How Busy Are You? Predicting the Interruptibility Intensity of Mobile Users

Fengpeng Yuan, Xianyi Gao and Janne Lindqvist
Rutgers University

ABSTRACT

Smartphones frequently notify users about newly available messages or other notifications. It can be very disruptive when these notifications interrupt users while they are busy. Our work here is based on the observation that people usually exhibit different levels of busyness at different contexts. This means that classifying users' interruptibility as a binary status, interruptible or not interruptible, is not sufficient to accurately measure their availability towards smartphone interruptions. In this paper, we propose, implement and evaluate a two-stage hierarchical model to predict people's interruptibility intensity. Our work is the first to introduce personality traits into interruptibility prediction model, and we found that personality data improves the prediction significantly. Our model bootstraps the prediction with similar people's data, and provides a good initial prediction for users whose individual models have not been trained on their own data yet. Overall prediction accuracy of our model can reach 66.1%.

ACM Classification Keywords

H.1.2. Models and Principles: User/Machine Systems; H.5.2. Information Interfaces and Presentation (e.g. HCI): User Interfaces

Author Keywords

Interruptibility; Notifications; Predictive Models; Context

INTRODUCTION

Smartphones frequently notify users about newly available information such as incoming calls, messages, emails, and notifications. It can be very disruptive when these notifications interrupt users while they are busy. Studies have shown that inappropriate interruptions not only annoy users, but also decrease their productivity [65] and affect their emotions and social attribution [1]. Hence, it is important to understand and select appropriate time and context to interrupt users.

Ideally, a smartphone notification management system should be capable like a human secretary. Various studies [31, 21, 22, 73, 61] have been conducted focusing on the contextual information that affects user's interruptibility, and the effects of interruption content on interruptibility also have received attention [19, 39, 54]. However, most of the previous works focus on the binary classification of the interruptibility to notifications, while recent work [50] found that only for 45% of the scenarios binary classification is appropriate. In other words, people's interruptibility to notifications is distributed among many levels. This means that classifying users' interruptibility as a binary status, interruptible or not interruptible, is not sufficient to accurately measure their availability towards smartphone interruptions.

In this work, we conducted a field study to predict users' interruptibility intensity to mobile interruptions. By identifying user's context and interruption content, we propose a two-stage hierarchical model to predict users' interruptibility intensity. In the first stage, our model predicts whether users will react to mobile interruptions. If users do not react, they are indeed uninterruptible and will not involve in the interruption. If they react, our model further predicts how interruptible the users are in the second stage based on the type of tasks they are able to perform. Different tasks require different time and efforts, knowing all the factors of a task is useful for predicting people's interruptibility.

One important advantage of building a hierarchical model is the ability to collect additional feedback data from the users. Several previous works have used only sensor data to predict user's interruptibility (whether a user would react to a notification) [61, 63]. Our first stage in the hierarchical model applies a similar approach. However, many applications prefer interacting further with users. This requires learning more about user's context and current state (e.g. their moods) to predict their availability. Our second stage prediction serves for this purpose on further predicting user's interruptibility intensity based on their feedback.

In this paper, we use tasks as an example to learn about the extent of users' interruptibility. The method we present can be applied to other scenarios. For example, a mobile operating system (OS) initiates interaction between users and applications. Based on the importance and attention demand of the interaction [47], an OS can prioritize the interaction initiations by knowing users' fine-grained interruptibility levels.

We present the following four major contributions:

1) We propose a two-stage hierarchical interruptibility prediction model. In the first stage, our model predicts (with 75% accuracy) whether a user will react to an interruption

CHI'17, May 06 - 11, 2017, Denver, CO, USA

ACM 978-1-4503-4655-9/17/05...$15.00
DOI: http://dx.doi.org/10.1145/3025453.3025946

or notification based on mobile sensor data and personality traits. If the user reacts, it further predicts the user's interruptibility intensity for various tasks (requiring user involvement) in the second stage based on mobile sensor data and user's self-reported contextual information.

2) The evaluation results showed that our model can achieve an overall accuracy of 66.1% for interruptibility intensity prediction (with 60.9% mean accuracy).

3) We are the first to introduce people's personality into an interruptibility prediction model. On average, it improves the major measures (accuracy, precision, recall and F-measure) of tested classifiers over 10 percentage points in the first stage.

4) Our model solves the initial prediction problem, that is, how to predict when you do not have user data. To achieve this, in the second stage, our model uses the data of people who share similar personality with the user. Compared to the models using all the data of other people [23, 61], this reduces the training time significantly while maintains comparable prediction accuracy.

As minor contributions, we implemented a smartphone platform for this study, and we used a mobile social networking application (Foursquare) checkins to infer users' semantic places. We collected over 5000 interruptibility records from 22 participants over four weeks.

RELATED WORK

Interruptions within task execution impact users in various ways (e.g. emotion, productivity) and several studies have been conducted to mitigate the impacts. Fogarty et al. [24] suggested capturing task engagement to create reliable interruptibility prediction models. Iqbal et al. [37] leveraged the task structure characteristics to predict the cost of interruptions, however, it is usually difficult to know the task structure in advance. Delivering interruptions at task breakpoints is considered as effective to reduce the cost of interruptions [1, 38, 72, 32, 59, 60], as the workload decreases when reaching task boundaries. Bailey et al. [5] found that disruption can be largely mitigated by deferring notification to coarse boundaries during task execution. Horvitz et al. [34] also found that deferring notification could balance the information awareness and the cost of interruption. Although deferring notifications could reduce disruptiveness or cost of interruption, this approach bears the risk of missing important time-sensitive notifications.

Context provides useful information when estimating opportune time to deliver interruptions. Hudson et al. [36] and Fogarty et al. [22] found that simple sensors can provide the context to construct interruptibility estimation model, which can make robust estimates [23]. Horvitz et al. [33] utilized the visual and acoustical information captured by microphones and cameras of computers to infer the cost of interruption (COI), and later they developed Busybody [35] to predict the COI based on user's environment. Mühlenbrock et al. [56] employed various sensor data from PC, PDA and phone to detect user's availability and assist face-to-face interactions in office environments. Begole et al. [6] presented a prototype system, Lilsys, to infer user's unavailability by sensing users' actions and environment. Also, wearable sensors can be used to provide useful data to make accurate interruptibility prediction [52, 40]. Recently, Kim et al. [41] used sensor data about drivers' states and driving situation to infer the drivers' interruptibility when they are driving.

All of the above works focus on the observed environments (lab or office) and desktop notifications, however, this is different from mobile use during people's daily lives. With smartphones, people tend to receive more interruptions due to its ubiquitous characteristics, and this sparks the research on interruptibility in the field.

Phone calls are considered as one of the major interruption sources. Ter Hofte [75] employed the ESM method to explore the context that can be used to predict people's availability to a phone call. Böhmer et al. [8] allowed users to postpone an incoming phone call and introduced a smaller notification screen that reduced the annoyance perceived with interruptions. Rosenthal et al. [66] used ESM to collect data and train personalized models to learn when to silence the phone to avoid embarrassing interruptions. Smith et al. [71] considered dataset imbalance, error costs, user behaviors to recognize disruptive incoming calls, and developed RingLearn [70] to mitigate disruptive phone calls. Fisher et al. [20] built an in-context application for smartphone to create personalized interruptibility prediction model for phone calls. Although they achieved high prediction accuracy (96.12%), similar to other works [66, 71, 70], their model only predicts phone's ringer modes (on and off), which is a rough measurement of interruptibility. Moreover, the ringer mode may not reflect users' actual interruptibility, for example, when they forget to switch the mode when their interruptibility changes.

Mobile notifications are more pervasive and common than phone calls. Among tons of notifications, selecting an opportune time to deliver them is critical to their reception. Ho et al. [32] explored the perceived burden of mobile notifications and found that the burden was reduced during the transitions of two different physical activities, such as sitting to walking. Fischer et al. [18] used mobile interaction as indicators of opportune moments to deliver notifications, they found the interruptions were responded more quickly after a user finished an interaction with mobile phone, such as phone call, text message. Pielot et al. [62] found that the interaction with notification center, the screen activity are strong predictors of the responsiveness to instant messages, with such simple features, their model predicts whether a message will be viewed within a few minutes with 70.6% accuracy. Poppinga et al. [63] investigated the context factors' effects on interruptibility, they proposed a decision tree-based model to predict the opportune time to deliver mobile notification with 77.85% accuracy. Pejovic et al. [61] used a similar approach but with more features obtained from ESM survey to determine different aspects of interruptibility, including reaction presence, response time and sentiment. Based on the work of Poppinga et al. [63], Sarker et al. [67] used wearable sensor in addition to smartphone sensors to collect data. Their model achieves an accuracy of 74.7% (against base accuracy 50%) in predicting opportune

time to deliver interruptions. Mehrotra et al. [54] showed that by considering the notification content, the response time to notifications can be predicted with 70% average sensitivity.

