitter.com/articles/20171990#`
`https://www.facebook.com/help/437430672945092`

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 post, she may instead decide to limit access to the post to only select friends or may even anonymize the post by removing any identifiable information. Here we generally consider the 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 the 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.

Limitations of inactivity-based withdrawal: However, inactivity-based withdrawal is a simple improvement over age-based withdrawal. Specifically, even in the case of inactivity-based withdrawal, all the past posts are eventually deleted. Thus, even in this case there is no preservation of part of historical content. Moreover 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.

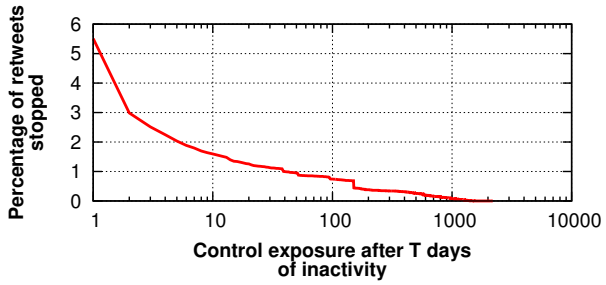


Figure 6.3: Percentage of lost retweets if tweets were withdrawn after T days of inactivity, for different values of T. When T is set to 180 days only 0.4% of the future retweets will be lost.

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 6.3 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 to make an informed decision.

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 6.4: Comparison of age and inactivity-based threshold when both have the same threshold. Retweets of more active tweets are stopped by age-based threshold.

A comparison between the inactivity-based withdrawal and the age-based withdrawal: To demonstrate advantages of inactivity-based withdrawal over the age-based withdrawal, we also simulated the 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 is withdrawn and all future retweeting is stopped. We closely investigated how many retweets would be affected by each of these policies if we set the same threshold. Table 6.4 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 a 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 6.4 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, respecting the intent of mediators and users in exposure set. 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 measured actual time when a tweet will be withdrawn when we set an inactivity-based threshold of T days for different values of T.

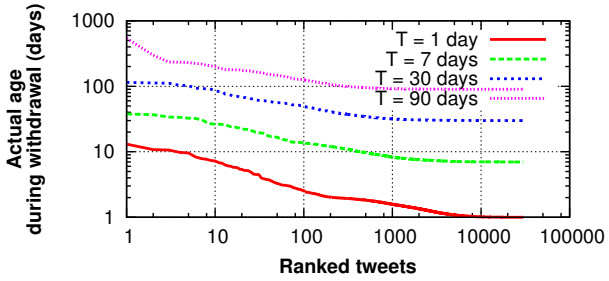


Figure 6.4: Actual time when a tweet will be deleted when we set an inactivity-based threshold of T days.

In Figure 6.4, 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).

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. However, the intent of the users in the exposure set might still not be met by inactivity-based withdrawal, since this mechanism too deletes all post over time.

Next, we propose anonymization as another effective method of longitudinal exposure control which preserves part of historical content and better honor the intent of users in the exposure set (e.g., researchers or users who replied to the content) as well as mediators.

6.5.3 Proposal 3: Anonymize withdrawn tweets

When an owner withdraws her posts, she explicitly expresses desire for not to be associated with that content any more. Thus a simple trade-off between respecting the owner's desire of withdrawal and still keeping a part of the data would be an anonymizing scheme that anonymize and unlink a post from the identity of its owner or uploader (the user who originally uploaded the withdrawn content).

In fact, as we briefly mentioned before, there are multiple systems today that decouple owner identities from content. Examples include fully anonymous social media systems like Whisper (`http://whisper.sh/`) or YikYak (`https://www.yikyak.com/home`), which omits the concept of associating user identities with posts altogether. Moreover, another OSM operator, Reddit, already employs a version of anonymization for their withdrawn content (`https://www.reddit.com/wiki/privacypolicy`). Reddit simply removes the user identity in a withdrawn reddit comment and replace it by a "deleted" string.

Based on these observations, we propose the following idea: OSMs should anonymize the withdrawn historical content to by unlinking the identities of owners from the content for withdrawn content, rather than removing both the identity and the content. This strategy provides a trade-off between the owner's intent of withdrawal and keeping the archive of historical content—anonymization removes user identity from posts, detaching an owner from her withdrawn content and the existence of anonymized version of posts (as opposed to complete removal) preserved a part of the historical data archive. Again we will use Twitter as a platform to instantiate our proposal of anonymizing withdrawal social content (tweets in this case).

6.5.3.1 Anonymization scheme

We propose a simple anonymization scheme for tweets withdrawn by owners; Twitter can recommend to just replace the owner identities (e.g., owner id, owner's user-name) in the withdrawn tweet with a random string.[3] In this way the studies that leverage the tweet content can still analyze the tweet text including URLs and hashtags. Note that here we consider tweets withdrawn by their owners. Twitter can choose to keep the tweets that redactors withdraw (e.g., by suspending accounts) as-is (perhaps with a flag that the user account who posted this tweet is suspended). After the withdrawn historical tweets are anonymized, users in the exposure set (e.g., researchers) can still view historical tweet data, and can obtain the URLs, hashtags and words to continue their analysis. Thus, the impact on the data quality for content analysis will decrease.

Limitations of anonymization: We acknowledge that our proposal is not a silver bullet. Specifically our particular scheme does not address two issues: First, since all owner identities are detached from their tweets, researchers focusing on specific parts from the population might not be able to collect data from those parts (e.g., collecting data from all female Twitter users posting about Brexit). However, we believe that this is a trade-off that a researcher have to make while respecting owner intent of data withdrawal. Second, personally identifiable information (PII) might still remain in the tweets (e.g. as proper nouns or even writing styles) and our simple anonymization scheme will not remove them. Twitter might provide users the option to remove this PII by automatically identifying them. However, finding the PII is highly context dependent and we leave it to future work for improving this aspect.

Aside from these two issues with our particular scheme, in general any anonymization scheme raises a general issue—the loss of context. In general when owners are allowed to delete posts that users in the direct/indirect exposure set have subsequently referred to the

[3] We address the question of whether the random strings are unique per post or just unique per user (and the same across all posts from that user) in Section 6.5.3.2.

deletion does not respect the intent of users in the exposure set. Even for anonymization schemes (when the content is partially retained) removing the author's identify from a post, can change the appearance of a response to the post and disrespect the intent of the users in exposure set. For example, take the case where the content owner is a staunch supporter of ruling party and still vehemently opposes government policies in an OSM post. In this case, anonymization omits the significance of the post considerably by omitting the context of owner's political leaning.

