UNDERSTANDING AND CAPTURING PEOPLE’S MOBILE APP PRIVACY PREFERENCES

THESIS PROPOSAL

Jialiu Lin

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
jialiul@cs.cmu.edu

August 2012

THESIS COMMITTEE

Norman Sadeh (co-chair)
School of Computer Science, Carnegie Mellon University

Jason I. Hong (co-chair)
Human-Computer Interaction Institute, Carnegie Mellon University

Mahadev Satyanarayanan

Computer Science Department, Carnegie Mellon University

Sunny Consolvo

Google, Inc.

ABSTRACT

A number of ongoing research efforts focus on protecting mobile users' privacy and security, using software analysis techniques or security extensions with app-specific privacy controls. These proposed extensions might overwhelm users with unnecessary and difficult to understand details. Unfortunately, there has been little work done to understand users' privacy preferences regarding mobile apps. A key question is whether it is possible to identify how apps' privacy-related behaviors impact users' privacy preferences in order to simplify the decisions users have to make without reducing their level of control over the decisions they really care about.

The proposed dissertation work aims to help answer this question. Specifically, we propose to use crowdsourcing and user-oriented machine learning techniques to capture and quantitatively model users' privacy preferences regarding mobile apps. We will perform detailed static analysis on a representative set of apps on the Android platform to understand their private resource usages. We will also use crowdsourcing to collect users' perceptions of these apps, including their expectations and levels of comfort in using these apps. The idea is to identify a relatively small number of sensitive data usage scenarios that most significantly impact users' privacy decisions when using a particular mobile app. By performing clustering, we expect to isolate different classes of mobile apps that elicit common privacy concerns and different groups of users with distinct privacy preferences. Based on these clusters, we want to see if we can identify a small number of user-understandable privacy profiles (or "personas") that can be used to simplify the privacy settings users could be exposed to.

The findings of this thesis can offer insight into improving current mobile privacy interfaces and settings. As a by-product, our resulting models and findings could also help mobile app developers estimate the user acceptance of their apps from a privacy perspective.

1 INTRODUCTION

Smartphone ownership has grown rapidly over the last few years. In 2012, global smartphone shipments are expected to reach 614 million units [27]. Nearly half of cell phone owners carry smartphone nowadays. The explosion in smartphone ownership has been accompanied by the emergence of App Stores that enable users to download a growing number of applications onto their devices. As of June 2012, the Google Play Store[1] offered more than 600,000 apps, with more than 20 billion downloads since its inception; the Apple App store offered more than 650,000 apps with more than 30 billion downloads since its launch. Mobile apps can make use of numerous capabilities of a smartphone, such as a user's current location and call logs, providing users with pertinent services and attractive features.

Inevitably, access to these capabilities opens the door to new types of security and privacy intrusions. Malware is an obvious problem [18, 36]; another serious problem is that mobile users, in general, are neither fully aware of nor have full control over how mobile apps access and transmit personal information. For example, the Pandora music app is under federal investigation for gathering location data, gender, year of birth, and unique device ID from mobile users and sharing this information with advertisers [19]. Social network applications, such as Facebook and Path, were found uploading entire contact lists onto their servers, which greatly surprised users and made them feel very uncomfortable [44, 76]. In fact, studies [38, 57, 59] have shown that users have a poor understanding of these sensitive resource usages, and existing interfaces fall short in terms of providing users with the information necessary to make informed decisions.

A number of ongoing research efforts focus on protecting mobile users' privacy and security using software analysis techniques or security extensions with app-specific privacy controls. (e.g., [15, 49, 87]). However, these proposed privacy controls were not grounded in user research. Asking users to systematically configure all these settings is unrealistic as it can overwhelm users with details they struggle to understand and may ultimately not care about. To date, not much work has been done to understand people's privacy preferences in using mobile apps and see to what extent a better understanding of these preferences could inform the design of interfaces that empower users to better manage their privacy.

The fundamental goal of this thesis is to complement existing mobile privacy research by providing important knowledge on the end-users' side. A key research question we aim to solve in this thesis is whether it is possible to simplify decisions users have to make without reducing their level of control over the decisions they really care about. Specifically, we propose to use crowdsourcing and user-oriented machine learning techniques to see whether it is possible to reduce and simplify the number of privacy decisions exposed to users without negatively impacting their sense of control. The idea is to identify a relatively small number of sensitive data usage scenarios that most significantly impact users' privacy concerns when using a particular mobile app. By performing clustering, we expect to isolate different classes of mobile

[1] Previously called "the Android Market."

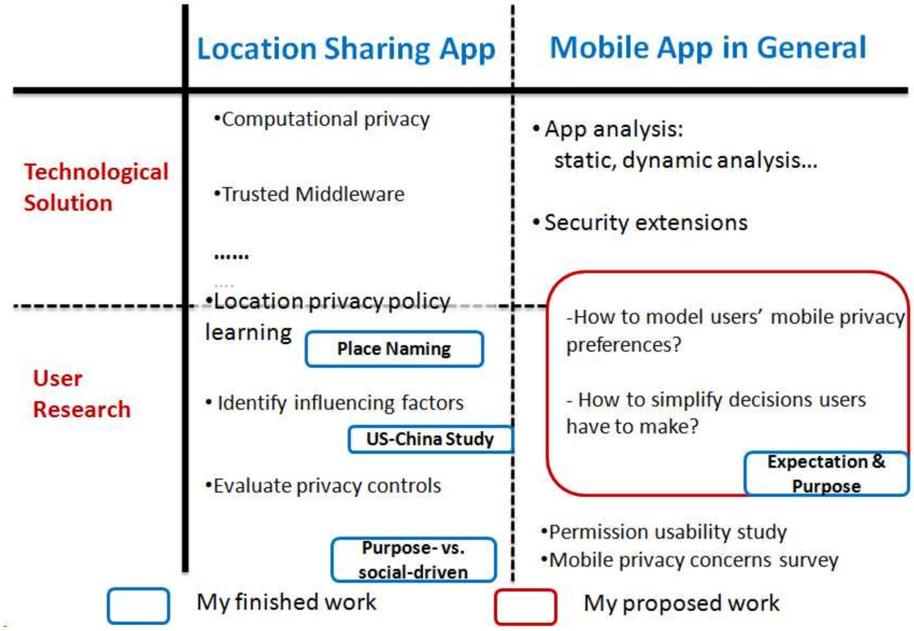

Location Sharing App
Mobile App in General
Technological Solution
User Research
•Computational privacy
•Trusted Middleware
……
•App analysis: static, dynamic analysis…
•Security extensions
•Location privacy policy learning
Place Naming
• Identify influencing factors
US-China Study
•Evaluate privacy controls
Purpose- vs. social-driven
-How to model users' mobile privacy preferences?
- How to simplify decisions users have to make?
Expectation & Purpose
•Permission usability study
•Mobile privacy concerns survey
My finished work
My proposed work

Figure 1. This figure illustrate where this thesis posit in the mobile privacy research. My proposed work complements existing mobile privacy research by providing important knowledge on the user side. My past work and the formative study are also marked in this figure, which will also be covered in this proposal.

apps and different groups of users with distinct characteristics. A small number of user-understandable privacy profiles (or "personas") will be learned based on the clusters of users with similar preferences.

Our previous work used location-sharing services as a prominent example to investigate users' privacy preferences in context sharing. We found that even when considering only this type of private resource, users' privacy preferences were complex and varied regarding a number of factors, which ranged from motivation and context to cultural influences. Yet, by leveraging machine learning techniques, we were able to identify certain patterns in users' privacy preferences and to some extent predict users' sharing behaviors under different context. This thesis will extend the discussion to general mobile apps other than context sharing services, in which users' personal information is not only shared with the members of their social network but also with app developers or 3rd parties. We foresee that this discussion will involve a significantly more complex problem space. Figure 1 shows where this thesis fits into the domain of mobile privacy research. We will conduct in-depth quantitative analysis on users' mobile app privacy preferences, especially how users' levels of comfort vary with different private resource usages.

More specially, my thesis will attempt to answer the following questions:

1. How can we capture users' mobile app privacy preferences in a scalable manner?
2. What are the key factors that affect users' privacy concerns about mobile apps?

3. To what extent can we identify meaningful collections of mobile apps that elicit similar privacy preferences among users?
4. To what extent can we identify groups of users who share similar privacy preferences for different collections of apps?

Given that there are more than half a million mobile apps in more than thirty categories, dozens of different private resources and presumably diverse privacy attitudes of users, conventional user study methods are not feasible for data collection due to their limited scalability and high management and time cost. Our initial results showed that crowdsourcing can scale up the data collection (i.e., capturing users' privacy preferences) in an efficient way [59]. Furthermore, initial results also suggested that despite the variety of sensitive data and functionality accessed by apps, users' privacy decisions are influenced by their expectations of these apps and the purposes of the resource usages. As part of this thesis, we propose to explore the validity and ramifications of these observations and leverage machine learning techniques to perform more in-depth analysis on both mobile apps' privacy-related behaviors and users' privacy preferences. We seek to determine to what extent different usage practices and groups of users could be identified to simplify the design of privacy control settings.

The central thesis aims at providing quantitative foundations and user perspectives to mobile privacy research, which can be summarized as:

> *By using crowdsourcing and user-oriented machine learning techniques, we can build accurate and understandable models of mobile apps and users' privacy preferences to inform the design of mobile privacy interfaces and settings, and to help developers build more privacy preserving apps.*

This thesis will contribute to mobile app privacy research in several ways including:

- We will compile a valuable dataset that includes attributes to describe privacy-related behaviors of mobile apps as well as how users feel about them though automated static analysis and crowdsourcing. This dataset will be an important foundation for all quantitative analyses in my thesis. It can also be used for other research purposes.
- We will develop detailed regression models of users' privacy preferences along with identified features that most critically impact users' comfort in using these apps.
- We will cluster apps based on their sensitive resource usage patterns and the resulting level of comfort expressed by users, and will present the predictive model generalized from these app clusters that could be used to estimate user acceptance of other apps.
- We will identify a set of default privacy settings by clustering users based on their preferences of different clusters of apps. Users can choose from these default settings when configuring their mobile app privacy settings. These default settings can greatly reduce user burden compared to other privacy settings that require users to specify their preferences for individual apps.

Collectively, these contributions should provide a scientific basis for starting to reconcile mobile privacy and usability and, in particular, helping inform the design of more usable privacy settings.

The remainder of this thesis proposal is organized as follows. In the next section, I will summarize my previous work in location sharing as an initial exploration of users' privacy preferences with a focus on how it links to the proposed thesis. Section 3 reviews current mobile app privacy research and how our work differs from it. Section 4 provides the preliminary results obtained from a formative study. Section 5 details the proposed work in terms of the steps involved to conduct data collection and analysis in investigating users' mobile app privacy preferences. In the remaining sections, I clarify the scope of this thesis and present a proposed schedule for completing the thesis work.

2 EXPLORING USERS' PRIVACY PREFERENCES IN LOCATION SHARING

Our initial exploration in users' mobile privacy preferences started with location sharing, focusing on how to understand and resolve users' privacy concerns in using location sharing applications (LSAs). These types of applications facilitate and encourage users to convey their location information to others in users' communications, which have recently attracted interest from both industry and academia [2-8, 17, 50, 51, 71, 84]. With the proliferation of smartphone ownership, most location-sharing services are available on mobile platforms (e.g., Google Latitude [5], Foursquare [4], Facebook Places [3]). As a special subset of mobile apps, where the users' location information is majorly consumed by people in their social networks,[2] studying the privacy issues in LSAs could provide important lessons from both methodological perspective and knowledge perspective.