Although the results of above works on predicting appropriate time to deliver notifications are promising, all of them mainly focus on the binary classification of interruptibility. However, Lopez et al. [50] found that for meetings, a binary classification was appropriate only 45% of the time. Pielot et al. [62] achieved the accuracy for two levels (binary) 70.6% and for three levels 61.6%. Züger [81] et al. predicted interruptibility of software developers in five levels; their model could achieve accuracy of 43.9% in the lab and 32.5% in the field. In addition to a binary prediction, Fogarty et al. [22] performed a 5-level interruptibility prediction. They achieved the prediction accuracy of 47.6% and 51.5% by using Naive Bayes and Decision Tree classifiers.

In our work, we conducted a field study towards understanding and predicting people's interruptibility intensity (levels) to mobile interruptions. We propose a two-stage hierarchical prediction model. In first stage, it predicts whether user is available to react to a notification. If the user reacts to the notification, it further predicts user's interruptibility intensity in second stage. In second stage, we take the ordering of the interruptibility ratings into consideration. The overall prediction accuracy can reach 66.1% (average 60.9%). Compared to previous works [22, 81], this is very competitive. We are also the first to take personality traits into interruptibility prediction model, and we found that the personality data significantly improve the prediction accuracy. For new users, our model uses data of other users who share similar personality. This solves the initial prediction problem of needing to train with the user before the app is usable. Our approach significantly reduces the training time of the model while maintains comparable prediction accuracy for predictions in the first days.

METHOD

Our study design focuses on investigating participants' interruptibility to mobile notifications in the field. We installed a smartphone app to the participants' phones to probe them during their daily lives.

Participants

We recruited participants using flyers, email lists and online advertisements. We required the participants to be at least 18 years old and active Android users. All the participants were compensated with a $30 gift card for completing the whole study, and they were enrolled in a raffle for two $50 gift cards. The study was approved by the Rutgers University IRB.

In total, we recruited 33 participants. Four participants withdrew during the study, and another seven participants were excluded from our analysis, as their survey response rates were less than 20%. This reduced our sample to 22 participants, in which we focus for the remainder of the paper.

Our participants' ages ranged from 18 to 27 (mean = 21.63, SD = 2.85, Mdn = 21); 9 participants were female and 20 were male. The participants' Android experiences ranged from one month to six years.

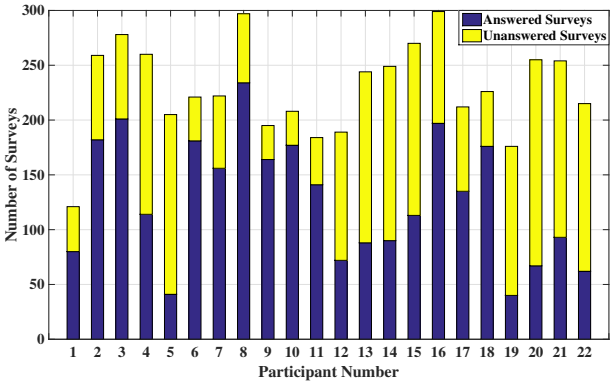

Figure 1. Answered and unanswered surveys of each participant. P8 received most surveys (297), while P1 only received 121 surveys. Because participants have different behavior and phone usage patterns, they received different numbers of surveys. P6 answered 85.1% of all the surveys, while P5 only answered 20%.

The field study was carried out for four weeks with each participant. Each of the participants received roughly 8 - 10 survey notifications everyday. In total, 5039 surveys were sent out, while 2804 (55.6%) were answered. Acceptable response rates considered in previous studies range from 11% to 60% [30, 13, 46]. Therefore, our data was valid for analysis. The number of completed surveys for interruptibility level 1 to 5 are 280, 284, 1047, 299 and 894. Figure 1 shows the distribution of answered and unanswered surveys.

Apparatus

We developed an Android app for Android version 4.4.2 while ensuring compatibility for all the later versions. The app was used to initiate interruptions via a popup survey, and capture the user's context and record self-reported interruptibility level and notification information. Our app periodically uploaded the collected data to our remote server automatically via a background service. The data was uploaded only when the phone was connected to a WiFi network. This avoided the potential cellular data cost for the participants.

Procedure

To investigate the interruptibility of participants, we used Ecological Momentary Assessment (EMA) in our field study. EMA [69] is a research method that is used to collect self-reports of participants' behaviors, physiological and psychological states during their daily lives. The data collected in our study includes the participants' self-reports data from the EMA surveys and various sensor data, as described below.

During the study, participants were asked to come to the lab twice, once at the beginning of the study and once at the end of the study. During the study, the participants performed their daily activities as usual and responded to the app prompts.

In the first visit, we gave a brief introduction about the study. Then participants read and signed the consent form. The participants were also asked usual demographic questions.

After the interview, we installed our app to participants' phones. While we were installing the app, participants were asked to take a personality test-Mini IPIP [16], a short measure

of the Big Five personality traits (Extroversion, Agreeableness, Conscientiousness, Neuroticism and Openness) [12].

Once the app was installed, it generated a dialog that notified the participant to take the survey when triggering conditions were satisfied. The prompts were triggered by the state changes of participants [32, 59, 60]. We used the Google Activity Recognition API [27] to detect participant's states in the app. Table 1 lists the states we included: *in vehicle*, *on bicycle*, *running*, *still*, *tilting*, *unknown*, and *walking*. Unknown state meant that the API was not able to categorize the state. A survey notification would be slated to pop up when two consecutive states were different.

Whenever a survey notification popped up, a new record was created in the database. The record was updated when participants responded to the notification. After participants finished the surveys, all the survey answers and timestamps would be inserted to the database.

If a participant did not respond to the notification in ten seconds, the notification was pushed into the system notification bar. It remained there until the participant clicked it and completed the survey, or it would be replaced when a new survey popped up. The app also provided an option to cancel the survey notifications when participants were not available at the moment. In this case, the app only updated the response time for that survey record in the database. This option was only available before the notification was pushed into notification bar. Additionally, the survey notifications were only allowed to pop up between 8:00 AM and 10:00 PM [45]. To reduce the workload of the participants, the time interval between two surveys was at least one and half hours. This ensured that the survey prompted no more than ten times a day [30].

After four weeks of study, the participants were requested to come back for an exit interview and debriefing. We uninstalled the app for them and destroyed the associated data on the phone. Meanwhile, participants were asked to finish a short survey about the experience of the study, and other thoughts of the study.

Data Type	Description
Time $^{*,+}$	Survey pop-up time, reaction time and survey completion time.
Current and previous state $^{*,+}$	In vehicle, on bicycle, running, still, tilting, unknown, walking.
Location $^{*,+}$	Latitude and longitude, Foursquare checkins grouped into 10 categories.
Personality traits $^{*,+}$	Extroversion, agreeableness, conscientiousness, neuroticism and openness.
Mood^{+}	Using BMIS survey, scaled from unpleasant to pleasant.
Transportation method $^{+}$	By car, by air, by bike, by bus, by train, by subway, by boat, running, walking.
Current activity^{+}	Doing exercise, having a meal, on the phone, playing games, studying, taking a rest, talking, watching video, working, writing/checking emails, bored, others.
Who would you like to do a task for (task sender)? (Select one or more) $^{+}$	1) Immediate family members 2) Extended family members 3) People you are close to 4) People you live with 5) People you work with 6) People you do hobbies/activities with 7) Strangers [78]
What type of tasks would you like to do? (Select one or more) $^{+}$	Educational activities, help colleagues, help family members, help strangers, household activities, leisure and sports, organizational, civic, and religious activities, phone calls and mails, purchase goods, others (Time Use Survey [58]).
Preferred task duration (slider) $^{+}$	1 minute to 120 minutes [58]
Interruptibility level $^{+}$	1) Highly interruptible 2) Interruptible 3) Neutral 4) Uninterruptible 5) Highly uninterruptible).

Table 1. Data types used to model interruptibility. The time, current and previous state, and location were recorded automatically by the study app, and the rest were reported by the participants with the pop up surveys. Participants could select more than one option for question "who would you like to do a task for?" and "what type of tasks would you like to do?". The first stage of our model uses all the features marked with *, and the second stage of our model uses all of the features marked with $^{+}$.

Collected Data

Table 1 gives a summary of all the data types we collected during the study: time, current and previous state, location, mood (BMIS scale), interruptibility level, current transportation method (e.g. by car, by air, walking and so on), current activity, and questions related to tasks they could perform.

We sampled the participants' current mood in the survey using a brief mood introspection scale (BMIS) measurement [53]. Previous works obtained participants' mood by either directly asking them how happy or sad they are [61, 67], or using an ECG sensor to measure the mental state [81, 67]. These approaches can be of limited validity for measuring actual mood or not feasible for mobile users during their daily lives over an extended time. BMIS is widely used to measure mood with a pleasant-unpleasant scale, which uses a four-point Likert scale for each adjective.

