ContextProbe: Exploring Mobile Privacy in Context

by

Fuming Shih

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2015

Signature redacted

Author . . . .
Department of Electrical Engineering and Computer Science
January 10, 2015

Signature redacted

Certified by . . . .
Hal Abelson
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Signature redacted

Accepted by . . . .
Leslie A. Kolodziejski
Chairman, Department Committee on Graduate Theses

ContextProbe: Exploring Mobile Privacy in Context

by

Fuming Shih

Submitted to the Department of Electrical Engineering and Computer Science
on January 10, 2015, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

My research investigates the following question: What factors affect people's privacy preferences for disclosing data that is collected about them when interacting with mobile applications? Research about information privacy has revealed that there are relations between the role of context and people's expectations about privacy. But it is unclear how those findings can be applied to the ubiquitous environment where mobile apps operate. In order to illuminate this problem I have developed a framework, ContextProbe, which supports both quantitative and qualitative investigations of how user context and other external factors jointly affect people's willingness to disclose personal information to mobile apps.

As a consequence of this work, I have learned that people use contextual factors in making decisions about disclosing personal information to apps. Some of the significant privacy contextual factors are people's frequently visited places, specific time slots, who is around, and activities people are engaged in. Although contextual factors help, they do not provide a complete explanation of people's privacy choices. More importantly, I found that other external factors such as purposes of data use and trust in the app itself outweigh contextual factors when considering information disclosure. My study showed that subjects were not aware of context in thinking about disclosure when purpose of data use was presented together in the question. Surprisingly, results drawn from in-situ responses are the exact opposite to previous survey-based approaches on the effect of apps' showing their purpose strings when requesting personal information: showing less information seems to result in greater willingness to disclose.

ContextProbe has three major parts: app-building platform, personal data store, and application server. The app-building platform allows experimenters to create apps for ESM studies easily within a visual programming environment. Apps built by ContextProbe can be used to collect sensor data on mobile phones as well as subject-reported data for representing subjects' context. In addition, the apps can probe subjects' privacy preference in-situ with the detected context. The personal data store holds all data collected from subjects' phones and is responsible for sending data automatically to the corresponding application server. It provides a one-stop "dashboard" approach that lets subjects review information collected by the ESM apps. The application server aggregates all collected data in the study and monitors the health status of data collection tasks running on subjects' phone. ContextProbe provides an automatic process for study subjects and experimenters to easily set up personal data store and application server without extra overheads comparing to other existing architectures for ESM studies.

My work has opened up the following new questions: how do we best represent the information of privacy-relevant contexts during preference solicitation? And how to balance the trade-offs between sampling in various contexts and the cost of subjects' times? Further research in fields such as behavioral economics that require real-time monitoring of user context, data collection, and in-situ responses might well be conducted using the ContextProbe framework.

Thesis Supervisor: Hal Abelson
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

Foremost, I would like to thank my thesis advisor Professor Hal Abelson for his support and guidance in the pursuit of this dissertation. His enthusiasm, dedication, and energy in pursuing research about democratization of technology and information has changed my view about the role of computer scientists in shaping the healthy development of an information society.

I would also like to thank the members of my committee: Peter Szolovits, Debrah Estrin, and Daniel Weitzner for their invaluable advice and insightful questions in improving this thesis.

I want to express my gratitude to my group-mates for their friendly support throughout these years. Their companionship has made my time in the lab much more enjoyable and bearable. Special thanks to Joe Pato who always listened patiently to all my research problems and encouraged me when I was stuck. To Ilaria Liccardi, who taught me how to write better research papers and pushed me to work harder while being one of my best friends at MIT. To Jose Flores, who spent endless time in proofreading my thesis, helping me debug the code, and sharing laughter during our daily trips to the Media Lab for free coffee. To Weihua Li, who worked diligently with me in accomplishing many challenging tasks throughout this research.

On the non-academic side, I am so thankful to many friends that I have met in the church life at Cambridge. Thank you for your continued prayers and support for our family. I also want to thank my parents who never fail to support and believe in me. I am greatly indebted to them for everything they gave me.

Finally, I want to thank my wife Ruth Peng for her love and constant support, for listening to complaints about my research with quiet patience, and for keeping me sane over the last few months. Thank you for reminding me to keep the faith in God and being my best friend in each and every moment of my life. To my two daughters, Stella and Kelly, your smiles are the fuel for my determination to finish the program.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Over the last decade, mobile phones have evolved from portable communication devices to information hubs that are deeply integrated into almost all aspects of daily life. With unprecedented amounts of sensitive personal data collected and processed on mobile phones, mobile apps can pose great risks to personal privacy. While people may seemingly grant informed permission for mobile apps to access their personal data, they nevertheless express concern when they later learn what information the apps have in fact collected. The disparity between peoples expressed privacy concerns and existing mechanisms for granting permissions, reflects the unsatisfactory situation in mobile application privacy.

One approach to improve mobile privacy is through user-preference mechanisms that let people decide what information to disclose and to which apps they are willing to disclose; the development of such mechanisms is the subject of much current work both by industry [16][35][63] and by the research community [39][62][26]. Yet, what really drives people's choices about disclosing personal information is still poorly understood. The research described in this thesis is aimed at improving our understanding of the factors that affect people's preferences about disclosing personal information that is collected when interacting with mobile applications.

I found that there are two major types of factors that affect people's willingness to disclose their personal information: 1) *user context*, which people use to assess the *sen-*

sitivity of the data to be disclosed, and 2) external factors such as the *purpose* for which information is requested and *trust* in the app requesting the information, which people use to assess the *appropriateness* of any information disclosure. Privacy decisions are jointly affected by these two factors and they vary across individuals. Given that people are typically uninformed about the purpose of data collection, it is important to raise the privacy awareness even before soliciting their preferences for information disclosure, so that their privacy expectations can be aligned.

Furthermore, people's answers to questions about their privacy preferences are influenced by how these questions are posed and in what situations these questions are posed. As a consequence, survey-based approaches that ask people to reflect on hypothetical situations, do not suffice for understanding privacy awareness in real experiences of using mobile apps. To help address this issue, I have implemented a framework, *ContextProbe*, that allows researchers to solicit privacy preferences *in situ*, by making it more convenient to build individual mobile apps tailored to particular investigations.

Much of this thesis can be viewed as exploring the role of context in determining people's attitudes toward disclosing personal information. Chapter 2 positions this research within the larger landscape of research on context and *contextual integrity*. This thesis adopts a limited notion of context, one that is tailored to mobile apps and information that can be sensed by mobile phones, but the approach to context is informed by a broader perspective.

Chapter 3 describes two studies that form a prelude to this thesis research. These studies highlight the challenges in probing people's privacy preferences. In the Privacy Attitude study [78], the results showed that presenting information about context is significant for people to determine the sensitivity of disclosed data, and thus affects their willingness to disclose data to mobile apps. In the Intrusiveness Study[1] [97], by monitoring apps' behavior in different usage contexts (when people are idle versus active), I present a metric for quantifying the "intrusiveness" of mobile apps. The outcomes of both studies provide evidence that context, an often overlooked component in privacy tools, can assist consumers by raising their privacy awareness and helping them to adjust their privacy expectations.

[1] This work was carried on in collaboration with Frances Zhang [96]

The observations from these preliminary studies revealed two critical requirements for soliciting people's preferences about disclosing personal information: 1) the content for probing privacy preferences should come from subjects' *real-world experience*, and 2) the questions about disclosure should be posed to subjects when they are still *in the situation* of interest, so as to avoid memory bias. Chapter 4 gives an overview of the ContextProbe framework and its support for the experience sampling methodology in probing privacy preferences. Chapter 5 gives a more detailed look at the architecture and implementation of ContextProbe, which is based on the App Inventor Integrated Development Environment [94] for creating mobile apps.

With ContextProbe, I conducted a study to investigate how interactions between user context and other external factors affect people's willingness to disclose their information to mobile apps. The results show that even though contextual factors are influential, they do not provide a complete explanation of privacy choices. More significantly, I found that the external factors *purposes of data use* and *trust in the app* itself outweigh contextual factors when considering information disclosure. My study showed that subjects were not aware of context in thinking about disclosure when purpose of data use was presented together with context. Another surprising result of the study is that people chose to *disclose more when purpose of data collection was omitted* than when a vague purpose was presented. Such a result is the precise opposite of what has been previously reported [84]. Taken together, the results are significant to privacy research in two ways: 1) soliciting people's privacy preferences *in situ* arguably leads to drastically different conclusions than a survey-based approach, and 2) the dominance of *purpose of data use* over *user context* suggests that people consider the appropriateness of information flow rather than the sensitivity of the information when disclosing information to mobile apps.

Chapter 7 presents the conclusion and future work of the thesis to discuss the next logical steps that are worth pursuing.

Overall, the core claims of my work can be summarized as:

> **With ContextProbe, privacy studies can benefit from capturing the contextually grounded reasons behind subjects' information disclosure. The purpose of data use plays a major factor in disclosure preferences, even**

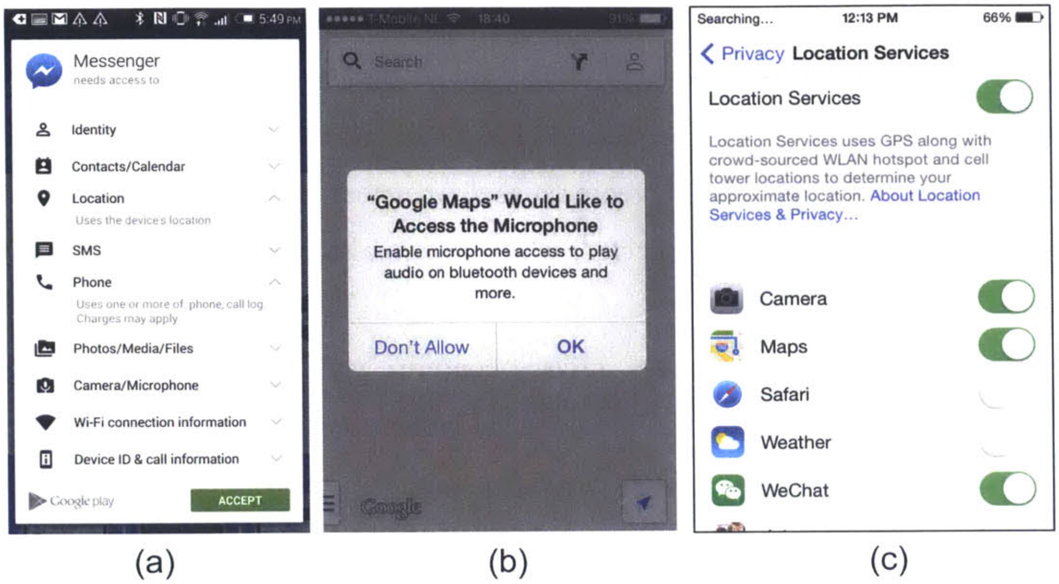
more than contextual factors. In addition, probing people's privacy preferences *in situ* is critical to accurately investigating privacy issues. The new approach enables observations of privacy preferences from both individual and systemic perspectives.

In the following sections, I will briefly describe the current state of mobile privacy with regards to information disclosure to apps and highlight the contributions of the thesis.

1.2 Information Disclosure on Mobile Apps

With pervasive use of smartphones and advances in sensor technology, context-aware services have become increasingly integrated into daily life. However, while extremely useful in serving and guiding us in our daily activities, smartphones can also silently collect data about us, allowing app companies to create digital dossiers for services like targeted advertisement, extrapolate behavioral patterns for market research, and engage in other activities. People today have to make decisions about disclosing personal information without being fully aware of the privacy implications of data collection [33]. At the same time, research has shown that most consumers regard their information created or stored on mobile data to be private and are strongly opposed to apps collecting information, such as when the apps track their locations without their consent [87].