6.5.3.2 Likelihood of de-anonymizing an owner using network structure

There is another technical concern that our scheme needs to address. Twitter is a social network and people converse about published content (e.g., in the form of replies to tweets). However, while anonymizing a withdrawn tweet, Twitter cannot simply anonymize these conversations, too; that will raise a complicated ethical concern — these conversations are posted by users in the exposure set and not the owner of the withdrawn tweet, so, ideally, anonymizing the conversations would require the explicit consent of all users participating in the conversation (i.e., all users in the exposure set). Still, these conversing users are highly likely to be connected to the original owner in Twitter (by follower/following relations) and might reveal owner identity thorough the network structure of the social graph. An obvious question is *how likely* is it for an analyst to identify the owner of an anonymized tweet by simply looking at the social connections of the users conversing (e.g., replying) with the anonymized tweet. Next, we will thoroughly investigate this question using data from the real world.

We will first describe our dataset of Twitter conversations around withdrawn content. Then we will consider two possible implementations of our anonymization scheme: (i) *anonymization per withdrawn tweet* – each withdrawn tweet is anonymized independently, i.e., owner identities in each withdrawn tweet is replaced by a unique random string and

(ii) *anonymization per owner* – where *all* withdrawn tweets from the same owner, owner identity is replaced by same random string.

Note that although anonymization per user better honors the intent of users in the exposure set and mediators (since all withdrawn tweets from same owner and conversations around them can be grouped), it also leaks more information about owner's identity (through information about multiple people conversing with tweets from same owner).[4] We will investigate the likelihood of deanonymizing owner using network structure in both of these cases.

Dataset to evaluate anonymization scheme: For our analysis in this part we require a dataset of withdrawn tweets from a large sample of Twitter users and the conversations around these tweets. We leverage the same dataset mentioned in Chapter 5.2.3 and Chapter 5.3.1.1. Recall that our dataset contains 8,950,942 historical tweets, posted by 97,998 users. Furthermore 33,925 (34.6%) users withdrawn 2,605,317 (29.1%) tweets from this collection. Aside from historical tweets, there are 41,618 conversations (tweets posted as replies) around 36,796 of these withdrawn tweets from 7,964 owners (23.5% of the owners who withdrew their tweets). Our data also contains the social connections of these conversing users as well as the tweet owners.

Using this dataset we seek to answer the question: How likely is it that the original owner of a withdrawn tweet is revealed by simply looking at the social connections of the conversing users? Or in other words, how likely is it that the original owner of a withdrawn tweet is the *only* common neighbor of the conversing users in the Twitter social graph?

Deanonymizing an owner when anonymization is done per withdrawn tweet: Recall that when each withdrawn tweet is independently anonymized, the user identity in each withdrawn tweet is replaced by a unique random string. So an analyst can only identify that conversations around each withdrawn tweet are addressed to a particular owner. In that scenario, we take each withdrawn tweet and collect the social connections (both followers

[4] Aside from the case when any single tweet by the owner makes the owner identity obvious.

and followings) of the users who conversed with that tweet. Then for each withdrawn tweet we check the number of common social connections for the conversing users.

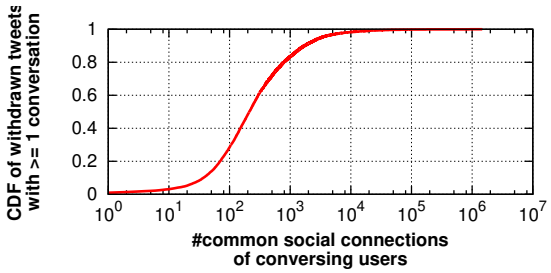


Figure 6.5: The number of common social connections between conversing users for withdrawn tweets with one or more conversations. Only 0.9% of the tweets have one common neighbor within conversing users. Also 98.6% of the 2,605,317 withdrawn tweets did not spur any conversation (not included in the figure).

The result is shown in Figure 6.5. We note that only for 0.9% of the withdrawn tweets with one or more conversation, there is exactly one common connection between the conversing users, and 96.9% of withdrawn tweets have more than 10. We concentrate on the withdrawn tweets, where there is only one common connection between the conversing users. We found that, only for 319 withdrawn tweets, the identities of 168 original owners are revealed, i.e., those owners are the only common social connection of the conversing users around those tweets. In other words, 99.5% of the 33,925 tweet owners who withdraw their tweet cannot be de-anonymized using the social connections of conversing users if Twitter leverage a simple *per withdrawn tweet* anonymization scheme.

We investigate further and find the main reason for this low likelihood of deanonymization: 98.6% of the 2,605,317 withdrawn tweets did not spur any conversation around them and 1.3% received only 1 conversation (i.e., only one reply). Thus, for most of the owners, an analyst does not have enough information from the social connections of the conversing users for revealing owner identity.

Now we investigate the likelihood of de-anonymization if we replace user identities of all withdrawn tweets belonging to a particular owner with a random string, i.e., the random string is not unique for each withdrawn tweet, but for each owner.

Deanonymizing an owner when anonymization is done per owner: In this case an analyst can identify and group multiple withdrawn anonymized tweets belonging to a particular anonymized owner. Moreover, she can also identify that *all* the conversations around those withdrawn tweets are addressed towards a particular user. Thus, intuitively, an analyst has more information available for deanonymizing an owner. We implemented this scheme and check in how many cases the owner is the common social connection of all the conversing users around all of the owner's withdrawn tweets.

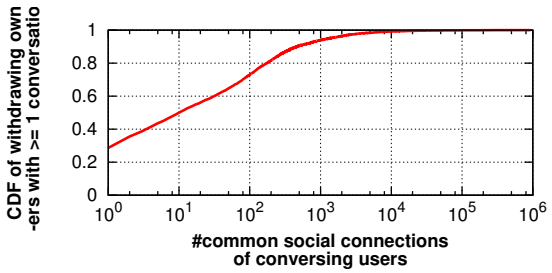


Figure 6.6: The number of common social connections between conversing users for withdrawing owners with one or more conversations. 23.5% of these owners have exactly one common connection between conversing users. Note that, 76.5% of the 33,925 owners with withdrawn tweets did not have any conversation (not included in the figure).