Previous work has indicated that people's responsiveness to notifications is high when they are changing their locations [43], and people's interruptibility is correlated to the semantic places [67, 54]. Our study app collected the GPS location every five minutes. In addition, we used the Foursquare API [25] in our app towards obtaining uniform semantic names for the places the participants would go. Based on their current location, participants were asked to confirm the place they were currently at. If they were at a new place, they were asked to check-in at this place. Our app provided the check in function via the Foursquare API, the participants simply selected the venues from the list of places to check in when taking the survey. After the study, we manually categorized the foursquare places into 10 categories: entertainment, health and medical, home, professional, church, restaurant, shopping, transportation, work and others [7, 79, 49].

Participants were asked to rate their interruptibility intensity. We used a five-point Likert scale to record the interruptibility intensity levels as: highly interruptible (1), interruptible (2), neutral (3), uninterruptible (4), and highly uninterruptible (5).

People's interruptibility can be influenced by the interruption time [1] and also by the content and context of the interruption [10, 19, 54]. Therefore, we predict whether people will react to mobile notifications and also want to predict the extent of their availability and busyness. Towards this end, we asked whether participants would be able to perform some tasks. During the survey, if participants did not want to take tasks at

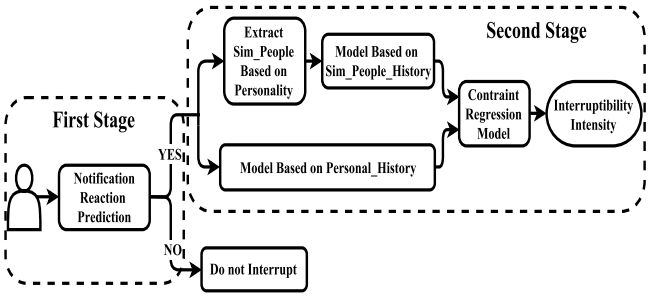

Figure 2. The overview of the two-stage hierarchical prediction model. The model first predicts whether a user is available to react to a notification, as shown in the left dashed box. If the user reacted to a notification, the model further predicts the user's interruptibility intensity for various tasks, as shown in the right dashed box. Otherwise, it will not allow disturbing the user. Most of previous works only focus on the first level (left dashed box) of the proposed model.

the moment, they did not have to do questions related to tasks. Otherwise, they were asked to answer whose task and what task they would like to perform, and the time they could spend on the task.

The *whose task* question asks about the task sender. The *what task* asks about the task content. We asked participants for whom they would be willing to do the task for because interpersonal relationships could affect how interruptions are perceived [29, 54]. We wanted to cover common activities that people do during their daily lives and used the American Time Use Survey [58] to find activities where and how people spend time. These common activities, listed in Table 1 can reflect people's availability. The time could be chosen from 1 to 120 minutes.

The task content in our survey can be mapped into the notification content from various applications, especially for crowdsourcing. For example, Chegg [9] app provides a platform for education tutoring and problem solving. Airtasker [3] app provides a platform for helping with chores, gathering activities, and errands such as cleaning, delivery, playing sports, and party planning. There are also web-based crowdsourcing systems involving physical activities and tasks, for example, *Pick-A-Crowd* [14], and other commercial apps for mobile on-demand workforce [74].

HIERARCHICAL PERSONALITY-DEPENDENT PREDICTION MODEL

In this section, we motivate and describe our hierarchical interruptibility prediction model.

Turner et al. [77] asked if there "could [be] a hybrid approach using personal and aggregated data reduce the training requirements for new users?" Our approach uses this simple idea, and we have implemented and evaluated its effectiveness. Further, we hypothesized that people's availability or busyness can also be related to potential tasks they could perform, in addition to other context and mental states. These ideas led us to build a model for predicting mobile users' interruptibility intensity.

Figure 2 gives a high-level overview of our two-stage hierarchical prediction model. We believe this approach elegantly solves the prediction problem. In the first stage, our model predicts whether a user is available to react to a notification. When a user reacts to a notification, our model further predicts what participants' interruptibility intensity is in the second stage. When the user does not react to a notification, they are classified as uninterruptible. This is also an important approach in itself.

Our first stage classification distinguishes situations when users are completely unavailable no matter what the interruptions are. In this stage, our model utilizes the sensor data to predict whether users present any reaction. In our study, we label situations in which participants do not complete our surveys as *completely unavailable* or *no reaction*. These cases mostly indicate bad moments to send interruptions, as users do not respond whether they miss them or ignore intentionally. On the other hand, the study would have less validity if we would ignore unresponded surveys as they also indicate unavailable situations. We label no response as unavailable in this stage to accommodate above tradeoffs.

Our second stage further predicts users' interruptibility intensity by using additional information provided when they interact with the interruptions, for example, mood. Such information is not available in the first stage. We ask the participants to report their task performing preferences at this moment in the survey. We use tasks to model the interruption content as task performing is a common example that reflects users' availability. It also acts as an example of users interacting with apps and providing additional information to assist on prediction. Participants need to determine whether or not they are available for listed tasks in Table 1. They could also just skip this part. This design has the advantage of evaluating multiple tasks at once compared to a single task prompt for a specific person.

The first step applies machine learning algorithms for a traditional binary classification of interruptibility. This is established by the sensor data and the personality test prior to installing the app. Thus, no self-reports beyond the personality test are used at this stage. The second stage is implemented with regression models, and is only applicable when the participant reacted to the pop up survey notification during the study. We used Weka for building the model and evaluating the classifiers and the regression models [28].

First Stage: Reaction Prediction

In this stage, we predict whether the user is available at all or completely busy. This is based on whether the user responded to the pop survey in our study or not. These reactions are used as prediction labels. For each survey record, we labeled it as *Reaction* if it was answered, otherwise, we labeled it as *Completely Unavailable*. No survey data is otherwise used in this stage. The prediction is based on the context information collected by the smartphone sensors, and users' personality traits. The contextual information includes weekend indicator, day of week, time of day, location, user's previous state and current state. User's personality traits include extroversion, agreeableness, openness, conscientiousness and neuroticism. Personality traits were obtained from the personality test when participants consented to join our study during their first visit.

The data used for classification has both numerical and categorical values. It is important to consider what types of classifiers would be able to perform well on such data since it is not obvious without testing them. SVM with nonlinear kernel functions (e.g. RBF) and tree based classifiers perform well and are generally robust on such kind of data. The attributes of our data may not be independent, examples of such data includes time and location. Domingos and Pazzani [15] showed that Bayesian classifiers can be used on such data and can achieve good performance. Additionally, tree, rule, and Bayesian based classifiers are widely adopted in the domain of interruptibility prediction, such as Decision Tree, SVM, and Naive Bayes [77].

Based on these considerations, we built the first stage of the model by using different classifiers: Naive Bayes, Bayesian Network, SVM and Decision Tree. We evaluated the model by using a 10-fold cross validation [42]. We used 90% of the data as training data, and left 10% of the data as testing data and the results are averaged over ten runs.

Second Stage: Interruptibility Intensity Prediction

After the first stage, when the model predicts the users have reacted to the notifications, the model further predicts their interruptibility intensity.

Prediction models, including interruptibility prediction specifically, can suffer from the problem that they cannot accurately predict when there is not enough training data [61, 77]. To solve this problem, we take advantage of the personality data in this stage. Studies have shown that personality has strong connection with human behaviors, for example, personality affects the task completion time [51], preferences and interests [76, 48, 17]. Further, personality traits influence the time people take to even view a notification and how disruptive notifications are perceived [55], which demonstrates the potential to consider personality in interruptibility prediction model. In our model, we utilize the data of people who share similar personality with the user and user's personal data to predict users' interruptibility intensity.

The second stage of our model is a constraint regression model. It consists of two components: the prediction from the data of people who share similar personality with the current user, and the prediction from user's personal data. The weighted combination of the above two components determines the interruptibility intensity. Equation (1) shows the model.

$$\begin{aligned} Interruptibility\_intensity = \omega_1 f(Sim\_People\_Data) \\ +\omega_2 f(Personal\_Data) \\ \text{s.t. } \omega_1+\omega_2=1, \omega_1 \geq 0, \omega_2 \geq 0 \end{aligned} \quad (1)$$

where *Sim_People_Data* is the data of the people who share similar personality traits with the current user, *Personal_Data* is the data of the current user, and ω_1 and ω_2 are the weights of predictions from people who share similar personality with the user and user's personal data. Function f refers to regression models, where interruptibility intensity (levels) is a dependent variable and contextual information (time, location, state changes, transition state, current activity, mood) and task information (type of task, whose task, task duration) are independent variables.