While mobile phones store tremendous amounts of personal data, mobile apps that collect data inappropriately can pose great risks to people's privacy. Today, platform providers such as Android and iOS are given the role to protect consumers' privacy by preventing apps from scraping people's data without consent. These two major platforms adopt different approaches to protect user privacy regarding how apps can have access to user data. Android uses the all-or-nothing permission model to restrict access to system features and user data. Users have to agree to permissions requests before downloading and installing the app. During the installation process, Android shows a list of permissions required by the app (see Figure 1-1(a)). Apple's iOS platform takes a different approach in granting apps' permissions for data access on the device. Users of iOS apps do not make choices about granting permission during installation time. Instead, the permission is requested by



"Google Maps" Would Like to Access the Microphone

Enable microphone access to play audio on bluetooth devices and more.

Don't Allow OK

Location Services uses GPS along with crowd-sourced WLAN hotspot and cell tower locations to determine your approximate location. About Location Services & Privacy...

Figure 1-1: Snapshots of the permission request windows in Android and iOS (a) Android's permission request window (b) iOS permission request window (c) iOS privacy setting panel

the app when users take an action that needs the data (see Figure 1-1(b)). The blanket permission model by Android treats privacy control simply as reading a product's (the app's) terms of service to merely meet the *informed consent* requirements. Meanwhile, iOS views user privacy as an integral component of user experience and enables its users to control their privacy in situ. While both approaches aim to give consumers some control over their data, they still fall short of the definition of privacy as stated by Alan Westin [91], that privacy is the ability for individuals to control *when*, *how*, and *to what extent* information about themselves they disclose to others. Today, users' confusion and ignorance of how privacy controls operate on mobile phones give malwares the opportunity to secretly collect data and transmit them to other third parties. One famous example that happened recently is the flashlight application on Android called Brightest Flashlight Free, which collected and shared its users' location with advertisers without their permission [73].

Privacy researchers have proposed a variety of approaches to provide more awareness about what information can be silently collected [48][25]. However, consumers have no way to understand how their data is actually used. A recent report released by the White House addresses these concerns, focusing on the idea of disclosing the range of **possible**

uses for peoples' data [93]. This recommendation may seem to conflict with the current practice of privacy policy being written so broadly by companies that collect user data, given the opportunities of turning personal data into a lucrative business. Yet, research by Hurwitz [43] has shown that users are more satisfied when they are given a choice of which data and for which purpose it is collected, which in turn also increases their trust in the service.

In order to properly address users' privacy concerns while simultaneously allowing companies to use mobile data, the World Economic Forum recently published a report [95] including suggestions for data collectors and data consumers to honor people's privacy and adopt the concept of *Personal Data Store*, a new model for disclosing and consuming personal data. The report emphasized that people should have control of their data in terms of how services can use it, rather than simply granting carte blanche access to data consumers. Instead, a model of "selective disclosure" should be adopted which changes and reverses the process of how personal data can be used by other organizations. The new process allows individuals to specify what information they are willing to disclose to third parties, and for what purposes. A number of research projects [61] [26] [64] as well as startup companies proposed a similar approach that aims to simultaneously improve user privacy and the utilization of personal data. The success of such efforts would rely on the design of a usable privacy control mechanism and a well-incentivized data market. However, people's behavior and their preference for information disclosure on mobile devices in a ubiquitous environment is still poorly understood. To fill this gap, this thesis is mainly motivated by the following research question:

> **What are important factors that affect people's preferences for information disclosure to mobile apps ?**

Researchers have long recognized that people's privacy preference for information disclosure is highly malleable and dependent on various factors. Research by Bansal et al [11] in online health services showed that the **perceived sensitivity** of subjects' health data affects the willingness to disclose information and also affects their trust in health web sites. Also, this sensitivity often changes dynamically according to the individual's **context**, such as location and activities. For example, in a study of subjects' behavior of using a photo-sharing web site, Shane et al. [6] found that subjects had context-derived patterns in making privacy decisions for attaching geo-tags to photos. Some subjects would not attach the geo-tags to photos that were taken at locations they deemed more private or sensitive. These examples suggest that the capability to **capture user context** at the time of information disclosure is critical to explaining what is observed about users' behavior. In Chapter 3, I explore the role played by context in people's attitudes about disclosing information to mobile applications. Findings in Chapter 3 reveal the lack of support in current mobile platforms to utilize context to address privacy issues revolving around data collection.

Many researchers have conducted surveys to explore consumers' attitudes and concerns about privacy while using mobile apps. However, research [3] also found that people act quite differently from their expressed concerns. One major reason for this disparity is because that the survey-based approach often lacks the reference to people's real experience. Surveys can hardly represent users' experiences of considering the trade-offs between privacy and other benefits when disclosing information. Further, the lack of awareness of ongoing data collection by different apps also leads to incorrect privacy expectations when answering privacy questions. The capability to **sample from people's daily experiences** and uncover the real privacy concerns for information disclosure would allow researchers to understand more of the "contextually-grounded" reasons behind people's behavior. Recently, some researchers [58][2][47] have adopted the Experience Sample Method [53] (ESM) approach to collect subjects' *in situ* responses and explore different factors that affect subjects' behavior in disclosing information while in the experience of using mobile apps.

The two capabilities, preserving context and asking in situ questions, could potentially improve the quality of subjects' responses for their privacy preferences. However, the inherent complexities in conducting research in the ubiquitous environment often require much effort to implement the study app and set up the experiment. In Chapters 4 and 5, I implemented a framework, ContextProbe, which enables experimenters to easily create mobile apps tailored to their privacy studies. Apps built with ContextProbe can capture user context using on-board sensors on mobile devices and trigger questions based on the detected context.

1.3 Contribution

In short, this thesis makes the following contributions to research in information privacy:

- Showed evidence that user context is a significant factor affecting people's perceived sensitivity in disclosing information to mobile apps.
- Demonstrated how subjects' app usage context can be used to identify context-deviated behavior of mobile apps and quantify the intrusiveness of apps.
- Showed the necessity and importance of soliciting people's privacy preferences *in situ*.
- Introduced a framework for experimenters to easily create mobile apps to conduct privacy studies using the Experience Sampling Method.
- Showed evidence that purpose of data use and trust toward the app outweigh contextual factors when considering information disclosure.
- Discovered that the mere appearance of a vague purpose significantly decreased people's willingness to disclose information to mobile apps when compared to the omission of purpose.
- Showed evidence that people become less willing to disclose personal information when more specific information is provided.

These contributions should improve the current approach of exploring people's preferences for information disclosure on mobile phones. Specifically, the framework allows experimenters to gain more insights from both individual and systemic perspectives, because of the capabilities to investigate privacy in context.

In the next chapter, I will survey the existing privacy frameworks that emphasize the importance of context when investigating privacy issues, followed by a review of the relevant literature to my thesis.

Chapter 2

Probing Privacy in Context

This thesis is motivated by the need to better understand peoples preferences about disclosing personal information to mobile applications. More specifically, I hope to make it easier for experimenters employ tools for probing disclosure preferences, these tools themselves implemented as mobile apps. At first blush, one might think that privacy preferences are purely a matter of what information would be disclosed. But that view is too simplistic. A host of additional factors come into play in determining privacy preferences, including the circumstances under which information is requested, who is request the information and for what purpose, how the information will be used, and other concerns not understood. We loosely refer to these additional factors as *context*.

The chapter begins by reviewing two conceptual frameworks for exploring privacy in the light of concerns for context: Altman's **Boundary Regulation** [7] and Nissenbaum's **Contextual Integrity**. I also discuss the framework of expectation of privacy, which plays a fundamental role in United States Fourth Amendment jurisprudence. I then present the (simplified) notion of context used in this thesis, namely, that context is a *representation of people's physical surroundings with the subjective interpretations they attach to it.*

Next, the chapter surveys recent approaches to understanding people's behavior regarding information disclosure on mobile platforms. Specifically, it highlights the challenges researchers face in capturing the "contextually-grounded" reasons behind information disclosure through traditional investigative methods that employ surveys, and it describes an alternative *experience sampling* methodology. These considerations motivate the building of a framework for soliciting people's privacy preferences through experience sampling, as will be described in chapters 4 and 5

2.1 Privacy Theories Related to Context

This section reviews two frameworks for exploring privacy, namely, **Boundary Regulation** [7] from Irwin Altman and **Contextual Integrity** [66] from Helen Neissenbum. The former explains individuals' privacy management as controlling the boundary of "personal space" in physical contexts; the later provides a systemic view for describing privacy violations in terms of information flow within a system. Specifically, these two frameworks provide complementary perspectives for interpreting people's privacy expectations and peoples behavior with regard to information disclosure on mobile devices. In addition, the section discusses the "**reasonable expectation of privacy**", a framework used in the American courts to address privacy rights protected by the Fourth Amendment.

2.1.1 Boundary Regulation: End-user Perspective

Irwin Altman is a well-known social psychologist whose research centers on the theory of "social interactions". Viewing privacy through the lens of social interactions, Altman treats privacy mechanisms as the ability for the self to regulate contact with others when so desired. He took this position to emphasize that privacy is important to personal autonomy, especially for individuals' function in society and psychological well-being.

Altman's theory of boundary regulation emphasizes that privacy management is intrinsically a dynamic process of adjusting the boundary of "personal space" through interactions with subjects' social worlds and environments. According to Altman, people change their degree of "openness" to allow others to have access to them in different circumstances and to change the level of privacy:

> *"Privacy is conceived of as an interpersonal boundary process by which a person or group regulates interaction with others. By altering the degree of openness of the self to others, a hypothetical personal boundary is more or*

> *less receptive to social interaction with others. Privacy is, therefore, a dynamic process involving selective control over a self-boundary, either by an individual or by a group."* [7, p. 6]

For example, this thesis was written mostly by closing myself off from others (social worlds) and finding an quiet place on campus to avoid unnecessary social interactions (environments). In Altman's theory, loss of privacy is the unwanted access by others to the self, which breaches the boundary of personal space. The theory of boundary regulation predates the existence of the Internet and mobile phones, and it originally only addressed non-technological situations. However, Altman's notion of **personal space** provides an important foundation for discussing people's perceptions of privacy, which is applicable today even with recent advances in technology.

Palen and Dourish [67] applied Altman's seminal work to discuss the privacy issues that emerge when people interact with technologies in the networked world. They specified three boundaries that are important to individuals' privacy management: 1) the disclosure boundary, 2) the identity boundary, and 3) the temporal boundary. The disclosure boundary is concerned with revealing information to or concealing information from others to maintain both personal and public personals . The identity boundary conveys the role that a person may consider playing when making decisions about information disclosure. The temporal boundary specifies how the expectation of unintended use of information in the "future" would affect "current" disclosure. For example, students will clean up their online presence ahead of applying for jobs, and stop posting inappropriate content. Palen and Dourish argued that privacy management in any information system does not happen through a set of fixed rules but rather through a continuous process of managing boundaries. Furthermore, these interrelated boundaries move dynamically when **context** changes. For example, decisions about disclosing one's location when traveling for job interviews (interview context) would very different than any normal day (work context).

In the mobile environment, Mancini et al. [58] discovered that social functions particular to a location and activities conducted in that location are factors used by subjects to determine boundaries for regulating information disclosure. Specifically, these **subjective and social-cultural meanings** for the physical environment are what constitute the **context** of the disclosed information that affects people's decisions. In addition, Mancini noted that context is an *emergent* entity, which is dynamically established by subjects' interactions with others in their physical environment. The implications of this **emergent context** are that 1) privacy management cannot be accomplished merely by static enforcement of rules and 2) studies of privacy preferences must take subjects' contexts into account.