We found that 23.5% out of 33,925 owners who withdrew tweets, have one or more conversation in total around her withdrawn tweets. Figure 6.6 shows the number of common connections between the conversing users for the owners with one or more conversation around their withdrawn tweets. We note that for 28.7% of these owners there is one common connection within the conversing users, and for 50% owners the number of common connections is more than 10.

We again focus on these owners for whom there is exactly one common connection between the conversing users around her withdrawn anonymized tweets. We found that 1,784 owners, i.e, 5.3% of the 33,925 owners who withdrew their tweets, can have their identity de-anonymized using social connections of the conversing users. This fraction is certainly higher than the case when each withdrawn tweet is independently anonymized, but

this anonymization scheme still protects identities of 94.7% of the owners who withdrew their tweets.

6.5.3.3 Improving our anonymization scheme: Future directions

So far, we considered a possible anonymization strategy for Twitter as a longitudinal exposure-control mechanism which partially retains the quality of historical data. Our strategy involved simply anonymizing the withdrawn tweets by replacing a user's identity with a random string. We found that our strategy provides partial protection against deanonymization upto certain degree when an attacker leverages the Twitter social graph and the social connection of users conversing with withdrawn tweets (via replying to tweets). In fact, in the case of per-withdrawn-tweet anonymization, 99.5% of owners would remain anonymized against such deanonymization attacks. In the case of per-owner anonymization, 94.7% of owners would remain anonymized against such attacks. The social media operators can choose which anonymization scheme will be suitable for their service, based on to the extent to which they want to honor the intent of users in the exposure set. They might further improve their anonymization schemes; we propose two concrete directions: First, the OSM operators can just keep a fixed number of residual activities associated with each withdrawn post (and remove the link to withdrawn tweet for other conversations), so that de-anonymization is nigh impossible. Second, the OSM operators can further strengthen anonymization by removing personally identifying information (PII) while anonymizing withdrawn posts. We leave the exploration of these further schemes as a potential direction of future work. Finally, an intermediate step before deploying this mechanism in platform like Twitter would be to give them the choice to either "anonymize" and "delete" old posts. This step would be useful in measuring the response of different actors regarding anonymization.

6.6 Discussion

In this chapter, we surveyed popular OSMs to identify the actors in the deletion ecosystem of these OSMs and the longitudinal exposure-control mechanisms available to these actors. We concretely identify the key research challenges to systematically explore and improve the usability of the longitudinal exposure-control mechanisms today. We proposed three improved exposure control mechanisms using real-world data to show that different longitudinal exposure control methods better respect the intent of some actors. Specifically, we evaluated the effectiveness of—age-based withdrawal, inactivity-based withdrawal and anonymization. We point out that inactivity-based withdrawal is an improvement over age-based withdrawal. However, anonymization provides a good balance between respecting the intent of owner as well as other actors. In the next chapter, we will conclude this thesis by summarizing the results of our investigation so far and pointing out the future work in this space.

CHAPTER 7

Concluding discussion

Today billions of online users use OSM sites like Facebook and Twitter to create and share hundreds of billions of pieces of content. The popularity of these OSM sites has renewed a discussion of online privacy. Particularly, OSMs created a fundamental shift in content sharing—prior to the advent of OSMs in the online world the majority of users were content consumers, however in OSMs users are content creators as well as managers. Thus, in addition to uploading content in these OSMs, users are also expected to manage the privacy of each and every piece of content they upload. Consequently, OSM users are facing a privacy management crisis today as pointed out by news reports, blog posts and research studies.

In this thesis, we addressed a key research challenge in this space: *How do we provide improved privacy controls to OSM users for managing privacy of their social content from other users?* We identified that the majority of OSMs today use access control as the dominant model to provide privacy management tools to their users. However, while useful, access control model is insufficient to capture many real-world privacy violations as well as multiple dimensions of privacy pointed out in prior privacy theories. To that end, in order to extend the current access control mechanisms, we introduce the model of exposure control. We define the exposure of a content as the set of people who actually view the content. We showed that exposure control in conjunction with access control effectively captures majority of the dimensions of privacy pointed out in earlier theories. Furthermore, exposure control captures privacy violations not captured by access control. Thus we argue that OSM

operators should provide users better tools to control the exposure of their content. To that end, we investigate how to better control exposure in three different real-world scenarios in this thesis.

Firstly, we identified that, today an easy and widely available method to better control exposure in OSMs for users is to specify social access control lists (social ACLs, or SACLs); SACLs are conservative access control lists which allow only a subset of their social contacts to access a content. However, so far no study attempted to understand how SACLs are used in the real world and how to improve their usability. To that end, in this thesis we present the first large-scale study of real-world in-use SACLS of more than 1,000 Facebook users. We found that the set of friends included/excluded in SACLs shows little correlation with either profile information, activity or social network structure—Thus making it hard to predict SACL membership. Fortunately, we found that SACLs are often reused, thus caching recent SACLs and making them available to users is likely to improve the usability of exposure control method via SACLs.

Secondly, we found that when users upload content in OSMs, third party crawlers or bots are quite possibly not in their expected exposure set. However, these crawlers collect and hoards the data of hundreds of millions of users and violate the exposure control of content. To that end, we propose Genie, a system for OSM operators to protect privacy via controlling exposure from large-scale third party crawlers. Genie is a credit network based solution which exploits the fact that the browsing patterns of honest users (i.e., non-crawler users) and crawlers are very different: even a crawler with access to many accounts needs to make many more profile views per account than an honest user, and view profiles of users that are more distant in the social network. Our experiments using real-world data gathered from Renren (a Chinese OSM) show that Genie frustrates large-scale crawling while rarely impacting honest users. Effectively, Genie provides a way to control the exposure of OSM users' content from third party crawlers.