For function f, we evaluated four different regression models: Linear regression, Additive regression [26], M5P [64] and k-Nearest Neighbors [4]. We used the Linear regression algorithm as a naive baseline. Additive regression treats the dependent variable as the sum of unknown functions of the independent variables. M5P is a model tree learner, which means it is a decision tree where each leaf is a regression model. M5P is considered good for categorical and numeric variables as in our case. k–Nearest–Neighbor algorithm (called IBk [2] in Weka) is a non-parametric regression method. It is robust to noisy data and requires no assumption of the data, and also suitable for low-dimensional data.

To obtain the data of people with similar personality, we can extract a group of people from the data pool with the knowledge of the personality of all participants, as shown in equation (2).

$$Sim\_People = f_0(personality\_of\_cur\_user) \quad (2)$$

where *Sim_People* means the people who share similar personality with the current user, $personality\_of\_cur\_user$ is the personality of the current user, f_0 refers to the similarity measurement of the personality. In our experiment, we employed the k-nearest-neighbor algorithm, and the distance function we used is Euclidian distance based on Big-Five personality traits, as shown in equation (3).

$$d(p_i, p_j) = \sum_{k=1}^{5} |p_{(i,t_k)} - p_{(j,t_k)}|^2 \quad (3)$$

where p_i is i'th person's personality, p_j is j'th person's personality, $p_{(i,t_k)}$ is the k'th personality trait of i'th person, $p_{(j,t_k)}$ is the k'th personality trait of j'th person.

As equation (1) shows, people's interruptibility intensity can be inferred from their own interruptibility history and the history of people who have similar personality to them. We need to guarantee that both ω_1 and ω_2 are non-negative, and the sum of them is 1 in that the range of the interruptibility intensity is from 1 to 5. Clearly, ω_1 is 1 and ω_2 is 0 at the beginning since we do not have any data about the current user, we can only rely on the data of the people who share similar personality with current user. When we have data of the current user, that personal data starts playing a role in the prediction. When that happens, ω_2 is set above zero. We note that both ω_1 and ω_2 are not static or preset, they are trained from the data.

RESULTS: STATISTICAL INFERENCE

In this section, we use Bayesian data analysis towards understanding how context factors and task related factors affect interruptibility. We consider participant's interruptibility ratings as an ordinal predicted variable, and use a Bayesian approach to model it with an underlying continuous variable [44]. We assume normal distribution for the underlying continuous value thus the interruptibility ratings are generated by the thresholded cumulative-normal model. For all the tests, we use 95% highest density interval (HDI), set the limit of the region of

practical equivalence (ROPE) on difference of means as (−0.2, 0.2), the limit of the ROPE on effect size as (−0.1, 0.1). These limits and settings are conventionally used in Bayesian data analysis [44]. We treat the categorical variable (location, activity, relations) as a nominal variable, and treat the continuous variable (mood, duration) as a numeric variable.

The more pleasant, the more interruptible: We model the mean (μ) of the underlying continuous variable as a linear regression of the mood scores:

$$\mu = \beta_0 + \beta_1 \times Mood$$

Marginal posterior distribution on β_1 (slope of the linear model) shows that the mode of β_1 is −0.065 with 95% HDI from −0.0755 to −0.0553. This indicates that the mean interruptibility rating decreases when mood score increases. In other words, participants were more interruptible when they were more pleasant (interruptibility rating 1 means highly interruptible, 5 means highly uninterruptible. The higher the mood score, the more pleasant the person is).

Interruptibility Differs at Different Places: We extracted five place categories that were commonly visited by most participants as: home, work, entertainment, transportation and shopping places. We examined the posterior distribution of underlying variable means at different places and the credible differences between means at different places.

The posterior distribution of underlying means at different places shows that participants did not want to be interrupted at shopping places with mode = 4.41. The posterior distribution on differences in the underlying means between shopping and entertainment place shows that the mean difference has a 95% HDI (−1.29, −0.226) excluding zero, and excluding the ROPE from −0.2 to 0.2 with mode = −0.708. The posterior distribution on the effect size of the mean differences has a 95% HDI (−0.621, −0.115) excluding a ROPE from −0.1 to 0.1, with mode = −0.369. This indicates that participants were less interruptible at shopping places than at entertainment places. Similarly, we found that participants are less interruptible at work places than at entertainment places, with 95% HDI of mean difference from −0.746 to −0.232, with mode = −0.475, and 95% HDI of effect size of mean differences from −0.432 to −0.13, with mode = −0.277.

We also found a few participants visited healthcare and medical facilities, and they were highly interruptible at such places (interruptibility mean = 1.59).

Interruptibility Differs with Different Activities: Table 1 lists all current activity types. The posterior distribution of underlying means for different activities shows that participants were most uninterruptible with mode = 4.36 when participants were studying, they were most interruptible when they were using the phone, with mode = 2.88. The posterior distribution on differences in the underlying means between exercising and other activities (talking, on the phone, gaming, watch TV/video, email, and bored) shows that the difference excludes 0.0 with 95% HDI from −2.11 to −0.26 with mode = −1.04, which completely excludes the ROPE from −0.2 to 0.2. The posterior distribution on effect size of mean differences has a 95% HDI (−0.872, −0.128) with mode = −0.506, excluding the ROPE from −0.1 to 0.1. Similarly, participants were less interruptible when they were they were studying than on the phone, with 95% HDI of mean difference from 1.22 to 1.75 with mode = 1.49, and 95% HDI of effect size of mean differences from 0.777 to 1.12 with mode = 0.936.

Personal Relations Influence Interruptibility: The posterior distribution on differences in the underlying means between different relations shows that the mean difference between stranger and other relations has a 95% HDI from 1.33 to 3.64 with mode = 2.28, which completely excludes the ROPE (−0.2, 0.2). The posterior distribution on effect size of mean differences also completely excludes the ROPE (−0.1 to 0.1) with 95% HDI from 0.394 to 0.823, mode = 0.59. This indicates that participants were more interruptible when interrupted by people they know or close to than strangers.

Short Interruptions Make People More Interruptible: We model the mean (μ) of the underlying continuous variable as a linear regression of the interruption duration:

$$\mu = \beta_0 + \beta_1 \times Duration$$

Marginal posterior distribution on β_1 (slope of the linear model) shows that the mode of β_1 is 0.0313 with 95% HDI from 0.0295 to 0.0327. This indicates that the mean interruptibility rating increases when interruption duration increases. In other words, participants became less interruptible when the interruption took longer time. (Interruptibility rating 1 means highly interruptible, 5 means highly uninterruptible).

RESULTS: PREDICTION EVALUATION

In this section, we present the evaluation results of our two-stage hierarchical model for interruptibility prediction.

First Stage: Predicting Reaction to Interruption

In this stage, we predict whether the participants react to the survey prompts. We used Naive Bayes, Bayesian Net, SVM and Decision Tree for performance comparisons.

Table 2 shows that SVM and Decision Tree outperform the other three classifiers; they can achieve prediction accuracy of 75.0%. SVM achieves better recall (76%) and Decision Tree achieves better precision (78%) and F-measure (62%). Accuracy is the ratio of correct predictions to all the predictions. Due to accuracy paradox [80], accuracy alone usually is not enough to measure the performance of a classifier. For example, a classifier with 95% accuracy is not useful if 95% of notifications are not answered and the 5% that are answered are misclassified. Therefore, we also reported precision, recall and F-measure. Precision is the ratio that true positive predictions to all the predicted positive predictions, it measures the exactness of the classifier. Recall, also called sensitivity, is the ratio of number of true positive predictions to the number of all positive class values in the data. It measures the completeness of the classifier. F-measure is the weighted average of the precision and recall, which measures the balance between the precision and recall.

Table 2 shows that the prediction result is improved on average over 10 percentage points with all of the important metrics

Classifier	Accuracy	Precision	Recall	F-Measure
Naive Bayes	0.66* 0.58	0.66* 0.59	0.69* 0.64	0.68* 0.61
Bayesian Net	0.72* 0.58	0.73* 0.59	0.73* 0.64	0.73* 0.62
SVM	0.75* 0.60	0.75* 0.61	0.78* 0.59	0.76* 0.60
Decision Tree	0.75* 0.60	0.76* 0.60	0.76* 0.60	0.76* 0.56
Baseline	0.56* 0.56	0.31* 0.31	0.56* 0.56	0.40* 0.40

Table 2. First stage prediction results of different classifiers with (* marked) and without personality traits. Both SVM and Decision Tree with personality traits can achieve 75% accuracy. The recall of SVM is slightly higher than Decision Tree, and the precision of Decision tree is better than SVM. The prediction results dropped largely when personality traits were not included. On average, all the important metrics of the classifiers dropped over 10 percentage points when we did not include personality. Baseline is a naive classifier that simply predicts the majority class in the dataset without considering the features.