Some of my past work falls into this line of research [60, 61, 75]. Our findings indicated that even only considering one type of sensitive resource, users' privacy preferences could be very complex and influenced by different factors. Our past work in location privacy provided us a sound foundation to extend the discussion to mobile apps in general, helping us proceed to a presumably more complex area. In this section, I briefly discuss three studies I conducted in this domain, and show how this line of research links to the current thesis.

2.1 Modeling People's Place Naming Preferences in Location Sharing [61]

In this work, we explored how users modulate their location information to cope with privacy concerns by analyzing the place names they used to convey location within a location sharing system. Specifically, we wanted to identify the general patterns of users' location naming preferences in different contexts and determine to what extent preferences were predictable.

[2] Though some location-sharing mobile apps also transmit users' location information to ad networks for advertising purposes.

To achieve this goal, we conducted a user study with 26 participants and captured their location traces and per-location sharing preferences by using the Day Reconstruction Method (DRM).

Based on the data we collected, we proposed a taxonomy based on the underlying information users want to convey to organize the place labels that we collected in a two-week user study in two cities. We observed that participants generally used two major techniques to tailor their location information. The first was to choose the perspective from which people address the places (i.e., semantic, e.g., "work"; geographic, e.g., "5000 Forbes ave."; or hybrid, e.g., "Starbucks on Craig st"). The second was to tune the granularity of the disclosure (i.e., the precision of a disclosure, from address level, e.g. "417 S. Craig st", to state level, e.g., "PA"). On average, we saw *2.78* place names per physical location (SD = 0.89, Med = 3, max = 7, min = 1). Participants considered multiple factors when they decided on what information to disclose, including their relationship with the recipients, their perceived levels of comfort in sharing specific locations, the recipients' familiarity with the places, and place entropy.[3] The latter two factors had not been examined previously.

Given the proposed taxonomy and these identified influencing factors, we demonstrated the feasibility of applying machine learning techniques to predict the way people manipulate their location information in various situations. In our experiments, we leveraged the J48 decision tree algorithm to predict the types of place naming methods people will use in different situations and validated prediction accuracies through a fivefold cross validation. The prediction of the top-level class (semantic, geographic, hybrid) yielded an average accuracy of 85.5%; granularity prediction yielded an average accuracy of 71.25%.

Along a similar direction, in collaborative work with Tang [75], we used the same taxonomy to analyze the location labels users selected in different scenarios and reframed the location-sharing applications (LSA) into two categories, based on the users' intention of sharing, namely purpose-driven LSAs and social-driven LSAs. Our findings indicated that people have distinct sharing preferences in using these two categories of LSAs in terms of (1) the types of location information they chose to share, (2) the different privacy concerns people had and strategies used to cope with these concerns, and (3) how privacy-preserving these location disclosures were. These results suggested that LSAs should consider which type of location sharing they primarily support in order to select the most appropriate data type and visualizations. Another important finding concerns the factors involved in users' location sharing decisions. In socially-driven sharing, users attempt a balance between maximizing their social capital and protecting their own privacy. Social-driven LSAs can leverage this information by playing to these factors to encourage users to share their location.

2.2 A Comparative Study of Location-Sharing Privacy Preferences in U.S. and China [60]

[3] Place entropy characterizes the diversity of users seen in a particular place. See [25]J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh, "Bridging the gap between physical location and online social networks," In Proc. *UbiComp*, 2010.

While prior studies had provided us with an initial understanding of people’s location-sharing privacy preferences, they had been limited to Western countries and did not investigate the impact of the granularity of location disclosures on people’s privacy preferences. In this study, we reported findings of a three-week comparative study collecting location traces and location-sharing preferences from two groups of university students in the U.S. and China with similar demographics. More specifically, we asked participants to specify their per-location privacy preferences in two conditions with four recipient types (i.e., close friends and family, friends on Social Networking Sites, university community, and advertisers). In the all-or-none condition, participants had to share an exact location or disclose no location at all. In the granularity condition, participants could choose to manipulate the level of granularity at which their location was disclosed.

On average, Chinese participants were more conservative about sharing their location, especially when sharing with their close friends and family and their friends on Social Networking Sites or when they were at work, compared to their U.S. counterparts. We also noticed that Chinese participants' privacy preferences were more differentiated during working and nonworking hours.

Another major finding was that these two groups of participants used granularity control differently. On average, U.S. participants shared less detailed location than their Chinese counterparts when the granularity control was given. The two groups used the granularity control with different intentions. Chinese participants leveraged the granularity control to open up more sharing opportunities where they refused to share in the all-or-none condition. However, U.S. participants harnessed granularity control to further limit location details in sharing. This finding suggests that, in the absence of granularity settings, U.S. participants are more willing than Chinese participants to relax their preferences and share their finest location details even when doing so is not their optimal choice, whereas Chinese participants are more likely to do the opposite. A significant implication of this finding is that granularity settings are likely to be more important for the adoption of location sharing among Chinese users than among American users.

Other findings we observed included the mobility differences and the subtle gender differences between the two groups. Our study was the first exploration into the differences of location-sharing preferences between participants of two countries, yielding several design implications for future location-sharing applications (LSA). First, LSAs should consider providing different levels of privacy assurance to users with different cultural backgrounds. Second, different cultures may have different control requirements for sharing their location data. For example, we observed that Chinese participants needed specific control over the time when their locations were shared, while data from U.S. participants suggest that the type of place they visit might be enough. Third, we found that participants’ sharing preferences were dramatically different when given additional control (e.g., granularity control) over the detail of their shared location information. This finding suggests that introducing a more complex control mechanism could increase users' comfort levels; however, it also might encourage or discourage users to share more information.

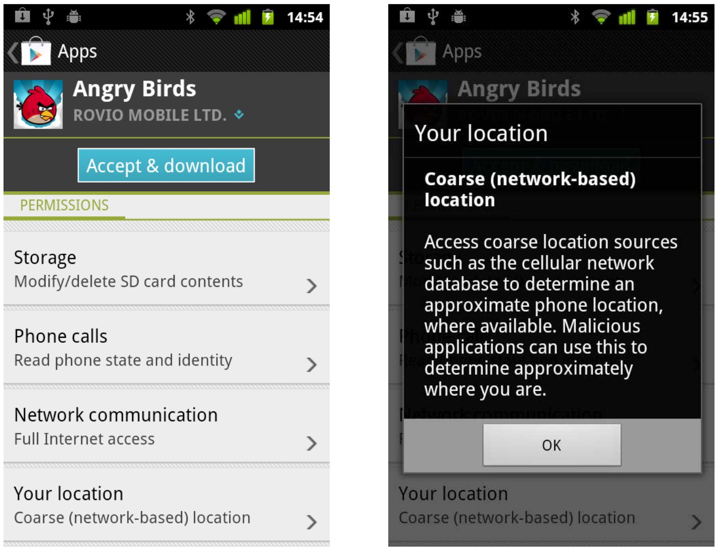
2.3 Summary and Lessons Learned

Previous research (including my own work) has provided important knowledge in understanding users privacy concerns and needs in mobile context-sharing. Although, location-sharing applications only take up a small portion of the spectrum in mobile services, lessons learned from this domain could still be inspiring and extended to mobile app privacy, in general and from multiple perspectives.

First, considering methodology, multiple user studies [14, 61, 79] have shown that remote auditing-based study methods (i.e., participants provide their responses remotely through web service) and the Day Reconstruction Method could provide equivalent data quality and are easier to manage compared to lab studies, interviews and the Experience Sampling Method in collecting participants' subjective feedback. However, we are fully aware of the limited scalability of the study methods we used in the past work given the number of mobile apps we want to investigate. Therefore, to tackle this challenge, we propose that mobile privacy user studies should take advantage of crowdsourcing to harvest user data with regard to privacy preferences. By carefully designing human intelligent tasks (HITs), a crowdsourcing platform (e.g., Amazon Mechanical Turk) could provide an efficient way to collect subjective feedback on mobile apps from a diverse population of users in a relatively short amount of time.

Second, we learned that users' privacy preferences are dynamic and complex. Considering location-sharing as an example, users typically consider multiple factors when they make decisions on whether and what to share. Factors that influence users' mental models range from type of requester, temporal or special context, current activity, motivation of sharing, cultural influence, etc. We believe that the complexity of the problem space addressed in this proposed thesis could be even higher given that more facets (e.g., app functionalities, different types of private resources) are involved. Previous research has suggested that users need multidimensional controls to manage sharing preferences [14, 22, 80, 81], which also might be true for mobile app privacy preferences. To make these controls really usable without overwhelming users, this thesis aims to identify the sensitive data usage scenarios that most significantly impact users' privacy concerns, in order to simplify the decisions users have to make.

Finally, several studies [23, 24, 61] have shown that user privacy preferences, though complex, are still predictable to some extent. This important lesson greatly inspired our proposed work. As such, in part of the proposed work, we want to determine to what extent machine learning techniques could help in modeling users' mobile app privacy preferences. As part of this study, we will explore the use of clustering techniques and investigate to what extent clusters of apps and users can be identified and generalized to inform the design of streamlined mobile privacy interfaces.



Angry Birds
ROVIO MOBILE LTD.

Storage
Modify/delete SD card contents

Phone calls
Read phone state and identity

Network communication
Full Internet access

Your location
Coarse (network-based) location

Angry Birds

Your location

Coarse (network-based) location

Access coarse location sources such as the cellular network database to determine an approximate phone location, where available. Malicious applications can use this to determine approximately where you are.

Your location
Coarse (network-based) location

Figure 2: Example permission screen in Google Play Store (left). When a user clicks on an entry from the permission list, more explanation will be shown (right). Previous research showed that most users click through the permission screen without carefully examining this list.

3 Related Work of Mobile App Privacy

For the purpose of my thesis, the related work of mobile app privacy falls into four major themes. First, I provide a brief overview of the permission framework used in the Android system[4]. Then, I survey existing research on mobile app analysis and security extensions. Finally, I briefly discuss relevant research in crowdsourcing and understandable machine learning.

3.1 Android Permissions

The Android permission framework is intended to serve two purposes to protect users: (1) limit the access of mobile apps to sensitive resources and (2) assist users in making trust decisions before installing apps. The latest Android 4.3 platform defines 11 permission groups with more than 130 permissions [43]. Android apps can only access sensitive resources if they declare permissions in the manifest files and obtain approval from users at installation. At the official Google Play Store, before installing an app, users are shown a permission screen that lists resources an app will access. It is this information that users must use to decide whether to trust the app (see Figure 2). In order to proceed to installation, users need to accept all the permissions. Once granted, permissions cannot be revoked unless the user uninstalls the app.

Several user studies have examined usability issues of permission systems in warning users before downloading apps. Kelley et al. [57] conducted semi-structured interviews with Android users and found that users paid limited attention to permission screens and had poor

[4] Throughout this thesis, all the studies and analysis are conducted on Android platform. We choose Android system due to its openness and its more outstanding privacy problems compared to other platforms.

understanding of what the permissions implied. Specifically, permission screens generally lack adequate explanation and definitions. Felt et al. [38] found similar results from Internet surveys and lab studies that current Android permission warnings do not help most users make correct security decisions. Later Felt et al. [35] also surveyed more than three thousand smartphone users about 99 risks associated with 54 permissions without considering specific apps. Their survey focused more on security risks that malicious apps can exploit rather than the potential privacy concerns caused by normal mobile apps. Similarly, an interview study by Chin et al. [21] probed smartphone users' concerns and fears with regard to privacy and security and offered several recommendations that could mitigate these threats.