Finally, we investigated longitudinal exposure control in OSMs. 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. Our study leveraged real-world data from Twitter to discover 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. However, we also discovered that residual tweets (replies to Twitter posts from other users or posts mentioning a user) pose a severe problem to longitudinal exposure control mechanisms in Twitter. These residual tweets leak significant information about the inaccessible tweets and user accounts. We developed and deployed a Twitter app (available at `http://twitter-app.mpi-sws.org/footprint/`) for Twitter users to check the residual activities around their accounts and posts.

We identified a key research challenge while investigating how to improve longitudinal exposure control mechanisms—In OSMs content from multiple users are effectively intertwined (via replies, comments, tags etc.). Thus in order to asses and improve longitudinal control methods we need to understand the deletion ecosystem in OSMs. To that end, we identified relevant actors other than the uploading user in the context of old content withdrawal and present a detailed survey of deletion mechanism in OSMs. We present concrete research questions in order to asses and improve longitudinal exposure control mechanisms which are part of deletion ecosystem. We present three proof of concept improvements in the deletion ecosystem of Twitter to show the potential of our approach.

One venue of future work is to provide a framework to compare and improve the longitudinal exposure control mechanisms via investigating the intent of all the actors involved (e.g., via semi structured surveys). Two more general venues of future work, pointed out by this thesis are (i) to expand the model of exposure control and (ii) to apply exposure control in other scenarios. Some researchers are already trying to do the first [42] by providing a probabilistic notion of exposure. A concrete example of the second direction is to improve search exposure—content retrieval mechanisms like search might increase

unwanted exposure of old content, e.g, by unearthing decades old usenet posts of a user. Thus, a concrete venue might be to use the concept of exposure control to argue about privacy violation via search and design exposure control mechanisms to limit search exposure. A very recent work tried to tackle search exposure [76].

To conclude, the concept of exposure presented in this thesis provides a novel way to argue about privacy breaches in OSMs and lays the ground work for improving the privacy management mechanisms. We strongly believe that, our exposure control model is an important step towards providing better privacy control to OSM users. In fact, exposure control provides a generic framework to the OSM operators and privacy researchers to better understand the privacy requirements of their users and improve upon the existing privacy management mechanisms.

BIBLIOGRAPHY

[1] arXiv.org. https://arxiv.org/. (Last accessed on June 2017).

[2] Cyber Dust. https://www.cyberdust.com/. (Last accessed on June 2017).

[3] 10 Amazing 4chan Statistics and Facts (December 2016). http://expandedramblings.com/index.php/4chan-statistics-facts/. (Last accessed on July 2017).

[4] 45,000 Facebook accounts compromised: What to know. http://www.pcworld.com/article/247370/ramnit_zeus_hybrid_compromises_45_000_facebook_accounts_what_you_should_know.html. (Last accessed on June 2017).

[5] 4chan. http://www.4chan.org/. (Last accessed on June 2017).

[6] Crawl packages for social networks. http://80legs.com/crawl-packages-social-networks.html. (Last accessed on December 2012).

[7] 83 million Facebook accounts are fakes and dupes. http://edition.cnn.com/2012/08/02/tech/social-media/facebook-fake-accounts/index.html. (Last accessed on June 2017).

[8] A standard for robot exclusion. http://www.robotstxt.org/orig.html. (Last accessed on June 2017).

[9] L. V. Ahn, M. Blum, N. J. Hopper, and J. Langford. CAPTCHA: Using Hard AI Problems For Security. In *Proceedings of the 22nd Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT'03)*, Warsaw, Poland, May 2003.

[10] 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):9:1–9:33, 2012.

[11] 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)*, San Antonio, Texas, USA, February 2013.

[12] I. Altman. *The Environment and Social Behavior: Privacy, Personal Space, Territory, Crowding*. Brooks/Cole Pub. Co., 1975.

[13] I. Altman. Privacy Regulation: Culturally Universal or Culturally Specific? *Journal of Social Issues*, 33(3):66–84, 1977.

[14] I. Altman. *Toward a Transactional Perspective: A Personal Journey*, pages 225–255. Springer US, 1990.

[15] S. Amershi, J. Fogarty, and D. Weld. Regroup: Interactive Machine Learning for On-demand Group Creation in Social Networks. In *Proceedings of the 30th SIGCHI Conference on Human Factors in Computing Systems (CHI'12)*, Austin, TX, USA, May 2012.

[16] Amazon Mechanical Turk terms of service. `https://requester.mturk.com/mturk/help?helpPage=policies`. (Last accessed on June 2017).

[17] G. M. Anderman and M. Rogers;. *Translation today : trends and perspectives*. Beijing : Foreign Language Teaching and Research Press, 2006.

[18] Anonymous Messaging App Whisper Raising New Funds. `http://fortune.com/2016/09/16/`

anonymous-messaging-app-whisper-raising-new-funds/. (Last accessed on July 2017).

[19] V. Arnaboldi, M. Conti, A. Passarella, and F. Pezzoni. Analysis of Ego Network Structure in Online Social Networks. In *Proceedings of the 4th ASE/IEEE International Conference on Social Computing (SocialCom'12)*, Amsterdam, The Netherlands, September 2012.

[20] Privacy is not Access Control (But then what is it?). http://33bits.org/2010/02/13/privacy-is-not-access-control/.

[21] 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)*, Newcastle, UK, July 2013.

[22] L. Backstrom, E. Bakshy, J. M. Kleinberg, T. M. Lento, and I. Rosenn. Center of Attention: How Facebook Users Allocate Attention Across Friends. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, Barcelona, Spain, July 2011.

[23] R. Baden, A. Bender, N. Spring, B. Bhattacharjee, and D. Starin. Persona: An Online Social Network with User-defined Privacy. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM'09)*, Barcelona, Spain, August 2009.

[24] A. Barth, A. Datta, J. C. Mitchell, and H. Nissenbaum. Privacy and Contextual Integrity: Framework and Applications. In *Proceedings of the 27th IEEE Symposium on Security and Privacy (IEEE S&P'06)*, Oakland, CA, USA, May 2006.

[25] 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)*, Berlin, Germany, November 2013.

[26] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing User Behavior in Online Social Networks. In *Proceedings of the 9th ACM/USENIX Internet Measurement Conference (IMC'09)*, Chicago, IL, USA, November 2009.

[27] 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)*, Paris, France, May 2013.