Classifier	NB	BN	SVM	DT
FPR	0.342	0.28	0.249	0.248
FNR	0.336	0.272	0.249	0.241

Table 3. False positive rate (FPR) and false negative rate (FNR) of different classifiers (NB: Naive Bayes, BN: Bayesian Net, DT: Decision Tree) for reaction prediction. DT and SVM have roughly the lowest FPR and FNR among the tested classifiers. Given the high accuracy, recall, precision and F-measure of Decision Tree and SVM, they are the best classifiers to predict the first stage reaction to a notification.

by including personality traits as features. To find the effect of personality traits on the prediction, we re-evaluated all the classifiers by removing the personality traits from the data. On average, the prediction accuracy, precision, recall and F-measure dropped by 13.8, 15.8, 11.2 and 12.6 percentage points. Personality indeed affects users behavior and knowing it can assist on how and when to interrupt users.

Table 2 shows that classifiers without personality traits perform only slightly better than a baseline classifier that simply predicts the majority class in the dataset without considering the features. In our case, 55.6% survey notifications were answered and 44.4% were not. The answered survey is the majority class. Thus, the prediction accuracy of the baseline classifier is 55.6%. If we do not consider personality traits, the accuracies of the tested classifiers are around 60%.

Table 3 shows the false positive and false negative rates of the tested classifiers. Decision Tree and SVM have roughly the lowest rates and they both had high prediction accuracy, precision, recall and F-measure.

Second Stage: Predicting Interruptibility Intensity

In this stage, our model predicts users' interruptibility intensity if users reacted to the interruptions. The interruptibility intensity is based on users's self-reports, it ranges from 1 (highly interruptible) to 5 (highly uninterruptible).

Figure 3 shows the results for predicting interruptibility intensity and that the Additive regression performs the best with an average prediction accuracy of 67.2%. We trained the model

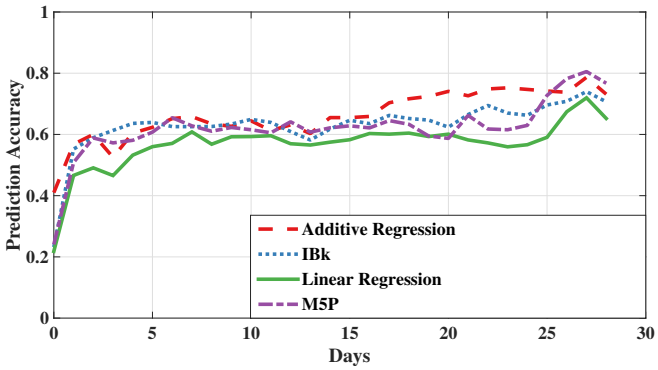

Figure 3. The prediction accuracy of model by using different regression algorithms and number of nearest neighbors k = 5. The prediction accuracy of different regression algorithms is increasing along with time. Our model using Additive regression performs the best in the initial stage, the accuracy of day 0 is 41.0%. On the average, Additive regression also achieves the best accuracy of 67.2%. The model using Linear regression performs the worst, as we cannot find a linear relation between the interruptibility and the independent variables.

by using the first N days' data of a user and the data of people who have similar personality with the user. Then we tested it by using the data of the user since day N+1. For example, the prediction in day 0 is for the initial stage when there is no data collected from the current user yet. At day 0, Additive regression algorithm achieves the best prediction accuracy. After about 16 days, the average accuracy tends to stabilize around 75%. Linear regression performs the worse across the whole period. IBk and M5P algorithms perform similarly, the average accuracy is around 60% for the first 25 days, and increases to 70% after that.

After comparing the performance of our model with all possible combinations of different k values (number of nearest neighbors) and different regression algorithms, we found that our model performs best when using Additive regression algorithm with k value of 5. Also, our model has the best prediction accuracy, 41.0%, for the initial prediction (day 0) when k = 5.

Figure 3 shows that the prediction accuracy of our model increases rapidly once we receive data from users. For example, the prediction accuracy can reach 56.7% after one day.

Figure 4 shows how the weight of prediction from current user (ω_2) and the weight of prediction from people who have similar personality (ω_1) change over time. Roughly on the fourth day, ω_2 outweighs ω_1. The prediction from user's own data is more important than the prediction from people have similar personality with the current user since day 5. After about 20 days, ω_1 and ω_2 stabilize around 0.25 and 0.75.

Figure 5 shows that recall, precision and F-measure are increasing along with the data collection when using Additive regression algorithm and k = 5. After roughly 21 days, the recall, precision and F-measure of the model stabilize around 65%. The second stage of the model can achieve relatively exact and complete predictions. When we tested if the prediction was correct or not, we translated the problem to classification. Therefore, we reported these major measures (accuracy, precision, recall and F-measure) of a classifier.

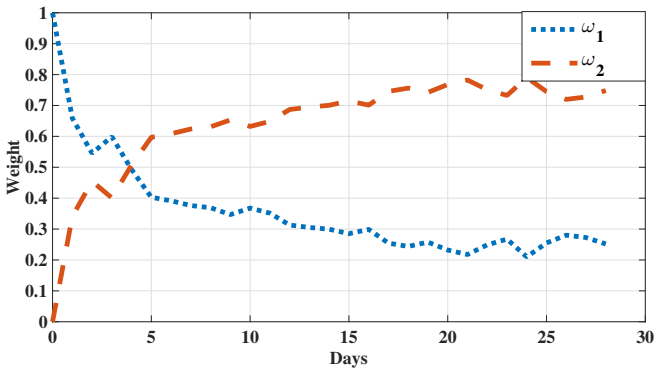

Figure 4. The weight of prediction from current user (ω_2) and the weight of prediction from people having similar personality with current user (ω_1) change over time when using Additive regression and k = 5. After about 4 days, ω_2 outweighs ω_1. That is the prediction from current user's own data is more important than the prediction from similar people since day 5. After about 20 days, ω_1 and ω_2 stabilize around 0.25 and 0.75. The prediction from user's own data contributes about 75% to the final intensity prediction, while the prediction from similar people contributes 25% to the final prediction.

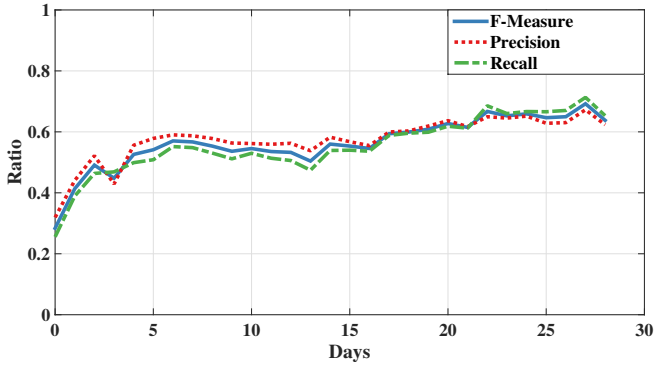

Figure 5. The changes of recall, precision and F-measure of second stage prediction over time when using Additive regression and k = 5. Recall, precision and F-measure are gradually increasing along with data collection. After about 20 days, all three metrics tend to be stable at 65%.

Root mean square error (RMSE) is one of the important metrics to measure regression models. It measures how close the predicted values to the observed values. Lower RMSE means better prediction. As our second stage is a regression model, we also use RMSE to evaluate how accurately the second stage can predict the interruptibility intensity.

Figure 6 shows that root mean square errors (RMSE) of the second stage of our model decrease over the time when using Additive regression algorithm and k = 5. At the initial stage when there is no data of the user, both RMSEs of training data and testing data are around 0.9. When we have more data of the user, RMSEs are gradually decreasing. That means our model fits the data better and predicts the interruptibility more accurately when we have more data of the user. After about 20 days, RMSEs of training data and testing data slowing approaches to 0.4 and 0.5. That shows that on average the predicted interruptibility intensity is less than one level apart from the actual level.

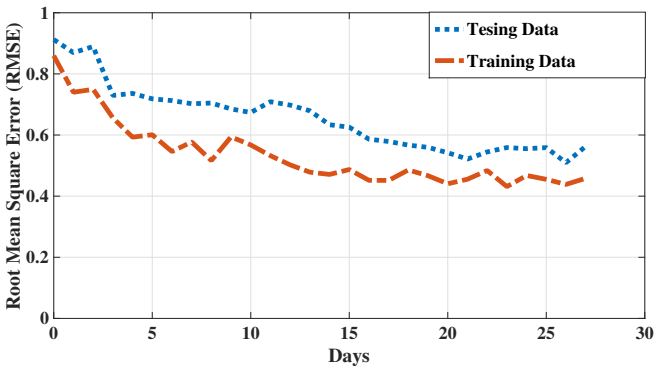

Figure 6. Root mean square error (RMSE) changes over time when using Additive regression algorithm and k = 5. Along with the data collection, RMSEs of training data and testing data are gradually deceasing. When having more data, the model can fit better to the training data and also make better prediction. After about 20 days, RMSEs of training data and testing data tend to stabilize around 0.4 and 0.5.