Different from their work, my thesis will leverage crowdsourcing to evaluate individual mobile apps within their specific context (e.g., an app's functionality), resulting in a more in-depth probing of users' mobile app privacy preferences.

3.2 Mobile Application Privacy Analysis

To handle the increasing rate of malware in the Android market, in Feb 2012, Google announced their "Bouncer" service that scans apps for malware, spyware, Trojans, and other suspicious behaviors [88]. Though its technical details are not publically available, Bouncer was developed to detect malicious apps rather than privacy intrusive apps.

Researchers have also developed many useful techniques and tools to detect sensitive information leakage in mobile apps [10, 13, 15, 20, 30-34, 37, 39, 49, 77, 82, 87]. Three methods are usually used in app analysis, namely permission analysis, static code analysis, and dynamic flow analysis. Permissions describe what an application can do once installed. Table 1 categorizes previous research and studies based on methods used and highlights the pros and cons of each method.

By analyzing the permission lists declared by app developers, potentially risky functionalities can be identified. This line of research has focused on how different permissions are used [13, 34, 39, 82] and highlights common usage patterns [13], misuses [37, 82], and potential implications to Android security and privacy [34, 37, 39]. Enck et al. [33] were the first to conduct permission analysis on the Android system. Among the 311 apps they examined, 10 apps were flagged with questionable private resource usage. Barrera et al. [13] performed permission analysis of 1,100 free applications in the Android Market and identified the exponential decay distribution in the number of applications that requested individual permissions (i.e., most applications require only a small number of permissions). Felt et al. [37] studied the effectiveness of Android's install-time permission. Specifically, Felt et al. found that developers sometimes made mistakes in declaring permissions requests (e.g., requesting unnecessary permission, non-existing permission, etc.). Hence, in follow-up work [34], Felt et al. proposed the Stowaway tool, which performs static analysis to detect over-privileged applications. Similarly, an Android SDK extension was developed by Vidas et al. [82], which assisted Android developers in including the minimum set of permissions required by their app's functionality. Permissions are valuable for performance efficient security analysis; however, permission lists could not provide detailed

	Permission Analysis	**Static Analysis**	**Dynamic Analysis**
Examples	Enck’09 [33] Barrera’10 [13] Felt &Greenwood’11 [37] Felt&Chin’11 [34] Vidas’11 [82]	Egele’11 [29] Chin’11 [20] Felt&Wang’11 [39] Enck’11 [32, 38] App Profiles [10]	Thurm’11[29, 77] Enck’10(TaintDroid) [31] Beresford’11 [15] Zhou’11 [87] Hornyack’11 [49]
Pros	Simple and efficient	Easy to automate, cover all possible execution patterns	Capture what actually happened, easy to interpret
Cons	Only high-level analysis cannot tell the whole story	Dead code problem; Depend on the performance of decompiler	Require human intervention, hard to automate

Table 1 : Categorization of existing work in mobile app analysis based on methodologies. The pros and cons of each method are highlighted. All methods assessed mobile apps’ behaviors from traditional security perspectives that cannot infer users’ perceptions of mobile privacy. Our proposed work makes use of the app analysis tool to obtain ground truth of mobile apps, aiming at bridging the gap between app analysis and users’ privacy preferences learning.

information concerning what purpose private resources would be used, but rather could only capture limited security and privacy risks.

Static program analysis can be conducted with or without source code. To date, most mobile app static analyses rely on decompilers to recover source codes of targeted apps (e.g., [9, 70]). Egele et al. [29] proposed PiOS to perform static taint analysis on iOS application binaries to identify potential privacy violations. Among the 1,400 apps studied, more than half leaked the privacy sensitive device ID without the users' knowledge. Chin et al. [20] proposed ComDroid, which operates on used disassembled DEX bytecode. Specifically, ComDroid identifies vulnerabilities in intent communications between applications, such as broadcast theft, service hijacking, malicious service launch, etc. Among 100 apps analyzed, Chin et al. found 34 exploitable vulnerabilities. App Profiles [10] developed by the RubustNet research group at the University of Michigan analyzed mobile applications offline to detect privacy-related actions written into the application source code. While static analysis provides a complete and automated scan of mobile apps, its accuracy might highly depend on the performance of the decompiler used or the code style of the developer. Another challenge for privacy research involving static analysis is that this method cannot automatically determine whether privacy-related behavior is desired.

Dynamic analysis can help resolve ambiguity in permission granularity as well as provide an intuitive way to monitor how applications run. The *Wall Street Journal* reported the results of 101 popular smartphone apps for iPhone and Android devices that were examined by monitoring network analyses [77]. Results showed that 56 apps transmitted the phone's unique ID to third party servers without user consent, and 47 apps transmitted the phone's location and other personal information such as age, gender, etc. TaintDroid [31] performed a thorough dynamic flow analysis to capture information leakage on Android devices in real time. The authors modified the Android's Dalvik VM to perform instruction-level taint tracking that captures how private information flows from its source to its destination (i.e., network interface). Other work has built on TaintDroid to provide more pertinent privacy analyses or controls [15, 49]. Dynamic analysis identifies what actually happens when an application is running. One drawback of dynamic analysis is that it is limited by scalability because human interventions (interactions with mobile apps) are needed to trigger certain behaviors of the apps in the process of analysis.

Though app analysis provides us with a better understanding of apps' behaviors, it cannot infer people's perceptions of privacy or distinguish between behaviors which are necessary for an app's functionality versus behaviors which are privacy-intrusive. Our work complements this past work by suggesting an alternative way of looking at mobile privacy from the users' perspective by leveraging crowdsourcing to bridge the gap between app analysis and resolving users' privacy concerns. To achieve this goal, we opt to use static analysis to capture the ground truth of apps with regard to type and purpose of information disclosed because of the scalability issue.

3.3 Security and Privacy Extensions

Many security extensions have been developed to harden privacy and security. MockDroid [15] and TISSA [87] substituted fake information into API calls made by apps, such that apps could still function, but with zero disclosure of users' private information. In addition to faking information, AppFence [49], a subsequent project of TaintDroid, allowed users to specify which resources should only be used locally. It also hashed the phone identifiers in a way that it no longer could be linked to users, while still being useful for application developers to track application usage.

Nauman et al. [69] proposed Apex, which provides fine-grained control over resource usage based on context and runtime constraints such as the location of the device or the number of times a resource has been used. They implemented an extended package installer named Poly that allows users to specify their policy at time of install. To enable wide deployment, Jeon et al. proposed an alternative solution that rewrote the bytecode of mobile apps instead of modifying the Android system [54]. When accessing sensitive resources, the modified apps talk to a privacy proxy layer instead of directly talking to Android APIs. Nauman et al. also proposed fine-grained permissions that could further control resource access.

These proposed privacy extensions aimed to provide users more control over apps and assumed that users are able to configure these settings perfectly. However, this assumption was not grounded by user studies. Dumping these settings on users and relying on them to specify their privacy preferences without adequate information could be questionable or even counterproductive.

3.4 Crowdsourcing and Human Computation

Crowdsourcing and human computation has gained attention as both a topic of and tool for research. Several methodological papers have addressed how to more effectively utilize crowdsourcing to yield better results [28, 52, 65, 66, 72]. Amazon's Mechanical Turk (AMT)[1] is currently the most popular crowdsourcing platform and the one used in this work. With AMT, requesters can publish Human Intelligence Tasks (HITs) for workers. A number of projects have successfully used AMT and have ranged from human assisted online tasks (such as image labeling) to surveys and user studies [16, 40, 48, 62, 63, 85].

Our work makes use of many findings and methodologies mentioned above and builds on past work by extending the application of crowdsourcing to a mobile privacy study. In so doing, we demonstrate the feasibility and potentials of crowdsourcing as a scalable tool for privacy studies.

3.5 Understandable Machine Learning

Most traditional learning problems are solved by a black box approach in terms of model selection and/or parameter estimation, aiming at optimizing the mapping between the input and output of the given data. The resulting models of these approaches are by and large obscure and not understandable by humans, which makes the knowledge discovery difficult. As suggested in [56], the interpretability of machine learning models depends strongly on the complexity of the model, and in general, the lower the complexity, the easier it is to understand. Furthermore, by constraining the complexity of resulting models, such as by penalizing the model complexity in the objective function, models' generalizability could be potentially improved [41, 55, 86].

Sadeh et al. first introduced the notion of understandable learning into privacy research [23, 68]. They used two types of user-oriented machine learning techniques, namely default personas and incremental suggestions, to identify users' privacy rules, resulting in a significant reduction of user burden. By restricting the level of control the user has over the policy model, their algorithm produced accurate and understandable learning results.

Our work inherits the rationals of their work, aiming at extracting logic and knowledge from users' mobile app privacy preferences. Hence, we take interpretability and generalizability as two of the crucial criteria in modeling users' preferences.

3.6 Distinction from Prior Work

Before moving on to the details of this thesis, a few high-level distinctions between my proposed thesis work and past related work are worth mentioning. From a technology standpoint, this thesis does not aim to produce new tools or techniques to analyze mobile apps' privacy related behaviors, rather it aims to link users' subjective feedback to various private resource usage patterns as identified through app analyses. Meanwhile, the security extensions mentioned above do provide users with more control over private data; however, these designs are not grounded in adequate user studies. Further, it is unclear if these are the settings users need and if lay users can correctly configure these settings to reflect their desired preferences.

My thesis will complement this past work by assessing mobile app privacy from the user perspective. We expect to identify a relatively small number of sensitive data usage scenarios that most significantly impact users' privacy concerns when using a particular mobile app.

From an HCI standpoint, this thesis probes much deeper in the users' privacy decision processes compared to previous permission usability studies [38, 57] or privacy surveys and interviews [21, 35]. By performing clustering, we expect to isolate different classes of mobile apps and different groups of users with distinct characteristics. A small number of user-understandable privacy profiles (or "personas") will be learned based on the clusters of users with similar preferences. These expected findings can provide practical suggestions to inform the design of simpler, easier-to-use interfaces and privacy control mechanisms that matter to users.