A number of privacy researchers in the HCI field use this notion of context, as a subjective interpretation of personal surroundings from a social-cultural perspective, to investigate people's privacy preferences. Among them, Khalil and Connelly [50] studied subjects' preferences for disclosing their contextual information in applications of context-aware telephony. They found that the *type* of disclosed information and the *relationship* with the inquirers are the two main factors that affect subjects' willingness to disclose. This indicates that the level of perceived sensitivity depends not only on the type of the disclosed information but also on the context.

Privacy researchers use various tools to capture environmental settings and subjective inputs for describing subjects' context, as I will discuss in Section 2.2.

2.1.2 Contextual Integrity: Systemic Perspective

Nissenbaum's theory of contextual integrity is based on the observation that in human society "social norms, or rules, govern the flow of personal information in distinct social contexts". Nissenbaum defines context as "structured social settings characterized by canonical activities, roles, relationships, power structures, norms (or rules), and internal values (goals, ends, purposes)" [66, p. 141]. For example, in a work context, sharing personal matters with colleagues would not be considered appropriate, because the governing norm is to share only information related to business.

Nissenbaum argues that the traditional "public-private" dichotomy fails to consider context when used to reason people's privacy expectations. Private-public dichotomy views making a secret or private matter public as the loss of privacy. However, in real life, privacy violations sometimes happen not at the moment of the disclosing of private information but when information flows between contexts. For example, a patient might disclose conditions of an illness to the physician during an appointment. But it would be awkward or inappropriate for the physician to ask patients about their conditions when they meet by chance in a public place and friends are around. Patients expect that the information stays "open" and "accessible" only within the healthcare context but not in other social contexts. According to Neissenbaum, privacy is about maintaining the contextual norms when information flow occurs. The medical example is a privacy violation not because secrecy is disclosed, but because it disregards the contextual norm when disclosing/accessing the information.

Privacy researchers use the framework of contextual integrity to examine privacy issues that arise when personal information is collected, processed, and disseminated on technology platforms in different domains. As noted by Neissenbaum [66, p. 141], applications of the framework need to first identify the **context-related information norms**, which are characterized by four main elements: contexts, actors, attributes, and transmission principles. *Actors* can be the subjects, senders, and receivers of the information. *Attributes* describe the nature of the information, while *transmission principles* serve as "rules" that restrict how information can flow to different actors. Barth et al. [13] applied contextual integrity to the medical domain by formalizing transmission of personal information using a logic framework. The logic framework can be used to express and reason about the constraints of information flow between entities with different roles as those described in regulations such as HIPAA [19] and COPPA [30]. Lipford et al. [57] explored different social contexts represented by information on social networking platforms and increased users' privacy awareness by making the flow of information more visible to the users.

For privacy issues on mobile platforms, Shklovski et al. [80] studied subjects' privacy perceptions and concerns when using mobile apps. The study used contextual integrity to query subjects about the types of data that they expect apps would collect in the context. The results revealed a significant gap between subjects' mental models of what is considered to be appropriate data access versus what apps actually request. Zhang and Shih [97], implemented a framework, as will be described in chapter 3, to reveal how apps' data access behavior deviates from the normal "app usage" context. For example, an app can still access subjects' location information even when subjects are not interacting with it. According to contextual integrity theory, this behavior violates the information norm of the app usage context: an app does not need consumers' data or does not collect data when it is not used. Using this framework, Zhang and Shih quantify the intrusiveness of famous apps by measuring apps' out-of-context access to user data on smartphones. More details about this study will be discussed in Chapter 3.2.

2.1.3 Reasonable Expectation of Privacy: Protecting Fundamental Rights

Another important core concept related to context is the **reasonable expectation of privacy** approach that is commonly used in American courts to address privacy rights protected by the Fourth Amendment. This approach was first mentioned in Justice Mashall Harlans concurrence in *Katz v. United States*[1], which highlighted the questions of whether "a person exhibited an actual expectation of privacy" and whether police eavesdropping on the suspect's conversations is subject to the "unreasonable search" clause in the Fourth Amendment. The ruling of the Supreme Court was in favor of Katz, and stated that the conversations in a public phone booth are protected by the reasonable expectation of privacy, and wiretapping is an intrusion on the individual's privacy interests. More importantly, in the concurrence, Justice Harlan described and formulated a test which was commonly cited by later court cases for determining whether a subject has a reasonable expectation of privacy: 1) whether "a person has exhibited an actual (subjective) expectation of privacy" and 2) whether "the expectation is one that society is prepared to recognized as reasonable".

The *Katz* test of reasonable expectation leaves unanswered questions about what is the socially recognized reasonable expectation of privacy. Privacy scholars acknowledge that the binary definitions of "private" and "public" do not adequately consider context when interpreting whether an appropriate "search" has occurred. For example, in the case *United States v. Jones*[2], the police tracked the suspect's vehicle on public streets with an installed GPS device. The Supreme Court ruled that such unwarranted monitoring is a violation to Jones' Fourth Amendment's rights, even though the information can be viewed as public.

Given the characteristics of emerging technologies, identifying what Neissenbaum referred as the "entrenched norm" becomes significant in attaining a more accurate descrip-

[1] 389 U.S. 347 (1967)

[2] 132 S. Ct. 945, 565 U.S. ___ (2012)

tion of the reasonable expectation of privacy. In the recent *Riley v. California*[3] case, the Supreme Court held that the warrantless search of cell phone data is unconstitutional, even when the search occurred during an arrest. The Supreme Court ruling recognized that a great amount of sensitive data about a person is stored and accessible on the smartphone.

> *[...] An Internet search and browsing history, for example, can be found on an Internet-enabled phone and could reveal an individual's private interests or concerns — perhaps a search for certain symptoms of disease, coupled with frequent visits to WebMD. Data on a cell phone can also reveal where a person has been. Historic location information is a standard feature on many smart phones and can reconstruct someone's specific movements down to the minute, not only around town but also within a particular building.*

Further, it also acknowledged that the use of smartphone apps is now deeply integrated into Americans' personal and social lives. These apps collect massive amounts of personal data, revealing people's social status, political preferences, health conditions, and many more sensitive pieces of information.

> *Mobile application software on a cell phone, or "apps", offer a range of tools for managing detailed information about all aspects of a person's life. There are apps for Democratic Party news and Republican Party news; apps for alcohol, drug, and gambling addictions; apps for sharing prayer requests; apps for tracking pregnancy symptoms; apps for planning a budget; [...] There are popular apps for buying or selling just about anything, and the records of such transactions may be accessible on the phone indefinitely.*

Debates revolving around reasonable expectation of privacy in the Supreme Court's cases are significant in several aspects. First, they acknowledge that individuals' privacy expectations do not change simply because they are in public. Second, the defense of "privacy in public" gives prominence to context and requires understanding of the information norm in it, especially in the ubiquitous technology driven world today. In other domains,

[3] SCt Slip Op No. 13132, 2014

the same concept can be applied for understanding individuals' privacy interests. For example, danah boyd studied teenagers' privacy attitudes and practices [15] on social networks and found that youths have strong privacy concerns even when they share information publicly on Twitter or Facebook. Another example is the concern for apps that run on Google Glass, a wearable head set with an optical display, to apply facial recognition technology and identify strangers in public [74].

2.1.4 Notion of context used in this thesis: an approximation with objective data plus subjective input *in situ*

The reviews of the three privacy frameworks above lead to one common conclusion: understanding of context (personal context or system norms) helps to clarify the seemingly obscure or unsatisfactory discussions of privacy interests and conflicts. Altman's boundary regulation was adopted by HCI researchers to discuss privacy management as managing personal space (boundaries), while the definition of a personal space depends on the social context of the people involved. Neissenbum's contextual integrity redefines privacy violations as the flows of personal information that violates the entrenched norm within and across contexts of an information system, refuting the public-private dichotomy in viewing privacy issues. Debates of "reasonable expectation of privacy" in Supreme Court's cases demonstrate that identifying context is critical for analysis of harm to privacy or privacy interests.

How about the analysis of people's privacy concerns for data disclosure while using mobile apps? Within the previous frameworks of privacy, I gradually came to speculate, as did Barkhuus' critics on the "mismeasurement" of privacy [12], that understanding people's real privacy concerns for using mobile apps would depend on identifying contexts where violations of information norms occur, and that presenting these contexts in privacy inquiries are critical to raise privacy awareness even before soliciting people's disclosure preferences. Specifically, Barkhuus argues that observations of information flows and entrench-norms within the system are overlooked in many investigations of privacy preferences. Privacy investigation should happen within a framework of *why* people disclose rather than *what* they disclose. The same principle is important and applicable to other topics in the broader context of privacy research. For example, information accountability framework [89] emphasizes the capability to monitor information use within a system and redress when violations occur (the *why*), rather than strict enforcement of access control (the *what*).

This line of thinking motivates further explorations of the roles and functions of context in the following chapters of the thesis. In Chapter 3, we will see two studies that illustrate the significance of capturing and presenting useful contexts for shaping people's privacy expectations in using mobile apps. Chapter 6 shows how presenting contextual information in privacy inquiries and asking questions *in situ* can affect people's willingness to disclose personal information.

The three frameworks provide the basis on which I develop my notion of context for exploring people's privacy preferences in information disclosure. Following Altman's concept of personal space, we can view mobile phones as an extension of self in the digital world. Therefore, discussions of privacy management on mobile phones must include the "subjective" part of how individuals view whether the disclosure of information is acceptable or not. On the other hand, we should not forget to satisfy the "objective prong" for analyzing privacy conflicts, with the view of Neissenbum's contextual integrity. Privacy studies should identify significant information flows within the information system, present them to users, and inquire subjects' responses while they are in the situation. In Chapter 3, I present two studies that illustrate uses of contextual information to probe subjective privacy attitudes and reveal significant information flows, respectively.

For the scope of this thesis, context is the **representation of people's physical surroundings with the subjective interpretations they attach to it**. Specifically, I built a framework that uses on-board sensors and other personal information on mobile phones to capture a degenerate case of *user context*. For example, with accelerometers, the GPS sensor, and the Bluetooth sensor, one can detect that someone is active at a certain location and surrounded by a number of people. With minimum input from study subjects, one can further determine the social-cultural functions of a place or activities carried in this place. For example, study subjects can provide annotations like "working alone" or "in a meeting" for the captured sensor data. In addition, sensing capabilities of mobile phones can be used to trigger privacy inquiries based on the detected context and prompt subjects for *in situ* responses. Chapters 4 and 5 describe the framework that allows experimenters to study individuals' preferences for information disclosure in various situations, using the above definition of context.

2.2 Research in User Behavior and Preferences for Information Disclosure

Given the above discussion of frameworks for addressing privacy and context as background, this section describes challenges to understanding people's behavior regarding information disclosure while using mobile apps. It reviews the recent approaches by privacy researchers in exploring information disclosures preferences. And it describes how privacy researchers try to use context as a crucial element in soliciting and explaining these preferences.