Overall Prediction Accuracy of the Model

The overall prediction accuracy is the ratio of correct predictions to all the predictions. Our model is a two-stage hierarchical prediction model. To calculate the overall prediction accuracy, we need to consider these two stages together:

$$\begin{aligned} &Overall\_accuracy = p(corr\_pred) \\ &= p(corr\_pred, reacted) + p(corr\_pred, not\_reacted) \\ &= p(corr\_pred|reacted)p(reacted) \\ &+ p(corr\_pre|not\_reacted)p(not\_reacted) \end{aligned} \quad (4)$$

Where *corr_pred* means correct predictions, *reacted* and *not_reacted* mean participant reacted to the interruption and did not react to the interruptions.

Equation 4 shows that the overall correct prediction consists of two parts: the correct prediction that participants did not react to notifications and the correct prediction when participants reacted to the notifications.

Our model achieves prediction accuracy of 75% in the first stage, and can achieve prediction accuracy of 78.7%, with an average of 66.2% in second stage. According to equation 4, the overall accuracy of our model can reach 66.1%, and on average it is 60.9%.

DISCUSSION

A traditional binary classification of users' interruptibility cannot predict the extent of users' availability and busyness. We solved this problem by proposing a two-stage hierarchical model, and our model can accurately predict users' interruptibility intensity. This is very useful for various applications. We introduced and took advantage of personality traits. We discuss the major results and findings below.

We found that utilizing the Big Five personality traits significantly improved the first stage prediction. On average, all the major measures (accuracy, precision, recall and F-measure) of the tested classifiers increased 10 percentage points when the personality traits were included as features. We believe this is because people who have similar personality traits can behave

similarly under similar context. For example, people who are more extroverted can be more likely to react to notifications.

One major implication of our finding of the personality traits is that we could use it to prime the second stage prediction. In the second stage, when we have no data of a user in the beginning, our model only uses the data of similar people to predict the interruptibility intensity. The result shows that in this case our model can have a prediction with accuracy over 40%. This is a significant result. When we have more data of the user, the prediction relies more on the user's personal data. After day four, the weight of the prediction from personal data outweighs the weight of prediction from similar people. After roughly 20 days, the prediction weight of personal data stabilizes around 0.75, while the prediction weight of data of similar people stabilizes around 0.25.

Based on how the weights change, the prediction power of people sharing similar personality traits is reduced over time. However, they still have some significance in the final prediction. This could be because when users receive new notifications that never seen before or change their behavioral pattern, the model cannot make an accurate prediction only based on the users' own data. In this case, the data used for prediction from the people who have similar personality traits would take effect. Indeed, people usually receive more notifications from already installed apps or messages from known people, and receive less notifications from new apps or strangers.

Personality traits can also be widely used in various applications. Personality traits can be obtained by asking users to take a short personality test after they install apps. With the acquisition of people's personality traits, systems can build a generic model for people sharing similar personality. For new users, we can use the model built on people sharing similar personality with them to make predictions.

Our results show that when people are in a pleasant mood, they are likely to be more interruptible than in a unpleasant mood. Currently smartphones cannot directly infer users mood, however, it may be possible to infer this from how they are interacting with their smartphones (e.g. finger stroke [68], motion gestures [11]) or from other sensors [57]. This presents an interesting avenue for further research.

We found that users' interruptibility varied among physical places. To our surprise, shopping places were found as the most uninterruptible place. A few participants were found highly interruptible at locations such as healthcare and medical facilities. This may be due to people waiting to see doctors. However, only few of our participants visited healthcare related places. In addition, we found that the people were less interruptible when at shopping and at work in contrast to places associated to entertainment.

We found that the relation between interrupters and interruptees plays an important role in estimating interruptibility. People were more interruptible when they would be interrupted by immediate family members or other people they know well and see frequently. Conversely, participants were reluctant to be interrupted by people they were not familiar with. This finding complements that disruption perception varies with senders of notifications [55]. Also, it complements previous findings that users are unlikely to click the notifications from distant senders [54] and notifications from immediate family members are more acceptable [19]. Participants were more interruptible when they were interrupted by shorter tasks than longer tasks.

We found that people's interruptibility differed when they were involving in different activities. For example, we found that participants were reluctant to be interrupted when they were studying. Compared to other activities, they were less interruptible when they were exercising. This dovetails well with previous findings that ongoing task type is associated with the perceived disruption [55].

LIMITATIONS

One limitation of EMA is that participants respond and self-report their availability and type of contents they are interested in. Participants can only see the interruption content when opening the app. Thus, our model may miss the contextual information embodied in the notifications that could be used in the first stage. Our model uses particular content (a survey) and may not necessarily generalize to other types of content. However, our results have shown the effectiveness of our two-stage model. Our approach could be applied to actual notifications with all types of different content.

Another limitation is that we classified all the non-responded EMA prompts as uninterruptible in the first stage. When users did not answer the survey, we cannot separate the situations whether they were available but ignored them intentionally or they were completely uninterruptible. One possible improvement for further studies is to make the first stage a ternary classification (unavailable, available but do not want to answer the survey, and completely available).

CONCLUSIONS

We developed a novel hierarchical model to predict interruptibility and its intensity. Our work highlights the importance of predicting interruptibility in different levels. Our model can achieve an accuracy of 66.1% (60.9% on average) for predicting interruptibility intensity, and 75% for first-stage binary interruptibility classification. In addition, we have presented the importance of personality traits in predicting people's busyness. Our work is the first to employ personality traits with predicting interruptibility. Our approach solves the important problem of initial prediction when the individual prediction model has not yet trained on user's data. This approach can be applied to various applications and platforms.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant Numbers 1211079 and 1546689. Xianyi Gao was supported by the National Science Foundation Graduate Research Fellowship Program under Grant Number 1433187. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

1. Piotr D. Adamczyk and Brian P. Bailey. 2004. If Not Now, when?: The Effects of Interruption at Different Moments Within Task Execution. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 271–278. DOI:http://dx.doi.org/10.1145/985692.985727
2. David W. Aha, Dennis Kibler, and Marc K. Albert. 1991. Instance-Based Learning Algorithms. *Mach. Learn.* 6, 1 (Jan. 1991), 37–66. DOI: http://dx.doi.org/10.1023/A:1022689900470
3. Airtasker. 2017. Airtasker Application. (2017). https://www.airtasker.com.
4. Naomi S Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* 46, 3 (1992), 175–185.
5. Brian P Bailey and Joseph A Konstan. 2006. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in human behavior* 22, 4 (2006), 685–708.
6. James "Bo" Begole, Nicholas E. Matsakis, and John C. Tang. 2004. Lilsys: Sensing Unavailability. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work (CSCW '04)*. ACM, New York, NY, USA, 511–514. DOI: http://dx.doi.org/10.1145/1031607.1031691
7. Frank Bentley. 2013. Investigating the Place Categories Where Location-Based Services Are Used. (2013).
8. Matthias Böhmer, Christian Lander, Sven Gehring, Duncan P. Brumby, and Antonio Krüger. 2014. Interrupted by a Phone Call: Exploring Designs for Lowering the Impact of Call Notifications for Smartphone Users. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3045–3054. DOI: http://dx.doi.org/10.1145/2556288.2557066
9. Chegg. 2017. Chegg Application. (2017). http://www.chegg.com.
10. Richard Cooper and Bradley Franks. 1993. Interruptibility as a constraint on hybrid systems. *Minds and Machines* 3, 1 (1993), 73–96.
11. Céline Coutrix and Nadine Mandran. 2012. Identifying Emotions Expressed by Mobile Users Through 2D Surface and 3D Motion Gestures. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp '12)*. ACM, New York, NY, USA, 311–320. DOI:http://dx.doi.org/10.1145/2370216.2370265
12. Boele De Raad. 2000. *The Big Five Personality Factors: The psycholexical approach to personality.* Hogrefe & Huber Publishers.
13. Philippe AEG Delespaul. 1995. *Assessing schizophrenia in daily life: The experience sampling method*. Ph.D. Dissertation. Maastricht university.
14. Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 2013. Pick-a-crowd: Tell Me What You Like, and I'll Tell You What to Do. In *Proceedings of the 22nd International Conference on World Wide Web (WWW '13)*. ACM, New York, NY, USA, 367–374. DOI: http://dx.doi.org/10.1145/2488388.2488421
15. Pedro Domingos and Michael Pazzani. 1997. On the Optimality of the Simple Bayesian Classifier Under Zero-One Loss. *Mach. Learn.* 29, 2-3 (Nov. 1997), 103–130. DOI: http://dx.doi.org/10.1023/A:1007413511361
16. M. Brent Donnellan, Frederick L Oswald, Brendan M Baird, and Richard E Lucas. 2006. The mini-IPIP scales: tiny-yet-effective measures of the Big Five factors of personality. *Psychological assessment* 18, 2 (2006), 192.
17. Greg Dunn, Jurgen Wiersema, Jaap Ham, and Lora Aroyo. 2009. Evaluating Interface Variants on Personality Acquisition for Recommender Systems. In *Proceedings of the 17th International Conference on User Modeling, Adaptation, and Personalization: Formerly UM and AH (UMAP '09)*. Springer-Verlag, Berlin, Heidelberg, 259–270. DOI: http://dx.doi.org/10.1007/978-3-642-02247-0_25
18. Joel E. Fischer, Chris Greenhalgh, and Steve Benford. 2011. Investigating Episodes of Mobile Phone Activity As Indicators of Opportune Moments to Deliver Notifications. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '11)*. ACM, New York, NY, USA, 181–190. DOI: http://dx.doi.org/10.1145/2037373.2037402
19. Joel E. Fischer, Nick Yee, Victoria Bellotti, Nathan Good, Steve Benford, and Chris Greenhalgh. 2010. Effects of Content and Time of Delivery on Receptivity to Mobile Interruptions. In *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '10)*. ACM, New York, NY, USA, 103–112. DOI: http://dx.doi.org/10.1145/1851600.1851620
20. Robert Fisher and Reid Simmons. 2011. Smartphone Interruptibility Using Density-Weighted Uncertainty Sampling with Reinforcement Learning. In *Proceedings of the 2011 10th International Conference on Machine Learning and Applications and Workshops - Volume 01 (ICMLA '11)*. IEEE Computer Society, Washington, DC, USA, 436–441. DOI: http://dx.doi.org/10.1109/ICMLA.2011.128
21. James Fogarty and Scott E. Hudson. 2007. Toolkit Support for Developing and Deploying Sensor-based Statistical Models of Human Situations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, NY, USA, 135–144. DOI: http://dx.doi.org/10.1145/1240624.1240645