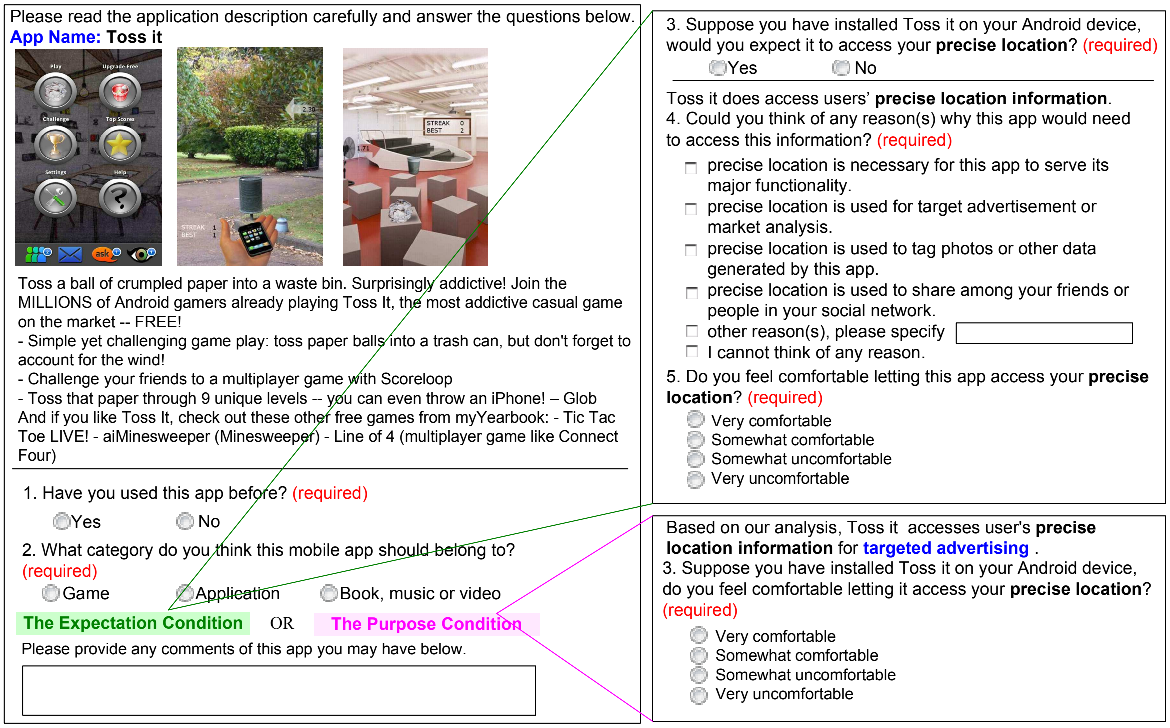
4 Preliminary Results: A Formative Study of Understanding Users' Mental Models of Mobile App Privacy [59]

Smartphone security research has produced many useful tools to analyze privacy-related behaviors of mobile apps. However, these automated tools cannot assess user perceptions of whether a given action is legitimate, or how that action makes them feel with respect to privacy. For example, is the use of a given app concerning one's location appropriate? The answer depends on the context[5]: for a blackjack game, probably not, but for a map application, very likely so. However, currently, end-users have very little support in making good trust decisions regarding what apps to install. The major goal of this formative study is to understand the design space and feasibility of our ideas.

More specifically, with this study we aimed to achieve the following four objectives. First, we investigated users' mental models in terms of their expectations about what an app does and does not do with a focus on where an app breaks people's expectations. We argue that by allowing the user to see the most common misconceptions about an app, we can rectify mental models and help users make better trust decisions regarding that app. Second, we demonstrated an efficient way to capture users' privacy preferences with a crowdsourcing platform (e.g., AMT), which produces results with similar quality as lab studies. Third, we identified two key factors that affect people's mental models of a mobile app, namely expectation and purpose and demonstrate how they influence users' subjective feelings. Finally, we evaluate a preliminary design of a new privacy summary that features expectations and purpose, which significantly increases users' privacy awareness and is easier to comprehend than Android's current permission interface.

4.1 Crowdsourcing Users' Mental Models

[5] Obviously the fact that an app requires access to one's sensitive resources does not guarantee that this information is used solely for the purpose of supporting its core functionality.



Please read the application description carefully and answer the questions below.

App Name: **Toss it**

Toss a ball of crumpled paper into a waste bin. Surprisingly addictive! Join the MILLIONS of Android gamers already playing Toss It, the most addictive casual game on the market -- FREE!
- Simple yet challenging game play: toss paper balls into a trash can, but don't forget to account for the wind!
- Challenge your friends to a multiplayer game with Scoreloop
- Toss that paper through 9 unique levels -- you can even throw an iPhone! – Glob
And if you like Toss It, check out these other free games from myYearbook: - Tic Tac Toe LIVE! - aiMinesweeper (Minesweeper) - Line of 4 (multiplayer game like Connect Four)

1. Have you used this app before? (required)

○Yes ○ No

2. What category do you think this mobile app should belong to? (required)

○Game ○Application ○Book, music or video

The Expectation Condition OR **The Purpose Condition**

Please provide any comments of this app you may have below.

3. Suppose you have installed Toss it on your Android device, would you expect it to access your **precise location**? (required)

○Yes ○ No

Toss it does access users' **precise location information**.
4. Could you think of any reason(s) why this app would need to access this information? (required)

□ precise location is necessary for this app to serve its major functionality.
□ precise location is used for target advertisement or market analysis.
□ precise location is used to tag photos or other data generated by this app.
□ precise location is used to share among your friends or people in your social network.
□ other reason(s), please specify
□ I cannot think of any reason.

5. Do you feel comfortable letting this app access your **precise location**? (required)

○ Very comfortable
○ Somewhat comfortable
○ Somewhat uncomfortable
○ Very uncomfortable

Based on our analysis, Toss it accesses user's **precise location information** for **targeted advertising** .
3. Suppose you have installed Toss it on your Android device, do you feel comfortable letting it access your **precise location**? (required)

○ Very comfortable
○ Somewhat comfortable
○ Somewhat uncomfortable
○ Very uncomfortable

Figure 3: Sample questions to capture users' mental models. Participants were randomly assigned to one of the conditions. In the *expectation condition*, participants' were asked to specify their expectations and speculate about the purpose for this resource access. In *the purpose condition*, the purpose of resource access was given. In both conditions, participants were asked to rate how comfortable they felt having the targeted app access their resources.

Taking a step back, there are four reasons why crowdsourcing is a compelling technique for examining privacy. Past work has shown that few people read End-User License Agreements (EULAs) [42] or web privacy policies [53] because (a) there is an overriding desire to install the app or use the web site, (b) reading these policies is not part of the user's main task (which is to use the app or web site), (c) the complexity of reading these policies, and (d) there is a clear cost (i.e., time) with unclear benefit. Crowdsourcing nicely addresses these problems because it dissociates the act of examining permissions from the act of installing apps. By paying participants, we make reading these policies part of the main task and offer a clear monetary benefit. Lastly, we can reduce the complexity of reading Android permissions by having participants examine just one private resource at a time, rather than all the permissions, and by offering clearer explanations of what the permission means.

We recruited participants using Amazon's Mechanical Turk (AMT). We designed each Human Intelligence Task (HIT) as a short set of questions about a specific Android app and resource pair (see Figure 3). Participants were shown one of two sets of follow-up questions. One condition (referred to as *the expectation condition*) was designed to capture users' perceptions of whether they expected a given app to access a sensitive resource and why they thought the app used this resource. Participants were also asked to specify how comfortable they felt allowing this app to access the resource using a Likert scale that ranged from very comfortable (+2) to very

MSE	Network Loc	GPS loc	Contact List	Unique ID
expectation [0,1]	0.0354	0.0303	0.0353	0.0363
comfort level [-2,+2]	0.7081	0.8136	0.6749	0.3067

Table 2: Crowd workers and experts have similar expectations toward targeted mobiles. In general, experts were slightly more skeptical about these privacy-related behaviors. Numbers in this table indicate the differences between the rating we obtained from the crowd workers and the experts, measured by the Mean Square Error.

uncomfortable (-2). In the other condition (referred to as *the purpose condition*), we wanted to see how people felt when offered more fine-grained information. Participants were told that a certain resource would be accessed by this app and were given specific reasons for the access. We manually identified these reasons by examining TaintDroid logs and using knowledge about ad networks. Participants were then asked to provide their comfort ratings as in the expectation condition. Finally, participants from both conditions were encouraged to provide optional comments on the apps in general. The separation of the two conditions allowed us to compare users' perceptions and subjective feelings when different information was provided.

We focused our data collection on four types of sensitive resources (as suggested by AppFence [49]): unique device ID, contact list, network location, and GPS location. We also restricted the pool of apps to the Top 100 most downloaded mobile apps on the Android market. We limited our participants to Android users and ensured a between-subjects design through a qualification test. This study included 179 verified Android users with an average lifetime approval rate of 97% (SD = 8.79%). The distribution of Android versions that our participants used was very close to Google's official numbers [83]. On average, participants spent about one minute per HIT (M = 61.27, SD = 29.03) and were paid at the rate of $0.12 USD per HIT.

4.2 Major Findings

4.2.1 Feasibility of Using Crowdsourcing to Study Privacy

Though we already adopted quality control questions and qualification tests to ensure the validity of the data collected, we want to prove quantitatively that the crowdsourcing approach would not bias the results in gathering users' subjective feedback. To this end, we recruited five Android experts[6] to come to our lab; then we presented them with the same questions in the expectation condition and asked them to complete the questions for every resource and app pair (i.e., 134 sets of questions in total). We used the Mean Square Errors (MSE) to measure the differences between the subjective feedback collected from crowd workers and experts (see Table 2). In general, crowd workers had a similar level of expectations as experts (i.e., MSE < 0.05). Experts on average appeared to be more skeptical about privacy-related behaviors of apps, which attributed to the slightly higher MSEs seen in the second row. Given the comfort level scaling from -2 to +2, these MSE were still considered acceptable. In other words, these results demonstrate the validity and feasibility of crowdsourcing as a method to collect users' subjective feedback to study privacy.

[6] Someone with security background and has development experience in Android OS.

Resource Type	comfort rating w/ purpose	comfort rating w/o purpose	df	T	p
Device ID	0.47(0.30)	-0.10(0.41)	55	7.42	0.0001
Contact List	0.66(0.22)	0.16(0.54)	24	4.47	0.0002
Network Location	0.90(0.53)	0.65(0.55)	28	3.14	0.004
GPS Location	0.72(0.62)	0.35(0.73)	23	3.60	0.001

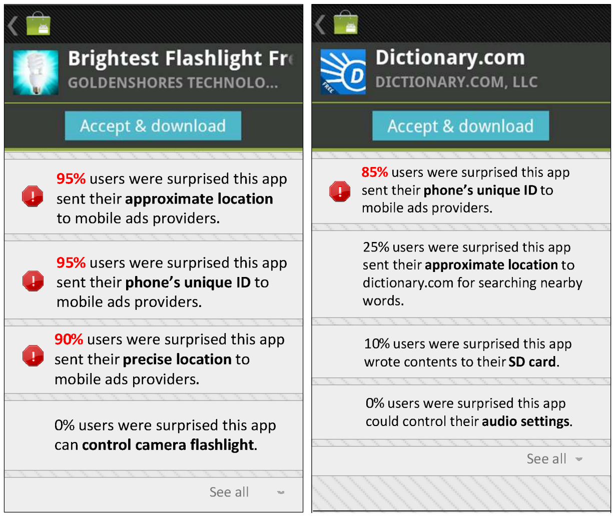
Table 3 Comparison of comfort ratings between the expectation condition (2nd column) and the purpose condition (3rd column). Standard deviations are shown between parentheses. When participants were informed of the purpose of resource access, they generally felt more comfortable. The differences were statistically significant for all four types of resources. The comfort ratings were ranging from -2.0 (very uncomfortable to +2.0 (very comfortable).

4.2.2 Expectation, Purpose, and Comfort Level

When participants were surprised by access to a sensitive resource, they also found it difficult to explain why the resource was needed. Note, in the *expectation condition*, participants were only informed about which resources were accessed without information on the purpose of access. This is similar to what the existing Android permission list conveys to users. In this condition, we observed a very strong correlation (*r* = 0.91) between the percentage of expectations and average comfort ratings. In other words, the **perceived necessity of resource access was directly linked to users' subjective feelings**, which guided the way users made trust decisions on mobile apps.