[28] M. S. Bernstein, A. Monroy-Hernández, D. Harry, P. André, K. Panovich, and G. G. Vargas. 4chan and /b: An Analysis of Anonymity and Ephemerality in a Large Online Community. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, Barcelona, Spain, July 2011.

[29] 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)*, Atlanta, GA, April 2010.

[30] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of community hierarchies in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[31] Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti. Characterizing and modelling popularity of user-generated videos. *Performance Evaluation*, 68(11):1037–1055, 2011.

[32] D. Boyd. Facebook's Privacy Trainwreck. *Convergence: The International Journal of Research into New Media Technologies*, 14(1):13–20, 2008.

[33] D. Boyd and E. Hargittai. Facebook privacy settings: Who cares? *First Monday*, 15(8), 2010.

[34] P. G. K. Brendan Meeder, Jennifer Tam and L. F. Cranor. RT@ IWantPrivacy: Widespread Violation of Privacy Settings in the Twitter Social Network. In *Proceedings of the Web 2.0 Privacy and Security Workshop (W2SP)*, Oakland, CA, USA, May 2010.

[35] C. Brooks. Facebook Horror Stories: How A Post Can Cost You Your Job. `http://www.businessnewsdaily.com/1329-facebook-horror-stories-how-a-post-can-cost-you-your-job.html`. (Last accessed on June 2017).

[36] C. Canali, M. Colajanni, and R. Lancellotti. Data Acquisition in Social Networks: Issues and Proposals. In *Proceedings of the International Workshop on Services and Open Sources (SOS'11)*, 2011.

[37] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media(ICWSM'10)*, Washington, DC, USA, May 2010.

[38] C. Charles. 3 Stories that should make you check your Facebook privacy settings. `http://www.thatsnonsense.com/3-stories-make-check-facebook-privacy-settings/`. (Last accessed on June 2017).

[39] S. Chen. Can Facebook get you fired? Playing it safe in the social media world. `http://www.cnn.com/2010/LIVING/11/10/facebook.fired.social.media.etiquette/index.html`. (Last accessed on June 2017).

[40] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70:1–6, 2004.

[41] D. Correa, L. A. Silva, M. Mondal, F. Benevenuto, and K. P. Gummadi. The Many Shades of Anonymity: Characterizing Anonymous Social Media Content. In *Proceedings of The 9th International AAAI Conference on Weblogs and Social Media (ICWSM'15)*, Oxford, UK, May 2015.

[42] A. Cortese and A. Masoumzadeh. Modeling Exposure in Online Social Networks. In *Proceedings of the 15th International Conference on Privacy, Security and Trust (PST '17)*, Calgary, Alberta, Canada, August 2017.

[43] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

[44] P. Dandekar, A. Goel, R. Govindan, and I. Post. Liquidity in Credit Networks: A Little Trust Goes a Long Way. In *Proceedings of the 12th ACM Conference on Electronic Commerce (EC'11)*, San Jose, CA, USA, June 2011.

[45] G. Danezis and P. Mittal. SybilInfer: Detecting Sybil Nodes using Social Networks. In *Proceedings of the 16th Annual Network and Distributed System Security Symposium (NDSS'09)*, San Diego, CA, USA, Feb 2009.

[46] Deleting My Teenage Tweets: A Student Journalists Perspective. `http://ajr.org/delete-tweets/`. (Last accessed on June 2017).

[47] Details of 100 million Facebook users published online. `http://www.nbcnews.com/id/38463013/ns/technology_and_science-security/t/details-million-facebook-users-published-online/`. (Last accessed on June 2017).

[48] 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)*, Lugano, Switzerland, March 2012.

[49] E. A. Dinic. An algorithm for the solution of the max-flow problem with the polynomial estimation. *Doklady Akademii Nauk SSSR*, 1970.

[50] D. do B. DeFigueiredo and E. T. Barr. TrustDavis: A Non-Exploitable Online Reputation System. In *Proceedings of the 7th IEEE International Conference on E-Commerce Technology (CEC'05)*, Munich, Germany, July 2005.

[51] R. I. M. Dunbar. The Social Brain Hypothesis. *Evolutionary Anthropology*, 6(5):178–190, 1998.

[52] A. Etzioni. Despite Facebook, privacy is far from dead. `http://www.cnn.com/2012/05/25/opinion/etzioni-facebook-privacy/`. (Last accessed on June 2017).

[53] Facebook. `http://www.facebook.com`. (Last accessed on June 2017).

[54] Facebook's Growth In The Past Year. `https://www.facebook.com/pg/facebook/photos/?tab=album&album_id=10151908376636729`. (Last accessed on June 2017).

[55] How do I know who's seen each post or message in a group? `https://www.facebook.com/help/409719555736128`. (Last accessed on June 2017).

[56] Facebook Messenger. `https://www.messenger.com/`. (Last accessed on June 2017).

[57] Facebook smart lists. `https://blog.facebook.com/blog.php?post=10150278932602131`. (Last accessed on January 2014).

[58] Number of active users at Facebook over the years. http://finance.yahoo.com/news/number-active-users-facebook-over-years-214600186--finance.html. (Last accessed on June 2017).

[59] L. Fang and K. LeFevre. Privacy Wizards for Social Networking Sites. In *Proceedings of the 19th International World Wide Web Conference (WWW'10)*, Releigh, NC, USA, April 2010.

[60] The day has come: Facebook pushes people to Go public. https://readwrite.com/2009/12/09/facebook_pushes_people_to_go_public/. (Last accessed on June 2017).

[61] What are the privacy options for groups? https://www.facebook.com/help/www/220336891328465. (Last accessed on June 2017).

[62] Facebook API. http://developers.facebook.com/docs/reference/api/, 2011. (Last accessed on June 2017).

[63] Detailed History of Facebook Changes 2004-12. https://www.jonloomer.com/2012/05/06/history-of-facebook-changes/. (Last accessed on June 2017).

[64] The Evolution of Privacy on Facebook – Changes in default profile settings over time. http://mattmckeon.com/facebook-privacy/. (Last accessed on June 2017).

[65] When I post something, how do I choose who can see it? https://www.facebook.com/help/120939471321735. (Last accessed on June 2017).