22. James Fogarty, Scott E. Hudson, Christopher G. Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny C. Lee, and Jie Yang. 2005a. Predicting Human Interruptibility with Sensors. *ACM Transactions on Computer-Human Interaction (TOCHI)* 12, 1 (March 2005), 119–146. DOI: http://dx.doi.org/10.1145/1057237.1057243

23. James Fogarty, Scott E. Hudson, and Jennifer Lai. 2004. Examining the Robustness of Sensor-based Statistical Models of Human Interruptibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 207–214. DOI:http://dx.doi.org/10.1145/985692.985719

24. James Fogarty, Andrew J. Ko, Htet Htet Aung, Elspeth Golden, Karen P. Tang, and Scott E. Hudson. 2005b. Examining Task Engagement in Sensor-based Statistical Models of Human Interruptibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05)*. ACM, New York, NY, USA, 331–340. DOI:http://dx.doi.org/10.1145/1054972.1055018

25. Foursquare. 2015. Foursquare API. (2015). https://developer.foursquare.com/.

26. Jerome H Friedman. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38, 4 (2002), 367–378.

27. Google. 2015. Google Activity Recognition API. (2015). https://developer.android.com/reference/com/google/android/gms/location/ActivityRecognitionApi.html.

28. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations Newsletter* 11, 1 (Nov. 2009), 10–18. DOI: http://dx.doi.org/10.1145/1656274.1656278

29. Rikard Harr and Victor Kaptelinin. 2012. Interrupting or Not: Exploring the Effect of Social Context on Interrupters' Decision Making. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design (NordiCHI '12)*. ACM, New York, NY, USA, 707–710. DOI: http://dx.doi.org/10.1145/2399016.2399124

30. Joel M Hektner, Jennifer A Schmidt, and Mihaly Csikszentmihalyi. 2007. *Experience sampling method: Measuring the quality of everyday life*. Sage.

31. Ken Hinckley and Eric Horvitz. 2001. Toward More Sensitive Mobile Phones. In *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology (UIST '01)*. ACM, New York, NY, USA, 191–192. DOI:http://dx.doi.org/10.1145/502348.502382

32. Joyce Ho and Stephen S. Intille. 2005. Using Context-aware Computing to Reduce the Perceived Burden of Interruptions from Mobile Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05)*. ACM, New York, NY, USA, 909–918. DOI: http://dx.doi.org/10.1145/1054972.1055100

33. Eric Horvitz and Johnson Apacible. 2003. Learning and Reasoning About Interruption. In *Proceedings of the 5th International Conference on Multimodal Interfaces (ICMI '03)*. ACM, New York, NY, USA, 20–27. DOI: http://dx.doi.org/10.1145/958432.958440

34. Eric Horvitz, Johnson Apacible, and Muru Subramani. 2005. Balancing Awareness and Interruption: Investigation of Notification Deferral Policies. In *Proceedings of the 10th International Conference on User Modeling (UM'05)*. Springer-Verlag, Berlin, Heidelberg, 433–437. DOI:http://dx.doi.org/10.1007/11527886_59

35. Eric Horvitz, Paul Koch, and Johnson Apacible. 2004. BusyBody: Creating and Fielding Personalized Models of the Cost of Interruption. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work (CSCW '04)*. ACM, New York, NY, USA, 507–510. DOI:http://dx.doi.org/10.1145/1031607.1031690

36. Scott Hudson, James Fogarty, Christopher Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny Lee, and Jie Yang. 2003. Predicting Human Interruptibility with Sensors: A Wizard of Oz Feasibility Study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*. ACM, New York, NY, USA, 257–264. DOI: http://dx.doi.org/10.1145/642611.642657

37. Shamsi T. Iqbal and Brian P. Bailey. 2006. Leveraging Characteristics of Task Structure to Predict the Cost of Interruption. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, New York, NY, USA, 741–750. DOI: http://dx.doi.org/10.1145/1124772.1124882

38. Shamsi T. Iqbal and Brian P. Bailey. 2008. Effects of Intelligent Notification Management on Users and Their Tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 93–102. DOI: http://dx.doi.org/10.1145/1357054.1357070

39. Shamsi T. Iqbal and Eric Horvitz. 2010. Notifications and Awareness: A Field Study of Alert Usage and Preferences. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW '10)*. ACM, New York, NY, USA, 27–30. DOI: http://dx.doi.org/10.1145/1718918.1718926

40. Nicky Kern, Stavros Antifakos, Bernt Schiele, and Adrian Schwaninger. 2004. A Model for Human Interruptability: Experimental Evaluation and Automatic Estimation from Wearable Sensors. In *Proceedings of the Eighth International Symposium on Wearable Computers (ISWC '04)*. IEEE Computer Society, Washington, DC, USA, 158–165. DOI:http://dx.doi.org/10.1109/ISWC.2004.3

41. SeungJun Kim, Jaemin Chun, and Anind K. Dey. 2015. Sensors Know When to Interrupt You in the Car: Detecting Driver Interruptibility Through Monitoring of Peripheral Interactions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in*

Computing Systems (CHI '15). ACM, New York, NY, USA, 487–496. DOI: http://dx.doi.org/10.1145/2702123.2702409