We also found that, even if users were fully aware of which resources were used, they still **had a difficult time understanding why the resources were needed**. We compared the reasons our participants provided in the expectation condition against the ground truth from our app analysis. In most cases, the majority of participants could not correctly state why a given app requested access to a given resource. When resources were accessed for functionality purposes, participants generally had better answers; however, accuracy never exceeded 80%. When sensitive resources were used for multiple purposes, the accuracy of answers tended to be much lower. Note that these results are for a situation in which participants were paid to read the description carefully. Many of them had even used some of these apps in the past. We believe for general Android users, their ability to guess answers would have been even worse.

Given the lack of clarity of why resources are accessed, users must deal with significant uncertainties when making trust decisions regarding installing and using a given mobile app. We observed that, for the four types of sensitive resources (i.e., device ID, contact list, network location, and GPS location), **participants, in general, felt more comfortable when they were informed of the purposes of a resource access** (see **Table 3**). The differences between the comfort ratings were statistically significant in paired *t*-tests. For example, concerning accessing the device ID, the average comfort rating in the purpose condition was 0.3 higher than in the expectation condition ($t(55) = 7.42$, $p < 0.0001$). This finding suggests that providing users with reasons why their resources are used not only gives them more information to make better trust decisions, but can also ease concerns caused by uncertainties. Note that informing users about the "purpose" for collecting their information is a common expectation in many legal and regulatory privacy frameworks. Our results confirm the importance of this information. This



Brightest Flashlight Fr
GOLDENSHORES TECHNOLO...

Accept & download

95% users were surprised this app sent their **approximate location** to mobile ads providers.

95% users were surprised this app sent their **phone's unique ID** to mobile ads providers.

90% users were surprised this app sent their **precise location** to mobile ads providers.

0% users were surprised this app can **control camera flashlight**.

See all

Dictionary.com
DICTIONARY.COM, LLC

Accept & download

85% users were surprised this app sent their **phone's unique ID** to mobile ads providers.

25% users were surprised this app sent their **approximate location** to dictionary.com for searching nearby words.

10% users were surprised this app wrote contents to their **SD card**.

0% users were surprised this app could control their **audio settings**.

See all

Figure 4: A mockup interface of our newly proposed privacy summary screen, taking the Brightest FlashLight and the Dictionary app as examples. The new interface provides extra information of why certain sensitive resources are needed and how other users feel about the resource usages. Warning sign will appear if more than half of the previous users were surprised about this resource access.

finding also provides us with a strong rationale to include the purpose(s) of resource access in our new design of privacy summary interface.

4.2.3 Impact of Previously Using an App

We also compared the responses between participants who had and had not used the app prior to this study. To make the comparison fair, we only examined apps that had at least five responses in both the used and not used categories.

According to our results, the differences between participants who had and had not used the apps were not statistically significant with respect to their expectation of sensitive resource access. Regarding their comfort level, the only significant difference we observed was the average comfort ratings for accessing the contact list. Participants who has used an app felt more comfortable allowing the app to access their contact list ($t(20) = 2.68$, $p = 0.015$). For the other three types of resources, the experiences with apps do not significantly impact participants' subjective feelings.

This finding suggests that **people who use an app do not necessarily have a better understanding of what the app is actually doing** in terms of accessing their sensitive resources. This finding also suggests that if we use crowdsourcing to capture users' mental models of certain apps, we do not have to restrict participants to people who are already familiar with the apps, which would allow us access to a larger crowd.

4.3 Preliminary Design and Evaluation of a New Privacy Summary Interface

We argue that our findings in exploring users' mental models can provide direct implications for the design of a future mobile privacy framework. To demonstrate this point, we drafted an initial design of a new privacy summary interface that features two crucial attributes identified in our formative study, namely expectation and purpose. In our new design, we directly leverage other users' mental models and highlight their surprises. By presenting the most common misconceptions about an app, we can rectify people's mental models and help them make better trust decisions. We also provide the *purposes of resource access* to give users more explanations in our new summary interface. Further, we use simpler terms to describe relevant resources, apply appropriate highlights, and prioritize the resource list based on the surprising levels. Figure 4 shows two examples of our new privacy summary interface. To make the comparison more symmetric, our design uses the same background colors and patterns that are used in the current Android permission screen. In this study, we used the data collected in our previously described crowdsourcing study to mock up the privacy summary interfaces for five mobile apps, namely Brightest Flashlight Free, Dictionary, Horoscope, Pandora, and Toss it.

We used AMT to conduct a between-subjects user study to evaluate our new privacy summary interface. Participants were randomly assigned to one of two conditions. In one condition, participants were shown the original permission screen that the current Google Play Store uses. In the other condition, participants were shown our new interfaces. We evaluated the new privacy summary interface from three perspectives. The first was *privacy awareness* (i.e., whether users were more aware of the privacy implications). This was measured by counting the number of participants who mentioned privacy concerns when justifying their recommendation decisions. The second was *comprehensibility* (i.e., how well users understood the privacy summary). This was measured by the accuracy in answering questions about app behavior. The third was *efficiency* (i.e., how long it took participants to understand the privacy summary), which was measured by the number of seconds participants spent reading the privacy summary screens.

Generally speaking, participants in the new interface condition weighted their privacy more when they made decisions about whether the app was worth recommending. More people in this condition mentioned privacy-related concerns when they justified their choices. When we asked participants in both conditions to specify the resources used by the target apps, those in the new interface condition demonstrated a significantly higher accuracy compared to their counterparts. Furthermore, except for the Pandora app, participants in the new interface condition, on average, spent less time reading the privacy summaries; however, the time difference was not always statistically significant. This finding suggests that we can provide more useful information without requiring users to spend more time to understand it.

4.4 Discussion and Connections to Proposed Work

The findings of this formative study provided several important implications for mobile privacy analysis. A major take-home of our work is that informing users of reasons why their sensitive resources are needed is crucial for users' decision making. In fact, users generally feel more comfortable when they were informed of these reasons. Our results quantitatively showed that properly informing users of the purposes of resource usage could actually ease their worries. In other words, it would potentially benefit all parties, including app developers, market owners, and advertisers to include such reasons. In our proposed work, we will continue to identify other factors that could potentially influence users' subjective feelings toward mobile apps.

Secondly, we observed that mobile advertising services were a consistent privacy concern for most participants. For all four types of resources, users felt the least comfortable when these resources were used for advertising or market analysis. We understand that many developers rely on ads for income; therefore, blindly blocking advertising APIs from accessing users' resources is not practical and not healthy for the mobile app market ecosystem. In our proposed work, we will further study sensitive usage patterns of third party APIs and how these patterns affects users' privacy concerns, based on which we will identify collections of mobile apps that elicit different privacy concerns. We hope these identified clusters could help future developers to estimate major adopter and user acceptance of their apps and nudge them to develop more privacy preserving apps.

Thirdly, an important contribution of this work was to demonstrate the feasibility of using crowdsourcing to capture users' perceptions and identify the strengths and weaknesses of the crowd in evaluating privacy. Based on our experiment, users were not very good at speculating the purpose of resource access, which is not surprising and might be compensated by leveraging existing mobile app analysis techniques. However, specifying subjective feedback is a relatively easy job for most people. In the proposed work, we continue to use crowdsourcing as a major tool to collect users' subjective feedback. We envision that if market runners (e.g., Google Play, Amazon App Store, etc.) could crowdsource similar user feedback from real users by incorporating behavior review into app rating and commenting mechanism, generating the proposed permission screens could be easily scaled up to the whole app market.

As a formative study, we only captured users’ perceptions of a small number (Top 100) of mobile apps with limited types of sensitive resources. In the proposed work, we plan to extend the application pool to around 400 mobile apps with a more representative distribution of both free and paid apps across all 30 categories. In addition, to make our findings more generalizable, we will leverage user-oriented machine learning techniques to discover interpretable models of both mobile apps and users' privacy preferences, through which we can obtain practical insights to inform the design of mobile privacy interfaces and settings.

5 Proposed Work: Mobile App Clustering and Users' Mobile App Privacy Preference Learning

5.1 Objectives and Challenges

To recap, the fundamental goal of this thesis is to complement existing mobile privacy research by providing important knowledge on the end-users’ side. Although previous system-oriented research contributed valuable tools in analyzing mobile apps, their proposed privacy controls were primarily not grounded by adequate user studies. As such, dumping these settings on users and relying on them to specify their privacy preferences without adequate information could be questionable or error-prone. Given the complex problem space, challenges exist in data collection, analysis, and interpretation. More specifically, this thesis attempts to answer the following challenging questions:

- **Scalable** --- How can we effectively capture users’ privacy preferences of individual mobile apps in a scalable and efficient way? Scalability is crucial as traditional user study methods fall short given the number of apps and resources we want to cover in this research. It is also an important criteria of making the data collection process extendable and repeatable.

- **Concise** --- How can we simplify the data model by reducing the dimensions of the variables? The conciseness makes sure our resulting models don't overfit data by constraints on the complexity of the model. It can also promote knowledge discovery as suggested by [56].
- **Interpretable** --- How can we extract human-understandable interpretations from potentially complex quantitative models? The major objective of my proposed work is to complement user research in the mobile app privacy domain, as which a black-box type of quantitative model is not our ultimate goal. Instead, the underlying knowledge and rationales are what we are aiming for.
- **Generalizable** --- How can we generalize findings to a larger mobile app pool and larger population, given that there are more than half million mobile apps and diverse user populations? Our dataset only covers 0.1% of the apps in the market. It is crucial to test if the knowledge obtained from the 0.1% could be applied to the remaining apps.

These four challenges pose specific design criteria throughout each step of the proposed work.

To address these challenges and achieve our objectives, we will take four major steps:

Step 1: Using application analysis and crowdsourcing, we will compile datasets that capture privacy-related characteristics from representative samples of mobile apps as well as information on how users feel about these apps.

Step 2: We will perform regression and feature selection on the produced datasets to identify attributes that most significantly affect users' level of comfort. This step is expected to reduce the dimensionality of our dataset and identify how different attributes affect users' decision making processes.

Step 3: We will cluster mobile apps based on mobile apps' sensitive data usage patterns based on the ground truth we captured through static code analysis, which also could be extended to a predictive model that could estimate user acceptance of an app.

Step 4: We will learn individual participant's privacy preferences over different types of apps in the form of rule-based models. Based on the data we crowdsourced in the step one, we will produce clusters of participants by grouping participants with similar privacy rules. These clusters could be used to generate a small number of user-understandable privacy profiles (or "personas"). The objective will be to determine the extent to which these personas could be used as default settings to simplify the number of privacy decisions exposed to users without negatively impacting their sense of control.

In the following subsections, I will discuss each step in detail.

5.2 Step One: Data Collection

Here, I present the data collection procedures in detail, including how we select a representative sample of mobile apps and how we gather the attributes of each app from multiple sources.