In general, apps require the disclosure of personal data to provide relevant services. Past research in consumer privacy shows that people are willing to disclose their personal information in exchange for customized services [20]. For example, consumers will turn on the GPS sensor on mobile phones and disclose their locations in order to receive recommendations for nearby restaurants. However, consumers also have hidden privacy concerns about being tracked by apps. Acquisti and Grossklags [3] highlighted that one main challenge in privacy decision making for individuals is the *incompleteness of information*. When individuals make privacy decisions, they often lack sufficient information to consider all possible factors. Therefore, as noted by Aquisti, consumers often trade long-term privacy for short-term benefits. Meanwhile, study by Spiekermann et al. [82] showed that even though consumers had stated strong privacy concerns, they still disclosed lots of personal information to shopping bots during the study. This disparity between consumers' privacy attitudes and actual behaviors make it harder for researchers to model what are consumers' real privacy perceptions.

2.2.1 Survey-based Approach

Many privacy studies use surveys to solicit people's preferences regarding information disclosure. Because privacy concerns are often subjective and experiential, approaches using surveys need to help study participants to recall their privacy experiences when answering the questions. In addition, there are various factors that can affect people's behavior in disclosing information. These factors include the degree to which the app is trusted, the sensitivity of the disclosed information, the perceived benefits to the user, and the appropriateness of disclosure as influenced by the context. Experimenters who conduct the study have to consider the interactions between these factors in order to correctly interpret the results.

Shilton et al. [79] conducted a context-based survey with 979 subjects to rate over 39,000 hypothetical vignettes to investigate people's privacy expectations about using mobile apps. Shilton's study tested different contextual factors including *who* (the data collector), *what* (type of disclosed information), *why* (application purpose), and *how* (use of data by data collector). For each participant, a series of questions was created with variations of the above factors. The questions explored different application contexts represented by different vignettes. The results of the study showed that scenarios of data harvesting by apps in different contexts often do not conform to users' privacy expectations.

Knijnenburg and Kobsa [51] explored whether justifications for information requested by a recommendation system affect subjects' willingness to disclose demographic and contextual information. Their study used a mockup interface to simulate an app running on a mobile phone. The results showed that for different genders, different strategies (order of data disclosure requests and justification types) could be used to make participants disclose more. For example, for female with high disclosure tendency, the best strategy is to show data requests for contextual information first than requests for demographic information, and should provide no justification. In contrast, a recent study by Tan et al. [84] found that purpose strings shown in the permission dialog on iOS increase the willingness of subjects' to disclose information. Tan's study also used mockups to simulate what subjects would see as requests of data access when using apps. Both studies recruited several hundreds of participants and let each participant make their privacy choice *only once*. The one-shot response from the survey-based approach is questionably representative of subjects' privacy preferences when making privacy choices in the real-world environment.

To observe subjects' information disclosure behavior, privacy researchers have implemented apps tailored to capture factors of interest, then conducted longitudinal studies by asking subjects to install and use the apps. Hurwitz [43] created an app for grocery shopping to investigate consumers' acceptance of mobile services that collect personal data. Subjects were asked to install the app and choose between different moneysaving programs. The savings program that collected more personal data would give subjects more discounts on grocery items. The study results showed that subjects who chose to disclose more were more satisfied with the mobile service.

The above examples illustrate some of the challenges in conducting experiments in privacy research: 1) the need to use appropriate methodologies to study privacy issues that arise from subjects' experience using apps, and 2) the need to preserve ecological validity so that the study approximates real-world experiences. For example, studies of location sharing behavior in social mobile apps would need an approach to record subjects' actions as well as the disclosure context. The goal is to capture subjects' realistic decisions with perceived real risks and benefits.

2.2.2 Experience Sampling Method

In this section, I will survey recent studies that used the approach of Experience Sampling Method (ESM) [53] or its variants such as the Daily Reconstruction Method [47] and the refined ESM [21] (rESM) to achieve that goal. The Experience Sampling Method is an approach to soliciting people's momentary responses about their feelings, preferences, or emotions throughout the day. Researchers use ESM to study subjects' "inner states" that can be affected by their surrounding environment. For example, research in psychology has used ESM to study factors that cause stress in the working environment [38]. The in-situ responses of ESM eliminate the cognitive bias that is introduced in the recall-based approach such as surveys and interviews. In the Human-Computer Interaction field, Consolvo and

Walker were among the first few researchers to apply ESM in studying user experiences of ubiquitous computing applications [24]. They view ESM as a "summative technique" for investigating the impact of ubiquitous computing in people's daily life.

The ESM approach allows privacy researchers to better understand the reasons behind subjects' behavior in disclosing information. For example, Mancini et al. [58] used ESM to explore how subjects attach "socio-cultural" meanings to a space when sharing information on a social networking platform (see Section 2.1.1). Abdesslem et al. [2] applied ESM on mobile devices to explore how contextual factors (location types, time, and data recipients) would affect subjects' willingness to disclose their current location. Carrascal et al. [17] studied how people evaluate personal information that is captured when browsing the Internet. The study presented specific questions to subjects based on their current "browsing context", i.e. the type of website they were visiting, to probe subjects' privacy expectations and evaluation of personal information. A recent study by Staiano et al. [83] explored how subjects value personal data that is captured while interacting with mobile apps. The study used the Daily Reconstruction Method [47], a hybrid approach of ESM and Diary method, to generate personalized surveys based on data collected during the previous day. For example, the survey question would first show information like "you were on the phone for seven minutes today" and ask subjects to place bids on how much they think the data is worth. The study results showed that subjects valued location data the most and valued their media usage data the least.

The above ESM studies highlight the importance of using mobile phones to investigate subjects' privacy experiences while using their apps. Using the onboard sensors of today's mobile phones, researchers can capture different contexts that are represented by sensor data. With these capabilities of mobile phones, Cherubini et al. [21] proposed to "refine" the ESM approach with user context that can be modeled and detected by sensors. In addition, the context can be logged automatically with less intervention from subjects and survey questions can be prompted when a specific context is detected.

Tools for Conducting ESM Studies

With the pervasiveness of mobile phones, most ESM studies now use participants' own devices to prompt questions and collect responses. Yet the time and effort spent in creating an ESM app can be a heavy burden for experimenters. To help experimenters create customized apps for ESM studies, several open source projects have introduced toolkits or platforms to simplify the tasks in conducting studies on mobile devices.

MyExperience [36] is a software tool running on Windows phones that allows experimenters to capture objective sensor data and self-reported survey data. Experimenters can define different triggers to activate predefined actions based on sensor readings on the phone. For example, a trigger could be configured to fire survey questions when the detected GPS location is within a certain region. The "Sensor-Trigger-Action" architecture allows experimenters to solicit subjects' *in situ* responses in specific contexts as represented by sensor data. The framework provides an XML interface to configure sensors and triggers without writing code. This allows researchers with limited knowledge of application development to customize MyExperience and tailor it to their studies. My thesis is motivated by the similar concept of removing the overheads in creating mobile apps to allow experimenters to focus on the design of their ESM studies.

Ramanathan et al. [72] built a sophisticated framework, *ohmage*, which provides a stack of tools for recording, analyzing, and visualizing personal data from self-reported user input, as well as from data streams collected from mobile phone sensors. The framework provides modularized components both on the phone and the server for conducting data collection tasks and ESM studies. Each app can be configured via a scriptable data schema that defines its data collection tasks (what types of data to collect and how often), as well as its corresponding visualization on the backend server. Another toolkit called Open Data Kit [8] (ODK) was built to help organizations with limited technical resources to conduct data collection tasks in developing countries. Through ODK, field workers who are not professionals in programming can quickly create surveys on mobile phones and upload the data for aggregation or analysis. Along the same line, some projects like Funf [5] and EmotionSense [71] provide software libraries for experimenters to easily create apps for conducting social science research on mobile phones.

A framework described in Chapter 4 and Chapter 5 adapts similar concepts to the above tools, but further simplifies the tasks and extends the capabilities for conducting ESM studies on mobile phones. For example, the framework gives more flexibility in creating ESM apps and removes overheads of setting up the backend system. Overall, the goal is to exploit more useful "contexts" by leveraging the available sensor data and allow experimenters to test different factors related to the contextually grounded reasons for information disclosure.

Chapter 3

Explore Privacy Expectations in Context

This chapter describes two studies, the Attitude Study and the Intrusiveness Study, that explore the role played by context in people's attitudes about disclosing information to mobile applications. The primary results of these studies are that context is a major factor in people's disclosure preferences, but it is not the only factor. Thus, investigations of privacy preferences must take context into account, and we will pursue this approach in subsequent chapters.

Currently, the lack of transparency of data collection in mobile applications leaves consumers vulnerable to excessive intrusion from their mobile phones and exposes them to the potential risk of that personal data being used inappropriately. There has been a considerable amount of work done at attempting to minimize this privacy risk. They include techniques to improve transparency by tracking the flow of sensitive data [27] and identifying applications whose access privileges exceed the actual needs of the apps [32]. In addition, platform providers supply various methods for granting permissions that let users control the extent to which apps can access personal data.

Useful as these tools may be, their effectiveness as privacy controls for consumers is questionable, especially given today's ubiquitous computing environments. Research by Felt and others has found that a majority of mobile phone users pay no attention to the process of granting permissions to applications [34]. Many users either wrongly interpret what privacy control can actually afford them or do not trust that the applications would handle the data appropriately [22].

To help consumers make informed decisions with regards to their privacy concerns, we need to understand people's privacy expectations when they are using mobile apps. Also, we need to identify the important factors that trigger their privacy concerns when disclosing their personal information on mobile phones. Chapter 2 of this thesis adapted Helen Nissenbum's contextual integrity theory [66] to describe the inherent challenges in protecting privacy in context-aware applications. Contextual integrity is a conceptual framework for understanding privacy concerns that considers various elements of situations such as roles, expectations, actions, and practices within an information system. Although, contextual integrity provides a theoretical perspective for viewing privacy expectations, it doesn't address the technology components required to capture or model people's attitudes toward privacy. The two studies described in this chapter – the *Attitude Study* and the *Intrusiveness Study* – were designed to help bridge this gap by exploring the role of context in people's privacy preferences.

The Attitude Study [78] explored people's privacy attitudes when disclosing contextual information to mobile apps. The Intrusiveness Study [97] investigated the "context-deviation behavior" of mobile apps that violate the principle of "respect for context" [92] when accessing user data. The goals of both studies are to understand the interplay of personal contexts and other factors in determining subjects' privacy preferences. The outcomes of both studies provide evidence that context, an often overlooked component in current privacy tools, can assist consumers by raising their privacy awareness and helping them to adjust their privacy expectations.

The following sections present details of these two studies and discuss how they shed light on the design of system architecture for user-centric data collection and the implementation of a framework to help researchers utilize context for exploring subjects' privacy preferences.

3.1 Attitude Study: Exploring Privacy Preferences for Disclosing Data to Mobile Applications

The Attitude Study explored people's willingness to disclose their personal data, especially contextual information collected on smartphones, to different apps for specific purposes. The goal was to identify the factors that affect people's privacy preferences. For example, study of location-sharing apps shows that user preferences vary depending on the recipients and the context (e.g. place and time). However, previous studies that used surveys and interview methods [58, 18] have had limitations in capturing the real causes for people's privacy concerns [12]. In the Attitude study, I used a hybrid approach of the *experience sampling method* [53] and the *diary study* to solicit people's willingness for disclosing information in different contexts. Specifically, I looked at possible contextual factors such as location, time, and people's activities at the moment they are asked to disclose the data. Additionally, I tackled the following challenges when conducting the study:

1. How do we collect information that can sufficiently represent people's contexts throughout the day?

2. How do we effectively solicit people's preferences for information disclosure that are related to their contexts?

3. What are the possible and common confounding factors introduced by people other than their contexts?

A three-week user study is conducted with 38 participants to collect contextual information and self-reported data using smartphones. In parallel to that, the study also solicited people's preference for information disclosure using contextualized questions. The questions specify the developers of the app, their purposes for data collection, benefits of disclosing, and most importantly subjects' context. I used the responses to build a preference model for each participant that reflects his or her privacy concerns in different contexts. I applied the J48 implementation[1] of C4.5 algorithm [70], a decision tree algorithm, to generate rules

[1] http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/J48.html

that could intuitively represent most relevant contextual factors. For some participants, the resulting models showed strong correlations between their decisions about information disclosure and their context, whereas others made decisions that were strongly biased toward other external factors such as the data requestor or rewarded benefits.