[66] New Tools to Control Your Experience. https://www.facebook.com/notes/facebook/new-tools-to-control-your-experience/196629387130/. (Last accessed on June 2017).

[67] Making Photo Tagging Easier. https://www.facebook.com/notes/facebook/making-photo-tagging-easier/467145887130/. (Last accessed on June 2017).

[68] Facebook's New Privacy Changes: The Good, The Bad, and The Ugly. https://www.eff.org/deeplinks/2009/12/facebooks-new-privacy-changes-good-bad-and-ugly. (Last accessed on June 2017).

[69] Tag Friends in Your Status and Posts. https://www.facebook.com/notes/facebook/tag-friends-in-your-status-and-posts/109765592130. (Last accessed on June 2017).

[70] F. Figueiredo, F. Benevenuto, and J. M. Almeida. The Tube over Time: Characterizing Popularity Growth of Youtube Videos. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM'11)*, Hong Kong, China, February 2011.

[71] Friendlist Manager. http://apps.facebook.com/friendlist_manager, 2012. (Last accessed on January 2014).

[72] Friendlist Manager. http://friendlist-manager.mpi-sws.org/. (Last accessed on January 2014).

[73] More Privacy Options. https://blog.facebook.com/blog.php?post=11519877130. (Last accessed on January 2014).

[74] Facebook's New Groups, Dashboards, and Downloads Explained. http://www.fastcompany.com/1693443/facebooks-new-groups-dashboards-and-downloads-explained-video. (Last accessed on June 2017).

[75] A. Ghosh, M. Mahdian, D. M. Reeves, D. M. Pennock, and R. Fugger. Mechanism Design on Trust Networks. In *Proceedings of the 3rd International Workshop on Internet and Network Economics (WINE'07)*, San Diego, CA, USA, December 2007.

[76] A. Ghosh, M. Mahdian, D. M. Reeves, D. M. Pennock, and R. Fugger. Learning to Un-Rank: Quantifying Search Exposure for Users in Online Communities. In *Proceedings of the 26th International Conference on Information and Knowledge Management (CIKM'17)*, Singapore, November 2017.

[77] M. Gjoka, M. Kurant, C. Butts, and A. Markopoulou. Practical Recommendations on Crawling Online Social Networks. *IEEE Journal on Selected Areas in Communications*, 29(9):1872–1892, 2011.

[78] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira. Exploiting Innocuous Activity for Correlating Users Across Sites. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*, Rio de Janeiro, Brazil, May 2013.

[79] A. V. Goldberg and R. E. Tarjan. A new Approach to the Maximum-Flow Problem. In *Proceedings of the 18th annual ACM Symposium on Theory of Computing (STOC'86)*, 1986.

[80] Google Plus rate limiting. https://developers.google.com/console/help/#cappingusage. (Last accessed on December 2012).

[81] D. Gregor and A. Lumsdaine. The Parallel BGL: A generic library for distributed graph computations. In *Proceedings of the Parallel Object-Oriented Scientific Computing (POOSC'05)*, 2005.

[82] S. Guha, K. Tang, and P. Francis. NOYB: Privacy in Online Social Networks. In *Proceedings of the 1st ACM workshop on Online social networks (WOSN'08)*, Seattle, WA, USA, August 2008.

[83] Hacker proves Facebook's public data Is public. `https://techcrunch.com/2010/07/28/hacker-proves-facebooks-public-data-is-public/`. (Last accessed on June 2017).

[84] R. A. Hill and R. I. M. Dunbar. Social network size in humans. *Human Nature*, 14(1):53–72, 2003.

[85] 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.

[86] L. Hong, O. Dan, and B. D. Davison. Predicting Popular Messages in Twitter. In *Proceedings of the 20th International World Wide Web Conference (WWW'11)*, Hyderabad, India, March 2011.

[87] A. Hutchinson. Social Media, Privacy and Scams – 3 Recent Cases That Highlight the Need to Take Care. `http://www.socialmediatoday.com/social-networks/adhutchinson/2015-08-08/social-media-privacy-and-scams-3-recent-cases-highlight-need`. (Last accessed on June 2017).

[88] Inside a Facebook botnet. http://www.itworld.com/article/2832592/it-management/inside-a-facebook-botnet.html. (Last accessed on June 2017).

[89] Instagram. https://www.instagram.com/. (Last accessed on June 2017).

[90] 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)*, New York, NY, USA, March 2016.

[91] J. Jiang, C. Wilson, X. Wang, P. Huang, W. Sha, Y. Dai, and B. Y. Zhao. Understanding Latent Interactions in Online Social Networks. In *Proceedings of the 10th ACM/USENIX Internet Measurement Conference (IMC'10)*, Melbourne, Australia, November 2010.

[92] 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)*, Washington, DC, USA, July 2012.

[93] L. Kagal and H. Abelson. Access Control is an Inadequate Framework for Privacy Protection. In *W3C Workshop on Privacy for Advanced Web APIs)*, 2010.

[94] S. Kairam, M. Brzozowski, D. Huffaker, and E. Chi. Talking in Circles: Selective Sharing in Google+. In *Proceedings of the 30th SIGCHI Conference on Human Factors in Computing Systems (CHI'12)*, Austin, TX, USA, May 2012.

[95] A. M. Kaplan and M. Haenlein. Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1):59–68, 2010.

[96] R. M. Karp. Reducibility Among Combinatorial Problems. In *Complexity of Computer Computations*, 1972.

[97] P. G. Kelley, R. Brewer, Y. Mayer, L. F. Cranor, and N. Sadeh. An Investigation into Facebook Friend Grouping. In *Proceedings of the 13th IFIP TC 13 International Conference on Human-computer Interaction (INTERACT'11)*, Lisbon, Portugal, September 2011.

[98] E. A. Kolek and D. Saunders. Online Disclosure: An Empirical Examination of Undergraduate Facebook Profiles. *Journal of Student Affairs Research and Practice*, 45(1):1–25, 2008.

[99] 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)*, Dublin, Ireland, June 2012.

[100] K. Lerman and T. Hogg. Using a Model of Social Dynamics to Predict Popularity of News. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*, Raleigh, NC, USA, April 2010.