5.2.1 Selection of Apps

For the purpose of this thesis, we plan to cover a representative sample of mobile apps in our datasets. In the formative study, we chose the Top 100 most popular free apps in the Google Play Store; therefore, we achieved better user coverage because participants in the formative

Category	% in market	Paid ratio	#paid	#free	Category	% in market	Paid ration	#paid	#free
Android Application					Sports	3.2%	24%	3	9
Books & Reference	7.4%	47%	13	15	Tools	7.4%	23%	6	22
Business	3.5%	9%	1	12	Transportation	1.0%	17%	0	3
Comics	1.0%	43%	1	2	Travel & Local	4.3%	33%	5	11
Communication	2.2%	19%	1	7	Weather	0.4%	17%	0	1
Education	5.6%	30%	6	15	Libraries & Demo	0.6%	10%	0	2
Entertainment	12.0%	20%	9	38	Total applications	86.8%		89	235
Finance	1.9%	15%	1	6					
Health & Fitness	2.2%	32%	2	5					
Lifestyle	6.3%	27%	6	18	Android Games				
Media & Video	1.8%	21%	1	5	Arcade & Action	3.3%	25%	3	9
Medical	1.0%	40%	1	2	Brain & Puzzle	5.0%	21%	4	15
Music & Audio	3.9%	13%	2	13	Cards & Casino	0.9%	27%	0	2
News & Magazines	2.9%	9%	1	10	Casual	3.2%	24%	3	9
Personalization	10.6%	60%	25	16	Sports Games	0.7%	26%	0	2
Photography	1.3%	25%	1	3	Racing	0.4%	15%	0	1
Productivity	2.9%	27%	3	8	Total games	13.4%		10	38
Shopping	1.2%	9%	0	4					
Social	2.3%	11%	1	8					

Table 4: Mobile apps covered in our dataset have a similar distribution as the official Google Play Store in terms of percentage of apps in each category and free vs. paid ratio. The first column shows the categories available in the Google Play Store, the second column is the percentage of apps that belong to each category, the third column shows the ratio of paid apps within each category, the fourth column is the number of paid apps that are selected in our datasets, and the last column shows the number of free apps selected. For example, there are 7.4% of total mobile apps belongs to the Book & Reference category; within this category, 47% are paid apps. Based on these ratios, we plan to select 13 paid apps and 15 apps in this category.

study were more likely to have experience with the target apps. However, this selection method is potentially biased toward apps with better quality and higher popularity. In the proposed work, we plan to extend the selection to around 400 apps, including both free and paid apps randomly selected from each category within the Google Play Store. The selection will also be based on the percentage each category takes within the whole market and the ratio between free and paid apps within each category. Table 4 summarizes the number of paid and free apps in our dataset selected from each category. The statistic of the official Google Play Store is obtained from Appbrain [12]. Within each category, apps will be randomly selected.

5.2.2 Dataset Construction

After determining the set of apps that will be covered in our dataset, we will collect the attributes of each app from the following three perspectives.

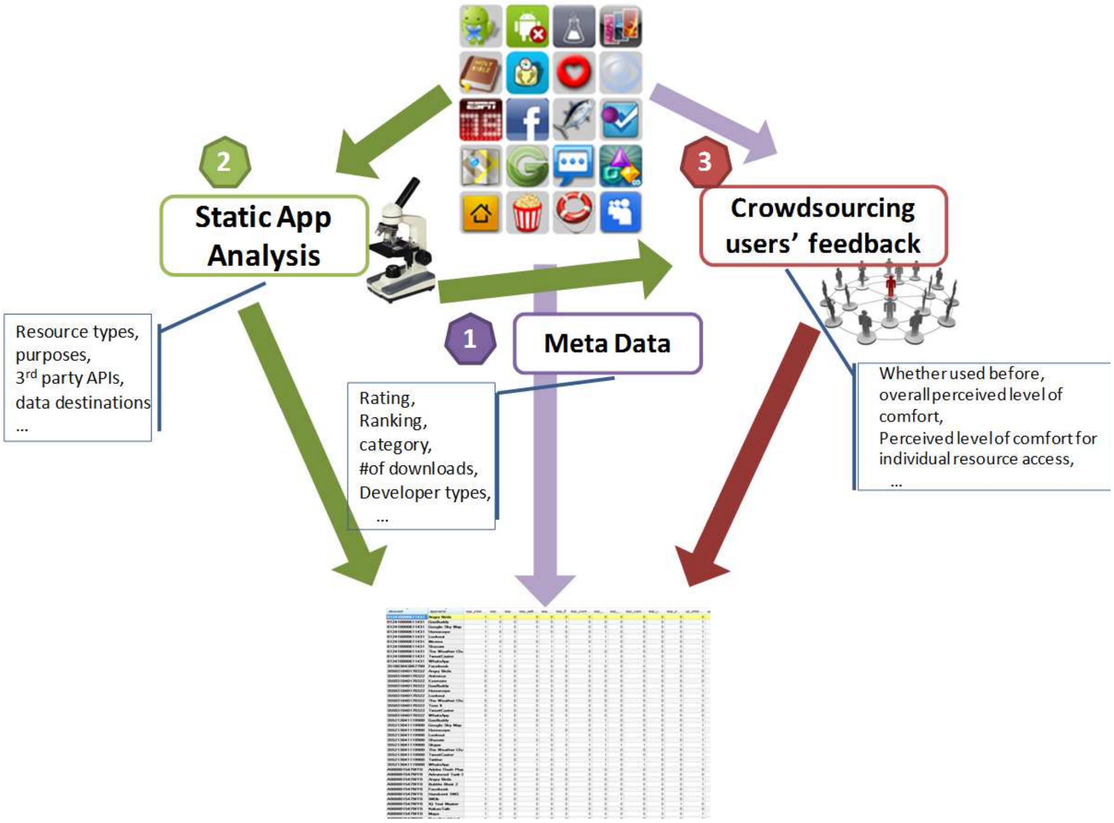


Figure 5: The dataset includes attributed of apps from three sources: (1) Meta data directly obtained from the official Google Play Store, such as app rating, number of downloads, category, etc., (2) behavior data obtained through static app analysis, such as sensitive resource usage, 3rd party API used, where data sent to, etc. and (3) users' subjective feedback captured through crowdsourcing, including whether they used these apps before and their perceived level of comfort in using these apps.

(1) App Meta data. These attributes include the name, description, rating, number of downloads, type of developer (developer badge), screen shots, etc. that can be easily obtained by crawling the relevant Google Play Store web pages.

(2) App behavior data. This data is obtained by conducting a detailed app analysis. It includes the type of sensitive resources an app consumes, for what purpose, the destination where an app sends users' sensitive data to, and the types of 3rd party APIs an app invokes. We will use static code analysis to scan the decompiled source codes of each target app and identify the function calls relevant to sensitive resource usages. We choose static analysis over dynamic analysis in this step because it is easier to automate. We will then examine the logs of static analysis and encode the results accordingly. For the purpose of this thesis, we will focus data collection on the Top 10 most sensitive resources, namely unique phone ID, external storage, coarse location, fine location, contact list, account info, call logs, SMS, audio recording, and camera.

(3) Users' subjective feedback of the target app. This data is obtained through crowdsourcing. Similar to our formative study, we will present crowd workers with apps' Meta data and behavior data and ask them to examine the sensitive resource usage of

the apps one-by-one by specifying how comfortable they feel about these usages. Participants will also be asked to specify their overall comfort level with the targeted app. For each app, approximately 20 unique responses are needed. To guarantee the quality of the crowdsourced data, crowd workers must be smartphone users to participate.

Figure 5 illustrates the data collection procedures. All attributes will be ultimately organized into a table, each line of which will describe one mobile app and how a particular participant feels about the app. This table may be sparse given that most apps access only a small number of sensitive resources.

5.3 Step Two: Preliminary Analysis

Following data collection, we will perform a preliminary analysis on the raw data. More specifically, we want to study the distributions of users' comfort levels of allowing mobile apps to access different types of sensitive resources. This analysis will greatly help us gain insights into users' privacy concerns and assist interpretations for further analysis. We will also perform linear regressions on overall user acceptance of a mobile app with comfort ratings of individual resources as independent variables to determine how different types of sensitive resources weigh in users' mental models.

After the preliminary probing, we will aggregate the data relevant to each app by averaging users' overall comfort levels. The resulting datasets will consist of two parts: (1) a matrix X, each row of which will describe app Meta data and attributes that describe privacy-related behaviors, (2) a vector Y, each entry of which will represent the average users' comfort level of the corresponding mobile app. Feature selection will be performed based on how each feature contributes to the predictability of the whole model and the degree of redundancy between these features. We will start with the simple correlation-based feature selection algorithm (such as [46]) and will also try other state-of-the-art feature selection algorithms such as [26, 45, 58, 67, 73, 78]. We opt not to use Principal Component Analysis (PCA) because this algorithm transforms existing features into another feature space, which will significantly increase the difficulty in interpreting the resulting model.

By performing feature selection, two objectives can be achieved. Feature selection can greatly reduce the complexity of our data model by eliminating redundant attributes and mitigating the risk of overfitting in the later analysis. On the other hand, as a by-product, feature select results can identify important factors that impact users' comfort level of an app, hence helping us better understand users' decision making processes.

5.4 Step Three: Mobile App Clustering

The purpose of performing clustering on mobile apps is two-fold. First, by categorizing mobile apps based on privacy-related resource usages, we can identify a set of common practices of how mobile apps consume users' sensitive data. These practices potentially differentiate users' privacy concerns and their acceptance in terms of how comfortable users feel about these information disclosures. Second, the set of clusters we identify could be extended to a predictive model to estimate user acceptance of apps. It could also demonstrate that the models we built on a small set of sample apps could be generalized to most mobile apps on the market. One thing we must keep in mind is that we are not aiming for a complicated model that will fit the model perfectly, rather we aim to develop a simplified version that can be interpreted and provide meaningful insights into the design space.

To cluster mobile apps, we need a clustering algorithm and a distance function. Because we are trying to identify clusters of apps that have different privacy implications (i.e., users' perceived level of comfort), we also need an objective function to guide the selection of clustering. This objective function measures the goodness of fit of the resulting clusters in estimating user acceptance and penalizes the complexity of the resulting models to prevent overfitting (e.g., AIC [11], TIC [74] measures). We will start with the simple *K*-mean clustering algorithm [64] with Hamming distance and explore other state-of-the-art clustering algorithms [47]. We opt to use Hamming distance as the distance function due to its simplicity and its semantic meaning.

The resulting clusters will be evaluated against the pre-defined objective function in cross-validation. The set of clusters with the highest predictability of user acceptance will be reported. We will also dissect the clustering models and try to explain why certain apps are grouped together. The resulting interpretation will provide us a better understanding of how users make privacy-related decisions over mobile apps.

5.5 Step Four: User Personas Generation

In brief, we will generate default personas using a three-step approach similar to [24] that involving the learning of individual users' policy and the clustering of all learnt policies.

Learning a Policy for Each User. Different from the approach used by Cranshaw et al. [24], we learn the policy of each user as follows. Suppose, in the previous step, we identified a set of meaningful clusters of mobile apps. We can use the resulting clusters to rearrange users' per-app preferences into a rule-based policy. In other words, a user's privacy policy of mobile apps is a set of rules with regard to each cluster of apps. These rules can be encoded into a vector, each entry of which represents the privacy preferences of a user with regard to one app cluster. To learn a user's privacy policy, we need to aggregate his or her comfort ratings of mobile apps in each cluster and organize the averaged comfort ratings into a feature vector.