42. Ron Kohavi. 1995. A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2 (IJCAI'95)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1137–1143. http://dl.acm.org/citation.cfm?id=1643031.1643047
43. Vlaho Kostov, Takashi Tajima, Eiichi Naito, and Jun Ozawa. 2006. Analysis of appropriate timing for information notification based on indoor user's location transition. In *Pervasive Computing and Communications, 2006. PerCom 2006. Fourth Annual IEEE International Conference on*. IEEE, 6–pp. DOI: http://dx.doi.org/10.1109/PERCOM.2006.8
44. John Kruschke. 2014. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.
45. Neal Lathia, Kiran K. Rachuri, Cecilia Mascolo, and Peter J. Rentfrow. 2013. Contextual Dissonance: Design Bias in Sensor-based Experience Sampling Methods. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*. ACM, New York, NY, USA, 183–192. DOI:http://dx.doi.org/10.1145/2493432.2493452
46. Mary Kate Law, William Fleeson, Elizabeth Mayfield Arnold, and R Michael Furr. 2015. Using negative emotions to trace the experience of borderline personality pathology: Interconnected relationships revealed in an experience sampling study. *Journal of Personality Disorders* (2015), 1–19.
47. Kyungmin Lee, Jason Flinn, and Brian Noble. 2015. The Case for Operating System Management of User Attention. In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications (HotMobile '15)*. ACM, New York, NY, USA, 111–116. DOI: http://dx.doi.org/10.1145/2699343.2699362
48. Cha-Hwa Lin, Dennis McLeod, and others. 2002. Exploiting and Learning Human Temperaments for Customized Information Recommendation. In *IMSA*. 218–223.
49. Janne Lindqvist, Justin Cranshaw, Jason Wiese, Jason Hong, and John Zimmerman. 2011. I'm the Mayor of My House: Examining Why People Use Foursquare - a Social-driven Location Sharing Application. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 2409–2418. DOI: http://dx.doi.org/10.1145/1978942.1979295
50. Hugo Lopez-Tovar, Andreas Charalambous, and John Dowell. 2015. Managing Smartphone Interruptions Through Adaptive Modes and Modulation of Notifications. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. ACM, New York, NY, USA, 296–299. DOI: http://dx.doi.org/10.1145/2678025.2701390
51. Gloria Mark, Daniela Gudith, and Ulrich Klocke. 2008. The Cost of Interrupted Work: More Speed and Stress. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 107–110. DOI: http://dx.doi.org/10.1145/1357054.1357072
52. Santosh Mathan, Stephen Whitlow, Michael Dorneich, Patricia Ververs, and Gene Davis. 2007. Neurophysiological estimation of interruptibility: Demonstrating feasibility in a field context. In *In Proceedings of the 4th International Conference of the Augmented Cognition Society*.
53. John D Mayer and Yvonne N Gaschke. 1988. The experience and meta-experience of mood. *Journal of personality and social psychology* 55, 1 (1988), 102.
54. Abhinav Mehrotra, Mirco Musolesi, Robert Hendley, and Veljko Pejovic. 2015. Designing Content-driven Intelligent Notification Mechanisms for Mobile Applications. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 813–824. DOI: http://dx.doi.org/10.1145/2750858.2807544
55. Abhinav Mehrotra, Veljko Pejovic, Jo Vermeulen, Robert Hendley, and Mirco Musolesi. 2016. My Phone and Me: Understanding People's Receptivity to Mobile Notifications. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1021–1032. DOI: http://dx.doi.org/10.1145/2858036.2858566
56. Martin Mühlenbrock, Oliver Brdiczka, Dave Snowdon, and Jean-Luc Meunier. 2004. Learning to Detect User Activity and Availability from a Variety of Sensor Data. In *Proceedings of the Second IEEE International Conference on Pervasive Computing and Communications (PERCOM '04)*. IEEE Computer Society, Washington, DC, USA, 13–. http://dl.acm.org/citation.cfm?id=977406.978698
57. Fatma Nasoz, Kaye Alvarez, Christine L Lisetti, and Neal Finkelstein. 2004. Emotion recognition from physiological signals using wireless sensors for presence technologies. *Cognition, Technology & Work* 6, 1 (2004), 4–14.
58. US Bureau of Labor Statistics. 2015. American Time Use Survey. (2015). http://www.bls.gov/tus/.
59. Tadashi Okoshi, Jin Nakazawa, and Hideyuki Tokuda. 2014. Attelia: Sensing User's Attention Status on Smart Phones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct)*. ACM, New York, NY, USA, 139–142. DOI: http://dx.doi.org/10.1145/2638728.2638802

60. Tadashi Okoshi, Julian Ramos, Hiroki Nozaki, Jin Nakazawa, Anind K. Dey, and Hideyuki Tokuda. 2015. Reducing Users' Perceived Mental Effort Due to Interruptive Notifications in Multi-device Mobile Environments. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 475–486. DOI: http://dx.doi.org/10.1145/2750858.2807517

61. Veljko Pejovic and Mirco Musolesi. 2014. InterruptMe: Designing Intelligent Prompting Mechanisms for Pervasive Applications. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14)*. ACM, New York, NY, USA, 897–908. DOI: http://dx.doi.org/10.1145/2632048.2632062

62. Martin Pielot, Rodrigo de Oliveira, Haewoon Kwak, and Nuria Oliver. 2014. Didn't You See My Message?: Predicting Attentiveness to Mobile Instant Messages. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3319–3328. DOI: http://dx.doi.org/10.1145/2556288.2556973

63. Benjamin Poppinga, Wilko Heuten, and Susanne Boll. 2014. Sensor-Based Identification of Opportune Moments for Triggering Notifications. *IEEE Pervasive Computing* 13, 1 (Jan. 2014), 22–29. DOI: http://dx.doi.org/10.1109/MPRV.2014.15

64. Ross J. Quinlan. 1992. Learning with Continuous Classes. In *5th Australian Joint Conference on Artificial Intelligence*. World Scientific, Singapore, 343–348.

65. Karen Renaud, Judith Ramsay, and Mario Hair. 2006. " You've got e-mail!"... shall I deal with it now? Electronic mail from the recipient's perspective. *International Journal of Human-Computer Interaction* 21, 3 (2006), 313–332.

66. Stephanie Rosenthal, Anind K. Dey, and Manuela Veloso. 2011. Using Decision-theoretic Experience Sampling to Build Personalized Mobile Phone Interruption Models. In *Proceedings of the 9th International Conference on Pervasive Computing (Pervasive'11)*. Springer-Verlag, Berlin, Heidelberg, 170–187. http://dl.acm.org/citation.cfm?id=2021975.2021991

67. Hillol Sarker, Moushumi Sharmin, Amin Ahsan Ali, Md. Mahbubur Rahman, Rummana Bari, Syed Monowar Hossain, and Santosh Kumar. 2014. Assessing the Availability of Users to Engage in Just-in-time Intervention in the Natural Environment. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14)*. ACM, New York, NY, USA, 909–920. DOI: http://dx.doi.org/10.1145/2632048.2636082

68. Sachin Shah, J Narasimha Teja, and Samit Bhattacharya. 2015. Towards affective touch interaction: predicting mobile user emotion from finger strokes. *Journal of Interaction Science* 3, 1 (2015), 1.

69. Saul Shiffman, Arthur A Stone, and Michael R Hufford. 2008. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.* 4 (2008), 1–32.

70. Jeremiah Smith and Naranker Dulay. 2014. Ringlearn: Long-term mitigation of disruptive smartphone interruptions. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*. IEEE, 27–35.

71. Jeremiah Smith, Anna Lavygina, Jiefei Ma, Alessandra Russo, and Naranker Dulay. 2014. Learning to Recognise Disruptive Smartphone Notifications. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services (MobileHCI '14)*. ACM, New York, NY, USA, 121–124. DOI: http://dx.doi.org/10.1145/2628363.2628404

72. Takahiro Tanaka and Kinya Fujita. 2011. Study of User Interruptibility Estimation Based on Focused Application Switching. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work (CSCW '11)*. ACM, New York, NY, USA, 721–724. DOI: http://dx.doi.org/10.1145/1958824.1958954

73. Justin Tang and Donald J Patterson. 2010. Twitter, sensors and UI: Robust context modeling for interruption management. In *User Modeling, Adaptation, and Personalization*. Springer, 123–134.

74. Rannie Teodoro, Pinar Ozturk, Mor Naaman, Winter Mason, and Janne Lindqvist. 2014. The Motivations and Experiences of the On-demand Mobile Workforce. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. ACM, New York, NY, USA, 236–247. DOI: http://dx.doi.org/10.1145/2531602.2531680

75. Henri ter Hofte. 2007. Xensible Interruptions from Your Mobile Phone. In *Proceedings of the 9th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '07)*. ACM, New York, NY, USA, 178–181. DOI: http://dx.doi.org/10.1145/1377999.1378003

76. Marko Tkalcic, Matevz Kunaver, Jurij Tasic, and Andrej Košir. 2009. Personality based user similarity measure for a collaborative recommender system. In *Proceedings of the 5th Workshop on Emotion in Human-Computer Interaction-Real world challenges*. 30–37.

77. Liam D. Turner, Stuart M. Allen, and Roger M. Whitaker. 2015. Interruptibility Prediction for Ubiquitous Systems: Conventions and New Directions from a Growing Field. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 801–812. DOI:http://dx.doi.org/10.1145/2750858.2807514

78. Jason Wiese, Patrick Gage Kelley, Lorrie Faith Cranor, Laura Dabbish, Jason I. Hong, and John Zimmerman. 2011. Are You Close with Me? Are You Nearby?: Investigating Social Groups, Closeness, and Willingness

to Share. In *Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp '11)*. ACM, New York, NY, USA, 197–206. DOI: http://dx.doi.org/10.1145/2030112.2030140

79. Mao Ye, Dong Shou, Wang-Chien Lee, Peifeng Yin, and Krzysztof Janowicz. 2011. On the Semantic Annotation of Places in Location-based Social Networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*. ACM, New York, NY, USA, 520–528. DOI: http://dx.doi.org/10.1145/2020408.2020491
80. Xingquan Zhu. 2007. *Knowledge Discovery and Data Mining: Challenges and Realities: Challenges and Realities*. Igi Global.
81. Manuela Züger and Thomas Fritz. 2015. Interruptibility of Software Developers and Its Prediction Using Psycho-Physiological Sensors. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 2981–2990. DOI: http://dx.doi.org/10.1145/2702123.2702593