[101] An Open Letter from Mark Zuckerberg. `https://www.facebook.com/notes/facebook/an-open-letter-from-mark-zuckerberg/2208562130/`. (Last accessed on June 2017).

[102] Linear regression. `http://en.wikipedia.org/wiki/Linear_regression`. (Last accessed on June 2017).

[103] 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)*, Berlin, Germany, November 2011.

[104] 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), Ann Arbor, MI, USA, June 2014.

[105] Y. Liu, M. Mondal, B. Viswanath, M. Mondal, K. P. Gummadi, and A. Mislove. Simplifying Friendlist Management. In *Proceedings of the 21st International World Wide Web Conference (WWW'12)*, Lyon, France, April 2012.

[106] 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/`. (Last accessed on June 2017).

[107] 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.

[108] G. Malewicz, M. H. Austern, A. J. Bik, J. C. Dehnert, I. Horn, N. Leiser, and G. Czajkowski. Pregel: A System for Large-scale Graph Processing. In *Proceedings of the International Conference on Management of Data (SIGMOD'10)*, Indianapolis, IN, USA, June 2010.

[109] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[110] 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, 2003.

[111] W. Mason and S. Suri. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44(1):1–23, 2011.

[112] 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)*, Washington, DC, USA, July 2012.

[113] W. Meng, R. Ding, S. P. Chung, S. Han, and W. Lee. The Price of Free: Privacy Leakage in Personalized Mobile In-Apps Ads. In *Proceedings of the 23rd Annual Network and Distributed System Security Symposium (NDSS'16)*, San Diego, CA, USA, February 2016.

[114] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Growth of the Flickr Social Network. In *Proceedings of the 1st ACM workshop on Online social networks (WOSN'08)*, Seattle, WA, USA, August 2008.

[115] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proceedings of the 7th ACM/USENIX Internet Measurement Conference (IMC'07)*, San Diego, CA, October 2007.

[116] A. Mislove, A. Post, K. P. Gummadi, and P. Druschel. Ostra: Leveraging Trust to Thwart Unwanted Communication. In *Proceedings of the 5th Symposium on Networked Systems Design and Implementation (NSDI'08)*, San Francisco, CA, USA, April 2008.

[117] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You Are Who You Know: Inferring User Profiles in Online Social Networks. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM'10)*, New York, NY, USA, Feb 2010.

[118] A. Mohaisen, A. Yun, and Y. Kim. Measuring the Mixing Time of Social Graphs. In *Proceedings of the 10th ACM/USENIX Internet Measurement Conference (IMC'10)*, Melbourne, Australia, November 2010.

[119] Most famous social network sites worldwide as of April 2017, ranked by number of active users (in millions). https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. (Last accessed on July 2017).

[120] M. Motoyama, K. Levchenko, C. Kanich, D. McCoy, G. M. Voelker, and S. Savage. Re: CAPTCHAs—Understanding CAPTCHA-solving services in an economic context. In *Proceedings of the 19th USENIX conference on Security (USENIX Security'10)*, Washington, DC, USA, August 2010.

[121] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker. Dirty Jobs: The Role of Freelance Labor in Web Service Abuse. In *Proceedings of the 20th USENIX conference on Security (USENIX Security'11)*, San Francisco, CA, USA, August 2011.

[122] L. Mullen. Predicting Gender Using Historical Data. https://cran.r-project.org/web/packages/gender/vignettes/predicting-gender.html, 2015. (Last accessed on June 2017).

[123] Netflix-AOL data leak. https://www.cnet.com/news/aol-netflix-and-the-end-of-open-access-to-research-data/. (Last accessed on June 2017).

[124] H. Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2010.

[125] D. Noyes. The Top 20 Valuable Facebook Statistics. http://zephoria.com/social-media/top-15-valuable-facebook-statistics/. (Last accessed on January 2014).

[126] X. Page. Contextual Integrity and Preserving Relationship Boundaries in Location-Sharing Social Media. In *Proceedings of the Measuring Networked Social Privacy Workshop*, San Antonio, Texas, USA, February 2013.

[127] L. Palen and P. Dourish. Unpacking "Privacy" for a Networked World. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*, Fort Lauderdale, Florida, USA, April 2003.

[128] S. Petronio. *Boundaries of Privacy: Dialectics of Disclosure*. State University of New York Press, 2002.

[129] S. Petrovič, M. Osborne, and V. Lavrenko. I Wish I Didn't Say That! Analyzing and Predicting Deleted Messages in Twitter. *arXiv preprint*, arXiv:1305.3107, 2013.

[130] P. Pons and M. Latapy. Computing Communities in Large Networks Using Random Walks. *Journal of Graph Algorithms and Applications*, 10(2):191–218, 2006.

[131] A. Post, V. Shah, and A. Mislove. Bazaar: Strengthening User Reputations in Online Marketplaces. In *Proceedings of the 8th Symposium on Networked Systems Design and Implementation (NSDI'11)*, Boston, MA, USA, March 2011.

[132] Public posting now the default on Facebook. `https://www.wired.com/2009/12/facebook-privacy-update/`. (Last accessed on June 2017).

[133] D. Quercia and S. Hailes. Sybil Attacks Against Mobile Users: Friends and Foes to the Rescue. In *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM'10)*, San Diego, CA, USA, March 2010.

[134] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[135] Rate Limiting for Yahoo! Search Web Services. `http://developer.yahoo.com/search/rate.html`. (Last accessed on June 2017).

[136] Rebecca Black phenomenon. http://youtube-trends.blogspot.de/2011/03/rebecca-black-phenomenon.html. (Last accessed on June 2017).

[137] Reddit. http://www.reddit.com. (Last accessed on June 2017).

[138] Renren. http://www.renren.com. (Last accessed on June 2017).

[139] J. Ronson. 'Overnight, everything I loved was gone': the internet shaming of Lindsey Stone. https://www.theguardian.com/technology/2015/feb/21/internet-shaming-lindsey-stone-jon-ronson. (Last accessed on June 2017).

[140] People Manage Their Privacy On Facebook Naturally. http://www.sciencedaily.com/releases/2009/04/090420084957.htm. (Last accessed on June 2017).

[141] More Facebook Friends Means More Stress, Says Report. http://www.sciencedaily.com/releases/2012/11/121126131218.htm. (Last accessed on June 2017).