A major challenge here is handling missing values. Because, in the crowdsourcing step, we cannot enforce a single crowd worker to express his or her feedback over all types of apps, for individual users and certain clusters of apps, it is very likely that there are no data to learn users' preferences. When this situation occurs, we propose two strategies. First, if the missing data situation is rare (e.g., occurs with fewer than 10% of the participants), we could potentially discard this portion of data and focus on learning policies from the participants with complete feature vectors. Alternatively, if the missing data situation is relevantly common, we could use the grand average of all users to substitute the missing data.

Clustering Policies into *K* Clusters Given that the policy vector of each user consists of continuous values in each entry, we choose to use the Euclidean distance as the distance function because of its simplicity. Additionally, we opt to use *K*-mean clustering and hierarchical clustering algorithms in this step because the resulting clusters are easier to interpret compared to other clustering algorithms. To restrict the *K* value to a small number, we can optimize model selection by penalizing the model complexity and the resulting number of clusters in the objective function. The clustering process is expected to identify groups of users with similar privacy preferences over different mobile app clusters.

Learning a Default Persona for Each Cluster. To generate a set of representative default personas, we need to aggregate policies within each cluster. In our model, we choose to compute the center of each cluster as a default persona. The center can be found easily by averaging the policy vectors in each cluster.

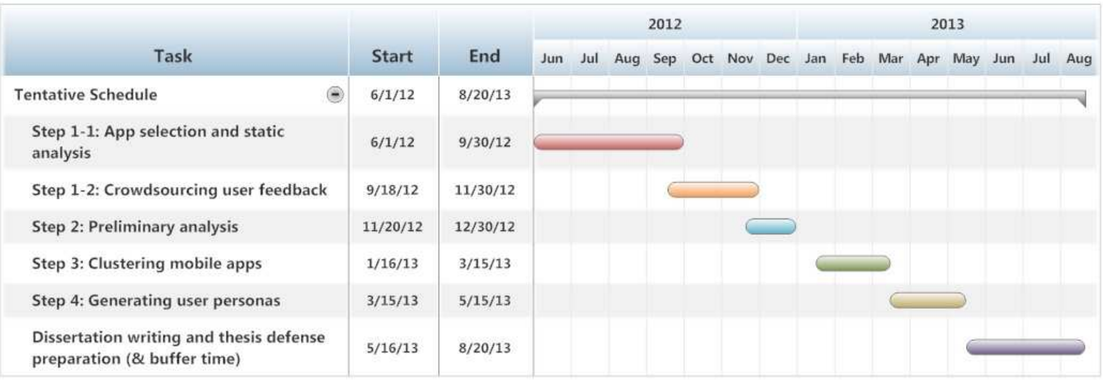
After we obtain the set of default persona, we will try to interpret the mathematical models into a set of human readable default settings. For example, a default persona might read, "I am willing to disclose (1) my location to mobile apps for the functionality purpose. I am NOT willing to disclose (1) my location for the advertising services or (2) my unique phone ID and call log to any service." We foresee that these default settings could dramatically reduce the user's burden in configuring his or her privacy preferences compared to specifying preferences for each individual app or each type of sensitive resource.

To evaluate how these generated personas match users' actually preferences, we will use the crowdsourcing methods described in 5.2.2 to capture the preferences of a group of new participants and to see whether their preferences could be covered by these personas. If time allows, a small scale lab studies with off-line participants can also be done by gathering their feedback of the apps actually running on their smartphones.

6 Scope

The proposed work can be extended in several directions that may NOT be included as part of this thesis but are worth pursuing as future work. These potential extensions include:

- Performing app analysis on a larger pool of mobile apps
  Extending the dataset to cover more mobile apps may improve the accuracy of resulting models. However, in the interest of time, the contribution of extending data collection might be marginal for the purposes of this work.
- Conducting a series of user studies to probe users' preferences in context-dependent scenarios
  My thesis focuses on investigating context-independent usage information of mobile apps. In reality, users' preferences may also vary in different context-dependent scenarios. For example, users might be OK disclosing most of their location to advertisers through mobile apps, but not some sensitive ones, such as clinic, rehab center. A series of user studies can



Task	Start	End
Tentative Schedule	6/1/12	8/20/13
Step 1-1: App selection and static analysis	6/1/12	9/30/12
Step 1-2: Crowdsourcing user feedback	9/18/12	11/30/12
Step 2: Preliminary analysis	11/20/12	12/30/12
Step 3: Clustering mobile apps	1/16/13	3/15/13
Step 4: Generating user personas	3/15/13	5/15/13
Dissertation writing and thesis defense preparation (& buffer time)	5/16/13	8/20/13

Figure 6: Proposed timeline for completing thesis work, ending in Aug 2013

be designed to cover these context-dependent cases. However, due to the time and the complexity of the study design, this thesis will not include these user studies.

- Investigating additional types of sensitive information
  The focus of this thesis is on the top ten sensitive data most frequently used by mobile apps. Other information, such as accelerometer readings, gyroscope readings and other derived information could be sensitive as well. Future work should take these types of information in consideration if more apps start to use them.
- Design and evaluate new privacy summary interfaces
  Although the finding of this work will provide valuable guidelines to inform the design of new privacy interfaces, this thesis may not include the actual design and evaluation of the interface as part of the deliverables.

7 Schedule

The proposed timeline for this thesis is shown in Figure 6. The data collection step is intended to be completed during the remainder of this summer and the early part of the Fall 2012 semester. I will then begin a preliminary analysis on the collected data and identify the most important attributes that affect users' privacy concerns by conducting feature selections. Starting in January 2013, I will focus on clustering and categorizing mobile apps based on their private resource usage. The goal is to have the clustering results ready for a Ubicomp 2013 submission. The remaining time will be spent on generating user personas, thesis writing, and defense preparation. I plan to graduate by Aug 2013.

8 Summary

The main propose of this thesis work is to complement existing mobile privacy research by providing important knowledge on the end-user's side given the lack of knowledge from the users' perspectives with regard to smartphone privacy. My previous work in location privacy revealed that people's privacy preferences are complex and varied, yet predictable to some extent by leveraging machine learning techniques. The proposed thesis extends this discussion to mobile app privacy, in general, and focuses on building better models of user privacy in the mobile context as well as using these models to inform the design of simpler, easier-to-use interfaces and privacy control mechanisms that matter to users.

The final deliverable of my thesis work will consist of (1) a valuable dataset that captures privacy-related characteristics of representative samples of mobile apps as well as information on how users feel about these apps; (2) a set of attributes that significantly impact users' privacy concerns of mobile apps; (3) collections of mobile apps obtained through clustering that elicit distinct privacy concerns; and (4) a set of default personas that users can select to configure their privacy preferences of mobile apps.

The findings of this thesis may provide important implications to improve current privacy frameworks. By identifying the attributes that influence users' level of comfort, my thesis will provide grounded suggestions concerning facts that should be presented to users before they install each app. The resulting clusters of apps could also help app developers understand how users make privacy-related judgments on mobile apps, thus nudging them to build more privacy-preserving mobile apps. Further, the generated personas will provide a more effective way for users to configure their mobile privacy settings and could potentially be refined and adopted in the design of future privacy frameworks.

9 Acknowledgement

I thank Prof. Norman Sadeh and Prof. Jason Hong for advising me on this work. Their expertise and insights greatly enriched this work. My committee members, Prof. Mahadev Satyanarayanan and Sunny Consolvo, have offered constructive questions and comments, which have greatly helped me improve the proposed work. I also thank other faculty and fellow students who provided valuable feedback on my thesis, in particular Janne Lindqvist, Joy Zhang, Nan Li, Shahriyar Amini, Jason Wiese, and Eiji Hayashi.

This work is supported by CyLab at Carnegie Mellon under grants DAAD19-02-1-0389 and W911NF-09-1-0273 from the Army Research Office and by Google. Support was also provided by the National Science Foundation under Grants CNS-1012763 and CNS-0905562.

10 References

[1]"Amazon's Mechanical Turk." Available: www.mturk.com
[2]"BrightKite." Available: http://brightkite.com
[3]"Facebook Places ". Available: http://www.facebook.com/places/
[4]"foursquare." Available: http://foursquare.com/
[5]"Google Latitude." Available: http://www.google.com/latitude
[6]"Locaccino: A User-Controllable Location-Sharing Tool." Available: http://www.locaccino.org/
[7]"Loopt." Available: http://loopt.com
[8]"Neer." Available: http://www.neerlife.com/
[9]"android-apktool." Available: http://code.google.com/p/android-apktool/
[10]"App Profiles." Available: https://play.google.com/store/apps/details?id=com.appdescriber&feature=search_result#?t=W251bGwsMSwxLDEsImNvbS5hcHBkZXNjcmliZXIiXQ..

[11]H. Akaike, "A new look at the statistical model identification," *Automatic Control, IEEE Transactions on,* vol. 19, pp. 716-723, 1974.

[12]AppBrain. "Most popular Android market categories." Available: http://www.appbrain.com/stats/android-market-app-categories

[13]D. Barrera, H. G. Kayacik, P. C. v. Oorschot, and A. Somayaji, "A methodology for empirical analysis of permission-based security models and its application to android," In Proc. *CCS*, 2010.

[14]M. Benisch, P. Kelley, N. Sadeh, and L. Cranor, "Capturing location-privacy preferences: quantifying accuracy and user-burden tradeoffs," *Personal and Ubiquitous Computing,* 2010.

[15]A. Beresford, A. Rice, and N. Sohan, "MockDroid: trading privacy for application functionality on smartphones," In Proc. *HotMobile*, 2011.

[16]M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, and K. Panovich, "Soylent: a word processor with a crowd inside," In Proc. *UIST*, 2010.

[17]B. Brown, A. Taylor, S. Izadi, A. Sellen, J. Kaye, and R. Eardley, "Locating Family Values: A Field Trial of the Whereabouts Clock," In Proc. *UbiComp*, 2007.

[18]I. Burguera, U. Zurutuza, and S. Nadjm-Tehrani, "Crowdroid: behavior-based malware detection system for Android," In Proc. *SPSM*, 2011.

[19]J. Cheng. "Pandora sends user GPS, sex, birthdate, other data to ad servers." Available: http://arstechnica.com/gadgets/news/2011/04/pandora-transmits-gps-gender-birthdate-other-data-to-ad-servers.ars

[20]E. Chin, A. P. Felt, K. Greenwood, and D. Wagner, "Analyzing inter-application communication in Android," In Proc. *MobiSys*, 2011.

[21]E. Chin, A. P. Felt, V. Sekar, and D. Wagner, "Measuring User Confidence in Smartphone Security and Privacy," In Proc. *Soups*, 2012.

[22]S. Consolvo, I. E. Smith, T. Matthews, A. LaMarca, J. Tabert, and P. Powledge, "Location disclosure to social relations: why, when, & what people want to share," In Proc. *CHI*, 2005.

[23]J. Cranshaw, J. Mugan, and N. Sadeh, "User-Controllable Learning of Location Privacy Policies with Gaussian Mixture Models," In Proc. *AAAI*, 2011.

[24]J. Cranshaw, J. Mugan, and N. Sadeh, "User-Oriented machine Learning for Capturing Privacy Preferences," In Proc. *under review*, 2012.

[25]J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh, "Bridging the gap between physical location and online social networks," In Proc. *UbiComp*, 2010.

[26]M. Dash and H. Liu, "Feature Selection for Classification," *An International Journal of Intelligent Data Analysis,* vol. 1, pp. 131-156, 1997.