3.1.1 Methodology

The study aims to accomplish the following two goals with the study approach: 1) to collect information that describes a participant's daily context and 2) to solicit people's answers that are as much contextually-bound as possible. The study lasted three weeks and was conducted in March and April of 2012. There are two tasks that the participants need to perform during the study. First, the participants were asked to install a program on their smartphone to collect sensor data, and respond to prompted questions for labeling their current location and activity. Next, the participants answered survey questions that were generated nightly and customized for each participant according to their daily contexts collected in the previous day. Each participant will receive a unique link to the personal survey everyday.

By the end of the study, qualified participants were called to join in-lab interviews. The interview provides more insights and details about the reasons of why participants disclosed or did not disclose their information.

Recruitment and Demographics

The study recruited 38 participants from one university campus through email-lists and flyers posted on bulletin boards. Twenty-eight participants were students (19 undergraduates and 9 graduate students) and ten were campus staff. Among all the participants, 18 were male and 20 were female. The participants were screened for their English proficiency and their use of the Android smartphone as their primary mobile device. About half of the participants lived outside of the campus; they had different lifestyle and composition of daily context (e.g. commuting between workplace and home) from those living on campus. Participants were compensated based on their level of participation in the study, including

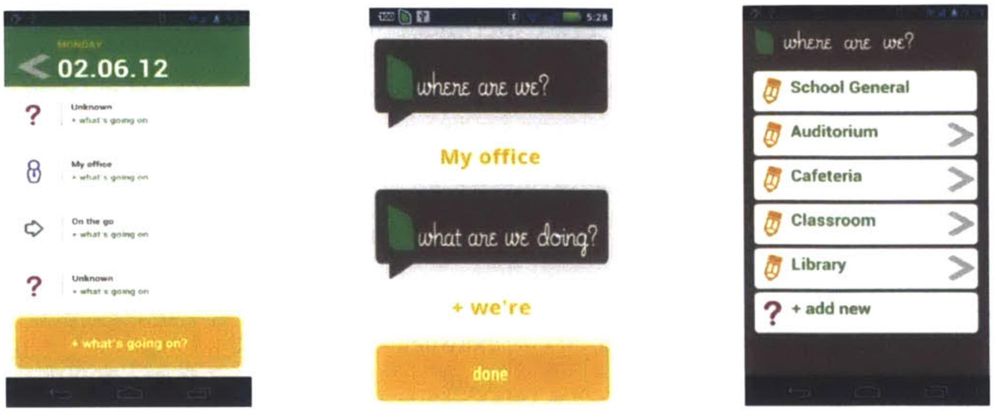
Table 3.1: Pre-experiment Survey

Q_1: How much time a day do you spend on using smartphone applications?	Q_2: How many Google services are you using currently?	Q_3: How many hours a day do you spend on Facebook?	Q_4: How often do you shop online on Amazon?
less than 30 minutes A_{11}: **15.7%** (6/38)	less than 3 A_{21}: **15.7%** (6/38)	less 0.5 hour A_{31}: **39.4%** (15/38)	seldom (e.g. only few times a year) A_{41}: **28.9%** (11/38)
between 30 minutes and 1 hour A_{12}: **21.1%** (8/38)	between 3 and 5 A_{22}: **34.2%** (13/38)	between 0.5 and 1 hour A_{32}: **31.5%** (12/38)	sometimes (e.g. about once a month) A_{42}: **31.5%** (12/38)
more than 1 hour A_{13}: **63.1%** (24/38)	more than 5 A_{23}: **50%** (19/38)	more than 1 hour A_{33}: **28.9%** (11/38)	very often (e.g. more than 3 times a month) A_{43}: **39.4%** (15/38)

hours of logging context data ($2.6 per day), numbers of survey questions answered ($2 per survey), and $10 for the final interview. An additional $2.6 was awarded to the participants for each week's completion of the two tasks. Besides the benefits, the participants needed to be compliant with the rules that ensure enough coverage of the self-reported data to correctly represent their daily context, or else they would not get their compensation for the day. The incentive structure was used to motivate the participants to contribute more data and stay in the study. Twenty-seven participants (14 undergrads, 7 graduate students, and 6 campus staffs) completed the full study and 11 of them joined the final interview.

Pre-experiment Survey

The participants were asked to fill a pre-experiment survey before the study to capture their familiarity with using smartphone apps and their experiences with major online web services (e.g. Google services such as Gmail, social networking sites like Facebook or online shopping sites like Amazon). Table 3.1 summarizes the questions and the statistics of the answers in the survey. The participants were also asked for their frequently visited local companies in three categories (e.g. banking, retail, and grocery stores) that were used later for generating personalized surveys. The survey results showed that more than half of the participants are heavy users of smartphone apps and Internet web services.



(a) Context history (b) Prompt for annotation (c) Options for annotation

Figure 3-1: Screenshots of the data logging application

Data Collection: Recording a History of Daily Contexts

The study used a hybrid approach of combining the experience sampling method with the diary method for acquiring in situ answers from the participants. The study was separated into two parts: 1) the *context recording* part using the experience sampling method and 2) the *experience reconstruction* part using the diary method of the study. The "context recording" part includes logging the contextual information as well as collecting annotations, tuples of location and activity, from the participants.

A data-logger program that was pre-installed in the smartphone would read various sensor data in the background to record contextual information such as location, time and proximity data (scanning of Bluetooth devices) of the participant. The data logger program, as shown in Figure 3-1, also detected frequently visited places and prompted the participants to provide annotations that they found meaningful to describe the moment when getting the prompt. For example, the participants received periodically a question like: *"Where are we? And what are we doing?"* They could answer the question by choosing a location and an activity label from a predefined list of choices or by creating new labels suitable for that situation. By doing so, the study was able to capture a history of contexts for different events that a participant encountered during the day.

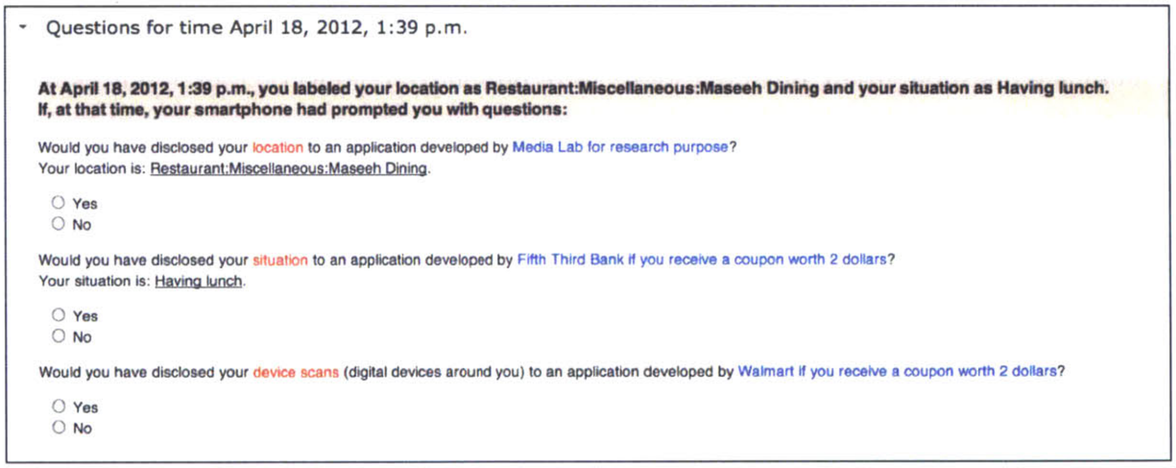
Diary Study

For the "experience reconstruction" part, a customized survey was sent to each participant every day with questions generated from the annotations of locations and activities each participant provided in the previous day. For example, if previously the participant entered "Messeeh Dining" as the location label and "Having Lunch" as the activity label, then the questions would be generated as shown in Figure 3-2. Each survey contained 4 to 10 question groups, depending on how many annotations the participant provided for that day. Although a participant might provide several annotations of their locations and activities within an hour, only one annotation was selected from that set. The study chose a one-hour window because people tend to regiment their life according to work-related schedule as described in previous research of life-logging applications [14].

For each question group, three questions were presented to collect the preferences for disclosing different contents: location data, situation data, and proximity data (Bluetooth scanning of the nearby devices). The question asked "Would you have disclosed..." to clearly indicate that the participants were asked to think about whether or not they would disclose the information. The contextual clues (time, location and activity labels) on top of each group help the participant recall the "context" when giving the answers. Participants were asked questions about their willingness to disclose the data to a particular entity with a specified purpose of data use.

The question simulated the situation of disclosing personal data to an application developed by a particular company or entity. For each question, the option for developer types was selected from three categories: ***academic entities***, ***companies***, and ***well-known large companies with web services*** with equal probability. In order to limit any bias that the participant might have for particular organizations, the question used multiple different organizations for each category of requestors. For the category academic entities, the question used *MIT*, *Media Lab* and *Harvard Medical*. For the category local companies, the question used *banking*, *retail store*, and *grocery store*. The specific grocery store, retailer, or banking company was customized for participants based on results of the pre-experiment survey, indicating which companies/stores they normally used/visited. I anticipated that this customization will make subjects' responses more representative of their actual disclosure preferences, since it brings the experiment closer in-line with their everyday life. Finally, the question used *Google*, *Amazon*, *Facebook* to represent well-know large companies with web services.

For each category of requestors, the question included the benefit or the purpose for collecting the information. For academic requestors, the survey questions told subjects that the data was being collected for research purposes. When the requestor was a company, participants were asked if they would disclose the information in return for a $2 coupon. Finally, for well-known large companies with web services, subjects were told that the purpose was for improving personal service. These purposes were used to communicate to subjects how the disclosed information would be useful for the requestor.



Questions for time April 18, 2012, 1:39 p.m.

At April 18, 2012, 1:39 p.m., you labeled your location as Restaurant:Miscellaneous:Maseeh Dining and your situation as Having lunch.
If, at that time, your smartphone had prompted you with questions:

Would you have disclosed your location to an application developed by Media Lab for research purpose?
Your location is: Restaurant:Miscellaneous:Maseeh Dining.

○ Yes
○ No

Would you have disclosed your situation to an application developed by Fifth Third Bank if you receive a coupon worth 2 dollars?
Your situation is: Having lunch.

○ Yes
○ No

Would you have disclosed your device scans (digital devices around you) to an application developed by Walmart if you receive a coupon worth 2 dollars?