[142] Sean Casey, President, Nielsen Social. 2016 nielsen social media report. http://www.nielsen.com/us/en/insights/reports/2017/2016-nielsen-social-media-report.html. (Last accessed on June 2017).

[143] Skype. https://www.skype.com/en/. (Last accessed on June 2017).

[144] 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.

[145] Snapchat. http://www.snapchat.com. (Last accessed on June 2017).

[146] D. J. Solove. A Taxonomy of Privacy. *University of Pennsylvania Law Review*, 154(3):477–560, 2006.

[147] D. J. Solove. *Understanding Privacy*. Harvard University Press, 2008.

[148] Spokeo privacy complaints and legal troubles. `http://en.wikipedia.org/wiki/Spokeo#Privacy_and_safety_concerns`. (Last accessed on June 2017).

[149] Reddit stats at a glance. `http://web.archive.org/web/20160416054155/https://www.reddit.com/about/`. (Last accessed on July 2017).

[150] T. Stein, E. Chen, and K. Mangla. Facebook Immune System. In *Proceedings of the 4th Workshop on Social Network Systems (SNS'11)*, Salzburg, Austria, April 2011.

[151] K. Strater and H. R. Lipford. Strategies and Struggles with Privacy in an Online Social Networking Community. In *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity (BCS-HCI'08)*, Liverpool, UK, September 2008.

[152] F. Stutzman, R. Gross, and A. Acquisti. Silent Listeners: The Evolution of Privacy and Disclosure on Facebook. *Journal of Privacy and Confidentiality*, 4(2):7–41, 2012.

[153] J. D. Sutter. Some quitting Facebook as privacy concerns escalate. `http://www.cnn.com/2010/TECH/05/13/facebook.delete.privacy/index.html`. (Last accessed on June 2017).

[154] G. Szabo and B. A. Huberman. Predicting the Popularity of Online Content. *Communications of ACM*, 53(8):80–88, 2010.

[155] M. Thelwall. Homophily in MySpace. *Journal of the American Society for Information Science and Technology*, 60(2):219–231, 2009.

[156] N. Tran, J. Li, L. Subramanian, and S. S. Chow. Optimal Sybil-resilient node admission control. In *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM'11)*, Shanghai, China, April 2011.

[157] N. Tran, B. Min, J. Li, and L. Subramanian. Sybil-Resilient Online Content Voting. In *Proceedings of the 6th Usenix Symposium on Networked Systems Design and Implementation (NSDI'09)*, Boston, MA, USA, April 2009.

[158] Z. Tufekci. Facebook, Youth and Privacy in Networked Publics. In *Proceedings of the 6th International Conference on Weblogs and Social Media (ICWSM'12)*, Dublin, Ireland, June 2012.

[159] Tumblr. `https://www.tumblr.com/`. (Last accessed on June 2017).

[160] Twitter. `http://www.twitter.com`. (Last accessed on June 2017).

[161] Twitter rate limiting. `https://dev.twitter.com/rest/public/rate-limiting`. (Last accessed on June 2017).

[162] Twitter team. The streaming apis. `https://dev.twitter.com/streaming/overview`. (Last accessed on June 2017).

[163] Viber. `https://www.viber.com/`. (Last accessed on June 2017).

[164] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the Evolution of User Interaction in Facebook. In *Proceedings of the 2nd ACM workshop on Online social networks (WOSN'09)*, Barcelona, Spain, August 2009.

[165] B. Viswanath, M. Mondal, A. Clement, P. Druschel, K. P. Gummadi, A. Mislove, and A. Post. Exploring the design space of social network-based Sybil defense. In

Proceedings of the 4th International Conference on Communication Systems and Network (COMSNETS'12), Bangalore, India, December 2012.

[166] B. Viswanath, M. Mondal, K. P. Gummadi, A. Mislove, and A. Post. Canal: Scaling Social Network-based Sybil Tolerance Schemes. In *Proceedings of the 7th European Conference on Computer Systems (EuroSys'12)*, Bern, Switzerland, April 2012.

[167] B. Viswanath, A. Post, K. P. Gummadi, and A. Mislove. An Analysis of Social Network-based Sybil Defenses. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM'10)*, New Delhi, India, August 2010.

[168] S. D. Warren and L. D. Brandeis. The Right to Privacy. *Harvard Law Review*, 4(5):193–220, 1890.

[169] A. Westin. *Privacy and Freedom*. Bodley Head, 1970.

[170] Whatsapp. `https://www.whatsapp.com/`. (Last accessed on June 2017).

[171] Whisper. `http://whisper.sh/`. (Last accessed on June 2017).

[172] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao. User Interactions in Social Networks and Their Implications. In *Proceedings of the 4th ACM European Conference on Computer Systems (Eurosys'09)*, Nuremberg, Germany, April 2009.

[173] C. Wilson, A. Sala, J. Bonneau, R. Zablit, and B. Y. Zhao. Don't Tread on Me: Moderating Access to OSN Data with SpikeStrip. In *Proceedings of the 3rd ACM SIGCOMM Workshop on Social Networks (WOSN'10)*, Boston, MA, USA, June 2010.

[174] Q. Xiao, H. H. Aung, and K.-L. Tan. Towards Ad-hoc Circles in Social Networking Sites. In *Proceedings of the 2nd ACM SIGMOD Workshop on Databases and Social Networks(DBSocial'12)*, Scottsdale, AZ, USA, May 2012.

[175] YouTube. http://www.youtube.com. (Last accessed on June 2017).

[176] How do I share a private YouTube video with someone? https://webapps.stackexchange.com/questions/7588/how-do-i-share-a-private-youtube-video-with-someone. (Last accessed on October 2017).

[177] H. Yu, P. B. Gibbons, M. Kaminsky, and F. Xiao. SybilLimit: A Near-optimal Social Network Defense Against Sybil Attacks. In *Proceedings of the 29th IEEE Symposium on Security and Privacy (IEEE S&P'08)*, Oakland, CA, USA, May 2008.

[178] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman. SybilGuard: Defending Against Sybil Attacks via Social Networks. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM'06)*, Pisa, Italy, September 2006.

[179] 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.

[180] 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)*, Montreal, Canada, April 2016.