[27]A. Diaconescu. "Smartphone Shipments to reach 614 Million Units Globally in 2012." Available: http://androinica.com/2012/01/smartphone-shipments-to-reach-614-million-units-globally-in-2012-according-to-new-projections/

[28]J. S. Downs, M. B. Holbrook, S. Sheng, and L. F. Cranor, "Are your participants gaming the system?: screening mechanical turk workers," In Proc. *Proceedings of the 28th international conference on Human factors in computing systems*, 2010.

[29]M. Egele, C. Kruegel, E. Kirda, and G. Vigna, "PiOS: Detecting Privacy Leaks in iOS Applications," In Proc. *NDSS*, 2011.

[30]W. Enck, "Defending Users against Smartphone Apps: Techniques and Future Directions," in *LNCS*. vol. 7093, ed, 2011.

[31]W. Enck, P. Gilbert, B.-G. Chun, L. Cox, J. Jung, P. McDaniel, and A. Sheth, "TaintDroid: An Information-Flow Tracking System for Realtime Privacy Monitoring on Smartphones," In Proc. *OSDI* 2010.

[32]W. Enck, D. Octeau, P. McDaniel, and S. Chaudhuri, "A Study of Android Application Security," In Proc. *USENIX Security Symposium*, 2011.

[33]W. Enck, M. Ongtang, and P. McDaniel, "On lightweight mobile phone application certification," In Proc. *CCS*, 2009.

[34]A. P. Felt, E. Chin, S. Hanna, D. Song, and D. Wagner, "Android permissions demystified," In Proc. *CCS*, 2011.

[35]A. P. Felt, S. Egelman, and D. Wagner, "I've Got 99 Problems, But Vibration Ain't One: A Survey of Smartphone Users' Concerns," UCB/EECS-2012-70, 2012.

[36]A. P. Felt, M. Finifter, E. Chin, S. Hanna, and D. Wagner, "A survey of mobile malware in the wild," In Proc. *SPSM*, 2011.

[37]A. P. Felt, K. Greenwood, and D. Wagner, "The effectiveness of application permissions," In Proc. *USENIX conference on Web application development*, 2011.

[38]A. P. Felt, E. Ha, S. Egelman, A. Haney, E. Chin, and D. Wagner, "Android Permissions: User Attention, Comprehension, and Behavior," In Proc. *Soups*, 2012.

[39]A. P. Felt, H. J. Wang, A. Moshchuk, S. Hanna, and E. Chin, "Permission re-delegation: attacks and defenses," In Proc. *USENIX conference on Security*, 2011.

[40]M. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin, "CrowdDB: Answering Queries with Crowdsourcing," In Proc. *SIGMOD*, 2011.

[41]F. Girosi, M. Jones, and T. Poggio, "Regularization theory and neural networks architectures," *Neural Comput.,* vol. 7, pp. 219-269, 1995.

[42]N. Good, R. Dhamija, J. Grossklags, D. Thaw, S. Aronowitz, D. Mulligan, and J. Konstan, "Stopping spyware at the gate: a user study of privacy, notice and spyware," In Proc. *SOUPS*, 2005.

[43]Google. "Android 4.0.x Platform." Available: http://developer.android.com/sdk/android-4.0.html

[44]S. Grobart. "The Facebook Scare That Wasn't." Available: http://gadgetwise.blogs.nytimes.com/2011/08/10/the-facebook-scare-that-wasnt/

[45]I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature selection," *Journal of Machines Learning Research,* vol. 3, pp. 1157-1182, 2003.

[46]M. A. Hall, "Correlation-based Feature Subset Selection for Machine Learning," University of Waikato, Hamilton, New Zealand, 1998.

[47]T. Hastie, R. Tibshirani, and J. Friedman, "14.3.Cluster Analysis," in *The Elements of Statistical Learning (2nd ed.)*, ed: New York: Springer., 2009.

[48]P. Heymann and H. Garcia-Molina, "Turkalytics: analytics for human computation," In Proc. *Proceedings of the 20th international conference on World wide web*, 2011.

[49]P. Hornyack, S. Han, J. Jung, S. Schechter, and D. Wetherall, "These aren't the droids you're looking for: retrofitting android to protect data from imperious applications," In Proc. *CCS*, 2011.

[50]S. Huang, F. Proulx, and C. Ratti, "iFIND: a Peer-to-Peer Application for Real-time Location Monitoring on the MIT Campus," In Proc. *CUPUM*, 2007.

[51]G. Iachello, I. Smith, S. Consolvo, G. D. Abowd, J. Hughes, J. Howard, F. Potter, J. Scott, T. Sohn, J. Hightower, and A. Lamarca, "Control, Deception, and Communication: Evaluating the Deployment of a Location-Enhanced," ed: unknown, 2008.

[52]P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on Amazon Mechanical Turk," In Proc. *Proceedings of the ACM SIGKDD Workshop on Human Computation*, 2010.

[53]C. Jensen and C. Potts, "Privacy policies as decision-making tools: an evaluation of online privacy notices," In Proc. *CHI*, 2004.

[54]J. Jeon, K. K. Micinski, J. A. Vaughan, N. Reddy, Y. Zhu, J. S. Foster, and T. Millstein, "Dr. Android and Mr. Hide: Fine-grained security policies on unmodified Android," 2012.

[55]Y. Jin, W. V. Seelen, and B. Sendhoff, "On generating FC<sup>3</sup> fuzzy rule systems from data using evolution strategies," *Trans. Sys. Man Cyber. Part B,* vol. 29, pp. 829-845, 1999.

[56]Y. Jin, B. Sendhoff, and E. Körner, "Evolutionary Multi-objective Optimization for Simultaneous Generation of Signal-Type and Symbol-Type Representations

Evolutionary Multi-Criterion Optimization." vol. 3410, C. Coello Coello*, et al.*, Eds., ed: Springer Berlin / Heidelberg, 2005, pp. 752-766.

[57]P. G. Kelley, S. Consolvo, L. F. Cranor, J. Jung, N. Sadeh, and D. Wetherall, "A Conundrum of permissions: Installing Applications on an Android Smartphone," In Proc. *USEC*, 2012.

[58]D. Koller and M. Sahami, "Toward optimal feature selection," In Proc. *13th International conference on Machine Learning*, 1996.

[59]J. Lin, S. Amini, J. Hong, N. Sadeh, J. Lindqvist, and Joy Zhang, "Expectation and Purpose: Understanding Users' Mental Models of Mobile App Privacy through Crowdsourcing," In Proc. *Ubicomp'12*, 2012.

[60]J. Lin, M. Benisch, N. Sadeh, J. Niu, J. I. Hong, B. Lu, and S. Guo, "A Comparative Study of Location-sharing Privacy Preferences in the U.S. and China," *PUC,* vol. under review, 2011.

[61]J. Lin, G. Xiang, J. I. Hong, and N. Sadeh, "Modeling people's place naming preferences in location sharing," In Proc. *UbiComp*, 2010.

[62]G. Little, L. B. Chilton, M. Goldman, and R. C. Miller, "TurKit: human computation algorithms on mechanical turk," In Proc. *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 2010.

[63]G. Liu, G. Xiang, B. A. Pendleton, J. I. Hong, and W. Liu, "Smartening the crowds: computational techniques for improving human verification to fight phishing scams," In Proc. *SOUPS*, 2011.

[64]J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations," In Proc. *5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281-297.

[65]T. W. Malone, R. Laubacher, and C. N. Dellarocas, "Harnessing Crowds: Mapping the Genome of Collective Intelligence," *SSRN eLibrary,* 2009.

[66]W. Mason and S. Suri, "Conducting Behavioral Research on Amazon's Mechanical Turk," *Behavior Research Methods, Forthcoming,* 2010.

[67]P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 24, pp. 301-312, 2002.

[68]J. Mugan, T. Sharma, and N. Sadeh, "Understandable Learning of Privacy Preferences Through Default Personas and Suggestions," *under review,* 2012.

[69]M. Nauman, S. Khan, and X. Zhang, "Apex: extending Android permission model and enforcement with user-defined runtime constraints," In Proc. *ASIACCS*, 2010.

[70]D. Octeau, W. Enck, and P. McDaniel, "The ded Decompiler. Technical Report NAS-TR-0140-2010, ," 2010.

[71]S. Patil and J. Lai, "Who Gets to Know What When: Configuring Privacy Permissions in an Awareness Application," In Proc. *CHI*, 2005.

[72]J. M. Rzeszotarski and A. Kittur, "Instrumenting the crowd: using implicit behavioral measures to predict task performance," In Proc. *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 2011.

[73]Y. Saeys, I. Inza, and P. Larraaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics,* vol. 23, pp. 2507-2517, 2007.

[74]K. Takeuchi, "Distribution of Informational Statistics and a Criterion of model fitting," *Mathematic Sciences,* vol. 153, pp. 12-18, 1976.

[75]K. P. Tang, J. Lin, J. I. Hong, D. P. Siewiorek, and N. Sadeh, "Rethinking location sharing: exploring the implications of social-driven vs. purpose-driven location sharing," In Proc. *UbiComp*, 2010.

[76]A. Thampi. "Path uploads your entire iPhone address book to its servers." Available: http://mclov.in/2012/02/08/path-uploads-your-entire-address-book-to-their-servers.html

[77]S. Thurm and Y. I. Kane, "Your Apps are Watching You," *WSJ,* 2011.

[78]R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal. Statist,* vol. Vol. 58, pp. 267-288., 1996.

[79]E. Toch, J. Cranshaw, P. H. Drielsma, J. Y. Tsai, P. G. Kelley, J. Springfield, L. Cranor, J. Hong, and N. Sadeh, "Empirical models of privacy in location sharing," In Proc. *UbiComp*, 2010.

[80]J. Y. Tsai, P. Kelley, P. Drielsma, L. Cranor, J. Hong, and N. Sadeh, "Who's Viewed You? The Impact of Feedback in a Mobile Location Sharing System," In Proc. *CHI*, 2009.

[81]J. Y. Tsai, P. G. Kelly, L. F. Cranor, and N. Sadeh, "Location-Sharing Technologies: Privacy Risks and Controls," In Proc. *TPRC*, 2009.

[82]T. Vidas, N. Christin, and L. Cranor, "Curbing android permission creep," *Proceedings of the Web,* vol. 2, 2011.

[83]A. Wagner. "Google Posts Refreshed Android Distribution Numbers." Available: http://www.twylah.com/surfingislander/tweets/177040176181288960

[84]R. Want, A. Hopper, and J. Gibbons, "The active badge location system," *ACM Trans. Inf. Syst.,* vol. 10, pp. 91-102, 1992.

[85]T. Yan, V. Kumar, and D. Ganesan, "CrowdSearch: exploiting crowds for accurate real-time image search on mobile phones," In Proc. *MobiSys*, 2010.

[86]J. Yaochu, "Fuzzy modeling of high-dimensional systems: complexity reduction and interpretability improvement," *Fuzzy Systems, IEEE Transactions on,* vol. 8, pp. 212-221, 2000.

[87]Y. Zhou, X. Zhang, X. Jiang, and V. W. Freech, "Taming Information-Stealing Smartphone Applications (on Android)," In Proc. *TRUST*, 2011.

[88]C. Ziegler. "Google unveils 'Bouncer' service to automatically detect Android Market malware." Available: http://www.theverge.com/2012/2/2/2766674/google-unveils-bouncer-service-to-automatically-detect-android-market