○ Yes
○ No

Figure 3-2: Example of a personalized questionnaire based on "contextual information"

3.1.2 Results

The 27 participants who completed the study answered 4781 question groups (14343 questions) in total. The participants answered an average of 24 questions per day. The results of those participants that started but quit the study early were not taking into consideration. The overall participant rate of the study, counting those who finished both the data collection and diary survey, was 71%. This section reports the main findings of the study, including results from analyzing quantitative data collected from the study and results from

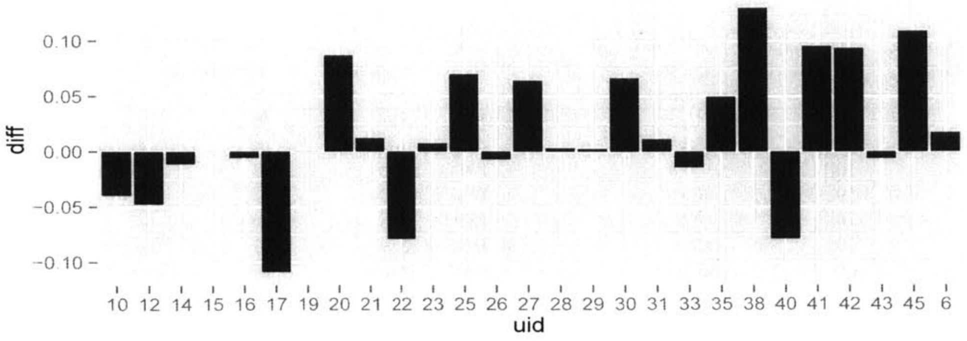


Figure 3-3: The percentage of *yes* responses for disclosing locations annotated as *home* vs. the percentage of *yes* responses specifically at time slots after 6pm or before 6am.

analyzing interview data. First, the section starts by describing the general outcome from the survey questions. Next, the section looks into the responses of each individual and how the results relate to contextual factors using outputs from the decision tree algorithm. Lastly, the section reports the results from the post interview to understand the privacy attitudes of participants that are often difficult to distill just from the quantitative data.

Type of data and context

The results showed that the participants were most likely to disclose activity data (62% yes), followed closely by location data (59% yes), but were less likely to disclose Bluetooth data (49% yes). The interview data revealed that the participants were more reluctant to disclose Bluetooth data due to the uncertainty of what information can be disclosed by Bluetooth data.

As for the general trend across individuals, study results showed that the preferences for disclosing information were dependent on participants' location at the time of disclosing. For instance, participants were most likely to disclose their locations when they were at places in the category *traveling* (79%), followed by *activities* (78%), *school* (65%), *work* (62%), *fun stuff* (58%), *on the go* (58%), *restaurant* (57%), *other* (54%), and lastly *home* (52%). The places that were deemed to be more private for personal activities such as *home* and the places in the *restaurant* category were shared less than public places such as *bus stops* in the *traveling* category or different classrooms in the *school* category. In

Table 3.2: Participant responses

User ID	Number of responses	Percentage of saying yes (%)	Affected by requestor type (R) or context (C)	User ID	Number of responses	Percentage of saying yes (%)	Affected by requestor type (R) or context (C)
P6	753	67	(C)	P10	819	38	(C)
P14	480	64	(R)	P12	1024	76	(C)
P15	363	100		P16	645	88	(C)
P20	555	59	(C)	P17	240	51	(C)
P22	522	76	(C)	P19	318	100	
P25	666	23	(R)	P21	579	10	(C)
P27	840	66	(R)	P26	438	79	(R)
P33	642	71	(C)	P29	585	35	(R)
P35	732	45	(R)	P30	771	36	(C)
P45	381	32	(R)	P31	210	69	(R)
P42	279	31	(C)	P38	675	78	(R)
P23	279	90		P41	333	36	(R)
P28	615	1		P43	390	99	
P40	210	76	(C)				

contrast, the results also showed that the difference in time did not significantly affect the participant's willingness to disclose location. For example, Figure 3-3 shows that there is only a small difference (10%) between the percentage of all *yes* responses for disclosing locations annotated as "home" and the *yes* responses for locations if the timestamps were after 6pm or before 6am. For example, participant 17 was 12% less likely to disclose his or her home locations if the time was before 6am or after 6pm. This might indicate that the participants considered only the type of locations when making privacy choices or failed to take the time factor into their considerations.

When considering the data requestor, the participants were most likely to disclose their data to academic entities (44%), followed by local companies (36%), and least likely to large companies with web services (20%). These results showed that subjects were more willing to disclose information to people who they were closer to – in this case local businesses as opposed to larger web services.

Individual preference model

I ran C4.5 decision tree algorithm and produced rules from each participant's responses. The results showed that about 81% (22/27) of the study participants have obvious patterns in their responses. Table 3.2 summarizes the results of participants' responses, and it shows that some participants are what Westin [90] called "privacy fundamentalist" in that they rejected most of the data requests in the survey questions, and some are "privacy uncon-

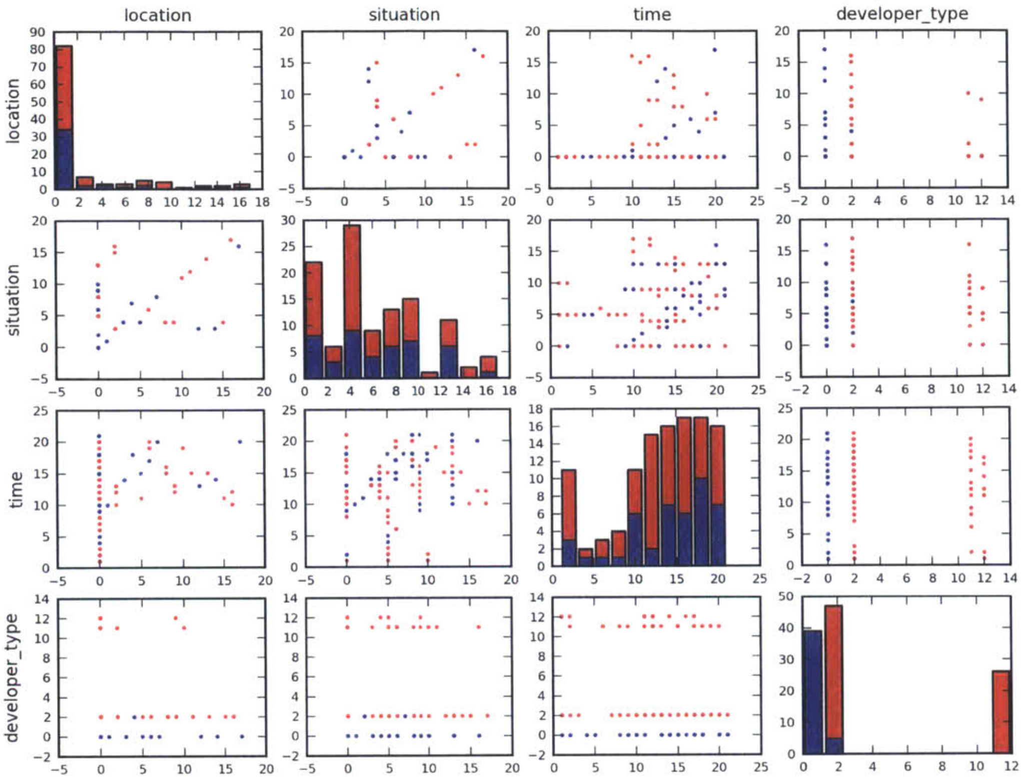


Figure 3-4: Responses (··· Deny; ··· Allow) showing the participant's privacy preferences are biased towards certain companies (*developer type*). This participant would always disclose information to *academic entities* but would not disclose to *companies* and *large companies with web services*

cerned", in that they accepted most of the requests. The participants marked with *(R)* in the table were those gave responses purely based on the type of data requesters. For example, some participants would always answer "yes" to disclose their data when the data requester represented in the survey question was of type *academic entities*. About 54% (12/22) of these participants (marked with *(R)* in Table 3.2) have decision rules that are related to contextual factors (location, time, and their activities), while the rest have decision rules related only to the data requesters. Figure 3-4 shows a scatter plot representing the results of a single participant's responses to disclosing their information in different contexts. Each box in the scatter plot represents one factor (location, situation/activity, time, and developer type) that affects the participant's responses. The x-axis represents indexes of locations, activities, hours, and developer types. Table 3.3 shows the user annotations for their loca-

Table 3.3: User Annotations of Locations and Situations

Location		Situation	
ID	User Label	ID	User Label
0	Home	0	Working alone
1	School:Classroom:PDL	1	In Lab
2	Restaurant:Maseeh Dining Hall	2	Walking
3	School:Classroom:10-250 Lecture Hall	3	Having lunch
4	Travel:Convention Center:MIT Energy Conference	4	In lecture
5	School:Classroom:Parker Lecture Hall	5	Sleeping
6	Other:Religious Center:Mosque	6	Praying
7	School:Classroom:Office Hours	7	Chatting informally
8	School:Classroom:TEAL Room	8	Studying Together
9	School:Classroom:Architecture Classroom	9	Chilling alone
10	Fun Stuff:Park:Kresge Lawn	10	Hanging out with friends
11	Fun Stuff:Stadium:Z-Center	11	Ultimate Frisbee
12	Restaurant:Black Seed	12	Playing Cricket
13	Fun Stuff:Mall:Gap	13	N/A
14	Restaurant:Student Center	14	Shopping
15	School:Classroom:Math Recitation	15	Having Breakfast
16	Fun Stuff:BC Porter Room	16	Having dinner
17	Restaurant:McCormick Dining Hall	17	Bhangra Practice

tions and situations. The id numbers for different locations and situations map to the index number of location axis and situation axis in the scatter plot respectively. For example, the top-left box in the scatter plot shows responses under location context "home" in which the participant responded with *yes* about 40% (33/82) of the total requests. The participant responded with *yes* when the *developer type* was of type academic entities (specified as index 0 in the *developer_type* box in the scatter plot) and *no* in the other categories. These results suggest that people have developed default policies based on other concepts such as trust of the companies rather than contextual information.

We found that the participants who incorporated contextual factors in their decisions have patterns based on: 1) location and time, 2) time and data requestors, and 3) location and data requesters. For example, P42 rejected all data request for location *Home* after midnight and before 6am. P17 would not disclose locations to data requestors from the category *grocery stores* between 12pm and 6pm. P17 later explained in the interview that she would not disclose work-related locations to a grocery store because locations from work are "unrelated" to understand shopping behavior. P29 would not disclose all locations labeled as *home* to requestors besides those from the category *academic entities*.

3.1.3 Post Interview

We invited the participants who completed the study to participate in a focus-group interview. Each interview was held in a conference room and lasted about 30 minutes with 3 participants attending. We asked questions concerning their reasons for rejecting or allowing the data requests, and details about the conditions (context) that triggered their privacy concerns. We first asked the participant to describe what were they thinking when they were answering the questions. Then we asked them to recall their rules, if any, for disclosing their information. We identified three characteristics of how some participants evaluate privacy risks based on their privacy expectations that are shaped by three factors: 1) subject interpretation of the context as being private or public, 2) sensitivity of the information disclosed, and 3) relations between the purpose of data collection and the context.

Mancini et al. discuss that people's information sharing behavior on social network is related to their interpretations of whether the context being private or public [58]. Yet, the post-interview data in this study showed that people have different interpretations of what is public and what is private. P30, for example, considered any location with "hanging out with friends" as its activity label a public context. On the contrary, P20 decided that all activities "hanging out with friends" are private. These two different views on the concept of "privateness" for a specific context resulted in two opposite rules in the decision tree algorithm. For example, the rule for describing P30's preference is:

YES -> location:ANY, activity:hanging-out-with-friends

while the rule for P20's preference is the opposite.

NO -> location:ANY, activity:hanging-out-with-friends

The post interview also revealed that failure in communicating what to disclose caused misjudgments on the sensitivity of the disclosed information. For example, several participants reported that they would not disclose Bluetooth data because they thought the term "device scans" in the questions means "all information on the smartphone". However, P33 and P38 who recognized this as Bluetooth technology would always disclose this information. Because, as they pointed out,

> P(33) *"I think device scans give information about the devices around me, and it is not personal."*

These remarks show that evaluating the sensitivity of information depends on participants' knowledge about the technology used in data collection. Lastly, participants tended to reject data requests if they failed to find "reasonable" connections between data collection and its possible purposes in a specific context. For example, P17 attempted to find reasons of why an app needs his workplace location:

> P(17) *"can't think of why an app needs my locations at work to figure out what I like to shop for food."*

Similarly, many participants said no to the companies with web services because they were unsure about how the disclosed information can be used by the data requestor.

Another interesting finding is how people developed their rules during the period of the study. Several participants reported that they started the study without obvious rules in mind, responding to the questions by just their instincts. But as the study continued, rules were introduced accumulatively through relevant contexts. For example, P22 become more privacy aware during the study and changed his preferences for disclosing his information in certain context.

> P(22) *"Before the study, I didn't think much about giving away my information. Then I realized that I would always say no when I am working in my office, so I* started saying no *at all places when I am working."*

3.1.4 Conclusion

The attitude study aimed to understand people's privacy attitudes in disclosing their information under different contexts. First, the study used an enhanced experience sampling method to record information that approximates an individual's daily contexts. The Experience Sampling Method program prompts subjects automatically for annotations of current location and activity.

Second, the study created a personalized survey in which each participant would answer questions with the help of contextual triggers. Participants would provide their answers of whether to disclose their data based on the given context in the survey questions while recalling the experience *in situ*. The decision tree algorithm C4.5 is then used to analyze the responses and generate a preference model for each participant.

The main findings of the study are:

1. The results of the study showed that although some participants' preference models are related to their contexts, not much could be gleaned about just how much contextual factors affect their decisions about data disclosure.

2. Both quantitative data from participants' survey responses and qualitative data from the post interview showed that other external factors such as types of the data requestors predominate over the contextual factors.

Some participants' responses showed that their answers for privacy preferences can be easily biased and overwritten by the existence of other confounding factors such as the subjective feeling about the data recipients (e.g. big companies such as Google vs. academic institutes such as MIT). In the post interview, participants remarked making privacy choices based on their intuitive judgments, such as company's brand, when they cannot understand the privacy implications behind the disclosure of their data.

The study also found that some participants had "discovered" their privacy preferences based on the context elements presented in the survey questions. In the post interview, some participants expressed that they became aware of how their decisions for disclosing information related to specific location (*home* vs. *workplace*) and time (*working hours* vs. *family times*). One participant revealed that he was surprised to learn that he was in a friend's house every weekend during the study and he didn't want to disclose that information to anyone.

The observations from the attitude study raised more questions to investigate further for understanding the effects of contexts on people's privacy preferences. Here is a list of them:

1. How to help subjects be aware of the activities for data collection happening on their phone so that they can adjust their privacy expectation?

2. How to solicit subjects' privacy preferences with respect to their contexts in a more systematic way for disclosing different types of information?

3. How to incorporate other factors such as for what purposes data was being collected to aid users evaluate the privacy implications for disclosing their data?

Participants' recollections for making sense of their own data and the process of preference forming suggest the need for a more systematic approach to assist subjects to discover their privacy preferences. This unmet requirement motivates the following research described in Chapter 6, specifically for mobile platforms.

3.2 Intrusiveness Study: Analyzing Apps' Behaviors in Privacy-relevant Contexts

The Intrusiveness Study[2] is motivated by the need for a more effective mechanism that will help users make informed decisions about choosing "privacy-friendly" apps. For consumers who are highly privacy aware, current platforms such as Android and iPhone don't provide adequate tools to help them select apps that are privacy-preserving and avoid those that are privacy-invasive. The existing transparency technologies such as Taindroid [27] proposed by Enck et al. mostly focus on tracking access to sensitive data and even the flow of privacy-sensitive data throughout a smartphone.

In measuring perceived privacy intrusions, we can go beyond just consideration of what data is collected by introducing the notion of *intrusiveness*. The study characterizes intrusiveness as the "out-of-context" actions of gathering personal information that result in a sense of "privacy loss" [37]. For example, an app is considered more intrusive if it reads a person's information at times that are inconsistent with the person's expectations [56]. Specifically, we integrate two main privacy concepts: identifiability of the requested data

[2] The study is a joint work with Frances Zhang who completed her master thesis at 2012. Zhang implemented the system for conducting the study and analyzed the data collected in the study. I advised Zhang to apply the notion of context for analyzing the data and designed the bag-of-context approach to categorize the data collected in the study. For more details of the system and the analysis of the results gathered in the study, the readers can refer to the paper [97] we published or Zhang's master thesis.

[46] and appropriateness of the context [66], as the central constructs for quantifying the intrusiveness of an app.

We formalize a framework that quantifies the intrusiveness of sensitive data accesses performed by each app. We quantify this subjective feeling based on more objective indicators gleaned from an app's data access, namely: 1) degree of personal identifiability of the data accessed (e.g., sms is more personally identifiable than a single gps location), and 2) the privacy-relevant usage contexts under which the sensitive data are obtained (e.g., the user is not using the app, but his/her location is continuously being collected).

We believe that measuring intrusiveness is a pragmatic methodology for measuring the privacy risk caused by app usage, as the level of intrusiveness can be reduced with self-regulatory behavior, as demonstrated by contextually appropriate data gathering. The availability of this measurement can prompt app developers to clarify the scope of their app's "context of interaction", as recommended in the recent FTC report [31], and this context will align well with users' expectations of privacy. The ultimate goal of developed framework is to provide smartphone users with a quantifiable metric that they can use to easily and effectively assess and compare the relative intrusiveness of apps.

3.2.1 Methodology and Implementation

To carry out the Intrusiveness Study, we modified the existing Android platform source code and created a data access monitoring framework called *AppWindows*. The framework contains two major components: the **Monitoring Framework** and the **Logger Application**. We inserted additional logging code to Android source code in various Android modules that implement public APIs for managing access to sensor data and personal information on the phone. For example, every time an application calls a method to read location data through LocationMananger, its actions will be logged. Each log contains 1) name of the requesting application, 2) type of the data requested, 3) a timestamp, 4) contextual information about screen status (on/off), 5) contextual information about app status at the time of request (in the foreground/background), and 6) contextual location. These transaction logs can later be used to analyze whether a data request made by the application

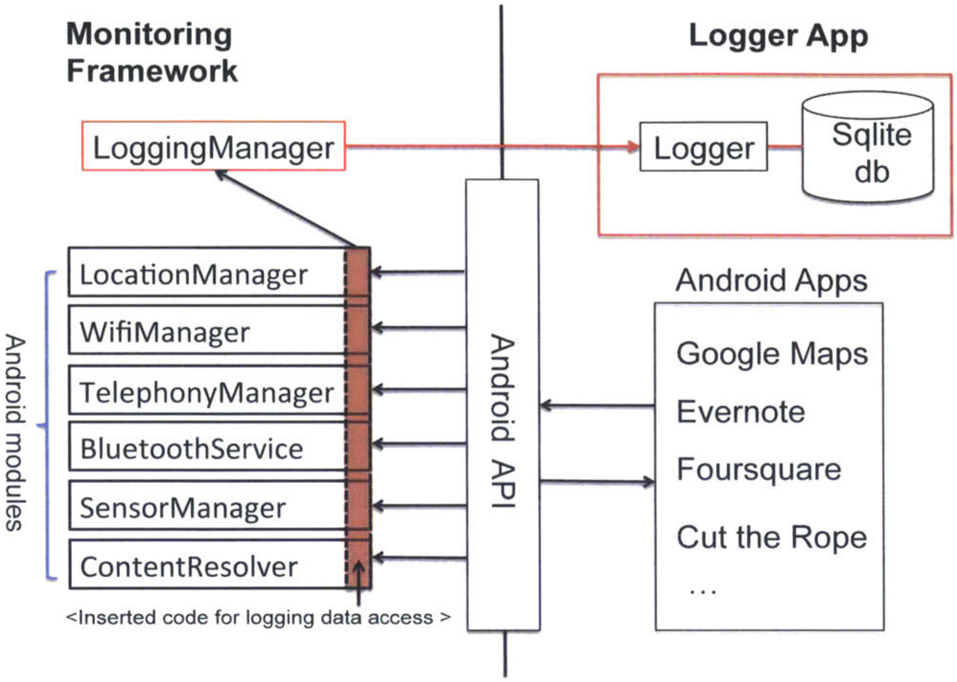


Figure 3-5: AppWindow Architecture

is a deviation from user perceived behavior or not. For example, by analyzing the logs, we may find that an app continuously reads device locations when it was in the background or even when the screen was off, indicating that the user was not interacting with their phone. Figure 3-5 shows how the Monitoring Framework (left of vertical line) and Logger App (right of vertical line) fit together in the overall architecture of AppWindow. Components outlined in red represent structures that are either a modified version of existing Android source code or newly created for AppWindow.

3.2.2 Quantifying Intrusiveness of the App

In this section, we discuss the design of our intrusiveness study and the approach for quantifying intrusiveness. The goal is to measure and compare each app's level of intrusiveness by monitoring its data access in different usage contexts. We analyzed the data access patterns obtained from real-time testing data of 6-10 apps in each of the following categories: Games (10 apps), Messaging (8 apps), Photography (6 apps), and Social (9 apps), for a total of 33 apps. All the apps ranked as the top 20 in Google Play in their category at the time of the experiment[3].

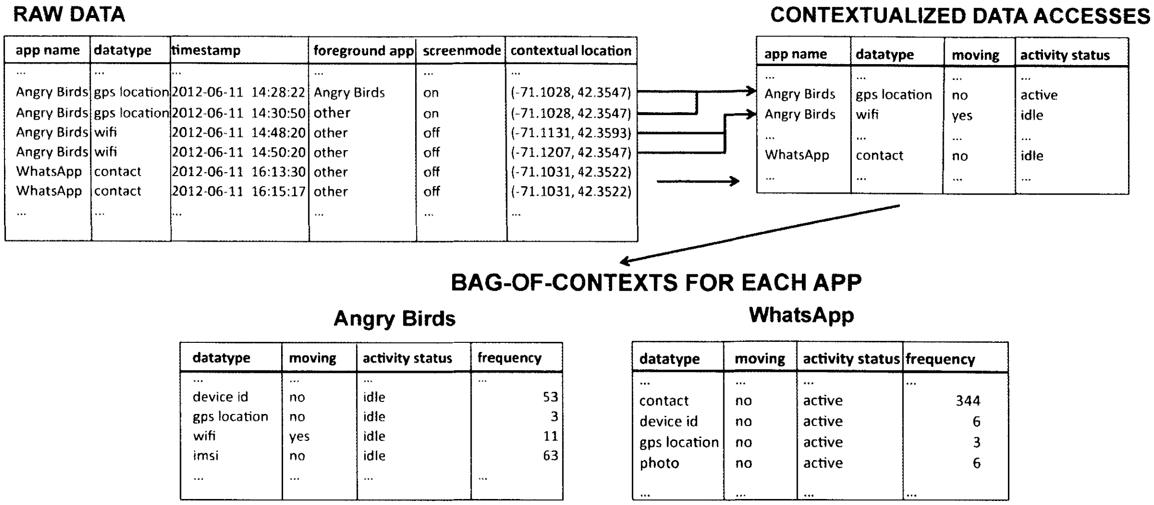
Based on preliminary testing with a handful of popular apps, we found that certain data access is triggered by particular usage patterns, namely: 1) the user's activity status (active or idle), and 2) the user's movement status at the time (moving/not moving). These usage factors are relevant to calculating intrusiveness because accessing personal information under certain usage conditions evoke different levels of negative privacy sentiment. If an app is accessing a user's location while his or her activity with respect to the app is idle (usage factor 1), then we deem that to be a highly intrusive data access. This is because the accesses occurring during a user's idle period are most likely unexpected.

We assigned two research assistants familiar with the study to test the 33 apps under 4 predefined **privacy-relevant usage contexts**: 1) User idle and user moving 2) User idle and user not moving 3) User active and moving 4) User active and not moving. We asked the experimenters to use the app under each of four usage contexts. To represent the "user idle state" (contexts 1 and 2), the experimenter clicks on the application and turns the screen off. To represent "user active" state (contexts 3 and 4), the experimenter will interact with the apps by exploring all the features that could trigger the permissions specified in the app's permission list. For the "user moving" state (contexts 1, 3), we asked the experimenters to move around while using the app. Lastly, for the "user not moving" state (context 2 and 4), they must stay in one location while testing the apps.

Intrusiveness Factors

In order to quantify intrusiveness, we first identified some factors that contribute to the intrusiveness of each data access. We choose to focus on two main factors: 1) datatype identifiability and 2) usage factors. Whereas datatype identifiability measures the inherent sensitivity associated with a particular type of personal data (e.g. GPS location, SMS, contact email), usage factors measure the intrusiveness of an access based on the condition that triggered it. We define datatype identifiability to represent how easily a particular type of data can be used to determine a person's identity. This term casts a broader net than definition of Personally Identifiable Information (PII) defined in the report written

[3] in August 2012



RAW DATA

app name	datatype	timestamp	foreground app	screenmode	contextual location
...	...	...	...	...	...
Angry Birds	gps location	2012-06-11 14:28:22	Angry Birds	on	(-71.1028, 42.3547)
Angry Birds	gps location	2012-06-11 14:30:50	other	on	(-71.1028, 42.3547)
Angry Birds	wifi	2012-06-11 14:48:20	other	off	(-71.1131, 42.3593)
Angry Birds	wifi	2012-06-11 14:50:20	other	off	(-71.1207, 42.3547)
WhatsApp	contact	2012-06-11 16:13:30	other	off	(-71.1031, 42.3522)
WhatsApp	contact	2012-06-11 16:15:17	other	off	(-71.1031, 42.3522)
...	...	...	...	...	...

CONTEXTUALIZED DATA ACCESSES

app name	datatype	moving	activity status
...	...	...	...
Angry Birds	gps location	no	active
Angry Birds	wifi	yes	idle
...	...	...	...
WhatsApp	contact	no	idle
...	...	...	...

BAG-OF-CONTEXTS FOR EACH APP

Angry Birds

datatype	moving	activity status	frequency
...	...	...	...
device id	no	idle	53
gps location	no	idle	3
wifi	yes	idle	11
imsi	no	idle	63
...	...	...	...

WhatsApp

datatype	moving	activity status	frequency
...	...	...	...
contact	no	active	344
device id	no	active	6
gps location	no	active	3
photo	no	active	6
...	...	...	...

Figure 3-6: Process of generating a bag-of-contexts

by National Institute of Standards and Technology (NIST) as a guideline to protecting the confidentiality of personal information [59]. Whereas PII only includes data that are directly linkable to personal identity, there are other types of data (e.g. GPS location, Wi-Fi, Bluetooth) [76] that can be useful in ascertaining a person's identity if collected over a period of time and examined in aggregate.

"Bag-of-Contexts" Model for Quantifying Sensitive Data Acceess

We adapt the "bag of words" model used in text processing (for computing the frequency distribution of words [75]) to construct an analogous model for assessing the distribution of privacy-relevant usage contexts of sensitive data access performed by apps. Specifically, we propose a "bag-of-contexts" model that counts the frequency of each context in the total sensitive access performed by a given app. Figure 3-6 outlines the process of generating a bag-of-contexts for each app, starting from raw testing data gathered by AppWindow for all apps installed on a smartphone. The "bag-of-contexts" model provides a different view to look at sensitive data access under the lens of context. For example in Figure 3-6, the table showing access by Angry Birds and WhatsApp in raw data and can be hard for users to compare whether these two apps behave appropriately in terms of their access to personal data on the smartphone. However, once we contextualize all data accesses using the bag-of-contexts approach, we can see that Angry Birds looks more intrusive because it makes frequent access to smartphone data when this user's context is idle which means the user is not using the app. In the contrast, data for WhatsApp shows that its data access happens mostly when the user is actively using the app.

To quantify the intrusiveness of an app based on its data access in different contexts, we designed a Intrusiveness Formula to compute the overall *Intrusiveness Score* for each application. The formula is presented in Equation below:

$$x = i * \sum_{n=1}^{k/2} w_n * p_n + a * \sum_{n=k/2+1}^{k} w_n * p_n \tag{3.1}$$

Variables

x = an app's overall intrusiveness

i = fraction of all data accesses that were conducted during the user's idle period

a = fraction of all data accesses that were conducted during the user's active period

k = total number of access contexts (in this study, $k = 48$)

w_n = intrusiveness weight assigned to the n^{th} context

p_n = the fraction of an app's total data accesses belonging to the n^{th} access context

The overall intrusiveness score of an app is the sum of the normalized Idle Intrusiveness Subscore ($i * \sum_{n=1}^{24} w_n * p_n$) and the normailzed Active Intrusiveness Subscore ($a * \sum_{n=25}^{48} w_n * p_n$). The Idle Intrusiveness Subscore is computed by summing all intrusiveness weights of all contexts corresponding to a user's idle state, normalizing each weight by its context's proportional frequency p_n. Similarly, the Active Intrusiveness Subscore is computed for all access contexts corresponding to a user's active state.

The intrusiveness weight assigned to each context is determined by two main factors: 1) the identifiability of the requested data and 2) the usage context under which the requested data is accessed. There are in total 76 unique contexts with different combinations of the above two factors. These unique contexts include 19 different data types, two modes of app status (active or idle), and two modes of users' movement status (moving or not moving). By removing trivial contexts that are not location-sensitive, specifically those that will never be triggered due to a user's movement, we are left with 48 remaining contexts. We then sorted the list of 48 contexts based on the relative intrusiveness represented by each of the context factors, in the following order: 1) user's app status (active before idle), 2) the identifiablity of the data type (in ascending order) and 3) user's movement status (not moving before moving). The resulting intrusiveness weights for all 48 contexts are shown in Table 3.4

The Intrusiveness Score represents the degree of how much an app makes data access that deviates from the appropriate usage context. The Intrusiveness Score of an app increases if the app accesses data that is highly identifiable within a context corresponding to users' idle status. The highest possible score for an app is 22.

In public view, an app's collecting data in the background without users knowing can be compared to stalking[4]. Research by Shklovski at el. showed that more than half of their survey respondents felt "deceived" and "creepy" about data collection by the apps in the background. Using the Intrusiveness Score, users have a quick indicator of how privacy invasive the app can be when accessing sensitive data on their mobile phone.

Evaluating App's Intrusiveness with Meaningful Contexts for Consumers

By quantifying the intrusiveness of an app, smartphone users have a tool to evaluate whether an app's behavior of data collection is consistent with their privacy expectations. Furthermore, they can compare apps that have similar functionality and choose a privacy-friendly app. In [97], we compared 33 popular Android apps across 4 app categories on Google Play (Google's app online market platform) by computing their Intrusiveness Score. The results showed that some popular apps exhibited a significant level of idle access activity. Thus, users should not automatically attach good privacy practices to the high reputability of app companies. Also, the level of idle access activity causes significant changes in relative intrusiveness ranking. This suggests that the level of idle activity is sometimes more impactful on relative intrusiveness than the types of data that are accessed.

The results of comparing different apps in Google Play motivated a better interface for users to choose privacy-friendly apps in the app store. We proposed a privacy dashboard that integrates both the Intrusiveness Score and the visualization of an app's data access activities created by AppWindow. Figure 3-7 illustrates an example of a privacy dashboard.

[4] http://ideas.time.com/2013/02/04/will-stalking-apps-be-stopped/

Table 3.4: Intrusiveness Weights for Access Contexts

Context ID	Datatype	Identifiability	Moving	User Activity	Context Weight
1	sms	10	no	idle	22
2	audio	10	no	idle	22
3	photo	10	no	idle	22
4	contact	10	no	idle	22
5	email	10	no	idle	22
6	phone number	10	no	idle	22
7	browsing	9	no	idle	21
8	voicemail	9	no	idle	21
9	gps location	5	yes	idle	20
10	gps location	5	no	idle	19
11	device id	5	no	idle	19
12	sim serial	5	no	idle	19
13	imsi	5	no	idle	19
14	msisdn	5	no	idle	19
15	wifi	4	yes	idle	18
16	wifi	4	no	idle	17
17	cell location	3	yes	idle	16
18	cell location	3	no	idle	15
19	neighboring cell	2	yes	idle	14
20	ip address	2	yes	idle	14
21	neighboring cell	2	no	idle	13
22	ip address	2	no	idle	13
23	sensor acce	1	yes	idle	12
24	sensor rot	1	yes	idle	12
25	sms	10	no	active	11
26	audio	10	no	active	11
27	photo	10	no	active	11
28	contact	10	no	active	11
29	email	10	no	active	11
30	phone number	10	no	active	11
31	browsing	9	no	active	10
32	voicemail	9	no	active	10
33	gps location	5	yes	active	9
34	gps location	5	no	active	8
35	device id	5	no	active	8
36	sim serial	5	no	active	8
37	imsi	5	no	active	8
38	msisdn	5	no	active	8
39	wifi	4	yes	active	7
40	wifi	4	no	active	6
41	cell location	3	yes	active	5
42	cell location	3	no	active	4
43	neighboring cell	2	yes	active	3
44	ip address	2	yes	active	3
45	neighboring cell	2	no	active	2
46	ip address	2	no	active	2
47	sensor acce	1	yes	active	1
48	sensor rot	1	yes	active	1

The Intrusiveness score in the dashboard can be further divided into two subscores: the Idle Intrusiveness Subscore and Active Intrusiveness Subscore, to highlight the substantial amounts of data access happened when smartphone users are not interacting with the apps directly. For example, in figure 3-7 the Idle Subscore is about 1.5 times (11.6/7.3) more than the Active Subscore. This indicates that a large portion of the intrusiveness score is contributed by data access when the user is not interacting with the app.

The visualization, Privacy Fingerprint, gives a more detailed and comprehensive view of all data access in context-specific lens. The two vertical stripes represent data access under idle and active usage contexts, while the horizontal bars within each strip represent the percentages of different types of data access within the usage context. The position of the type of data access in each strip is ranked in ascending order according to its sensitivity and identifiability. For example, access of GPS data is considered more sensitive and can be used to identify a person more easily than access of Wi-Fi signals. By glancing at the Privacy Fingerprint, the consumer can have an contextual view of an app's data access and make informed privacy choices. For example, as depicted in Figure 3-7, consumers who are more privacy-aware might not choose the application *PlacesForYou* after they learn that 91% of its data accesses in the idle context are for reading contacts on the phone.

3.2.3 Conclusion

The Intrusiveness Study demonstrates that apps with similar functionality can vary greatly in their behavior of data access under various usage conditions. *AppWindow* provides a new method to measure the context deviations from what users perceive as normal behavior. The framework provides a quantitative method without relying on users' subjective feelings about privacy [56] to account for the intrusiveness of an app. Usage context (idle or active mode) becomes an effective tool to determine the appropriateness of information flow from the mobile phone (as a storage place that contains personal data) to the apps (data requester and consumer of personal data) as described in Neissenbum's contextual integrity theory. This suggests that the representation of context, an often overlooked component in proposed transparency tools, can be used to inform users about the privacy implications of

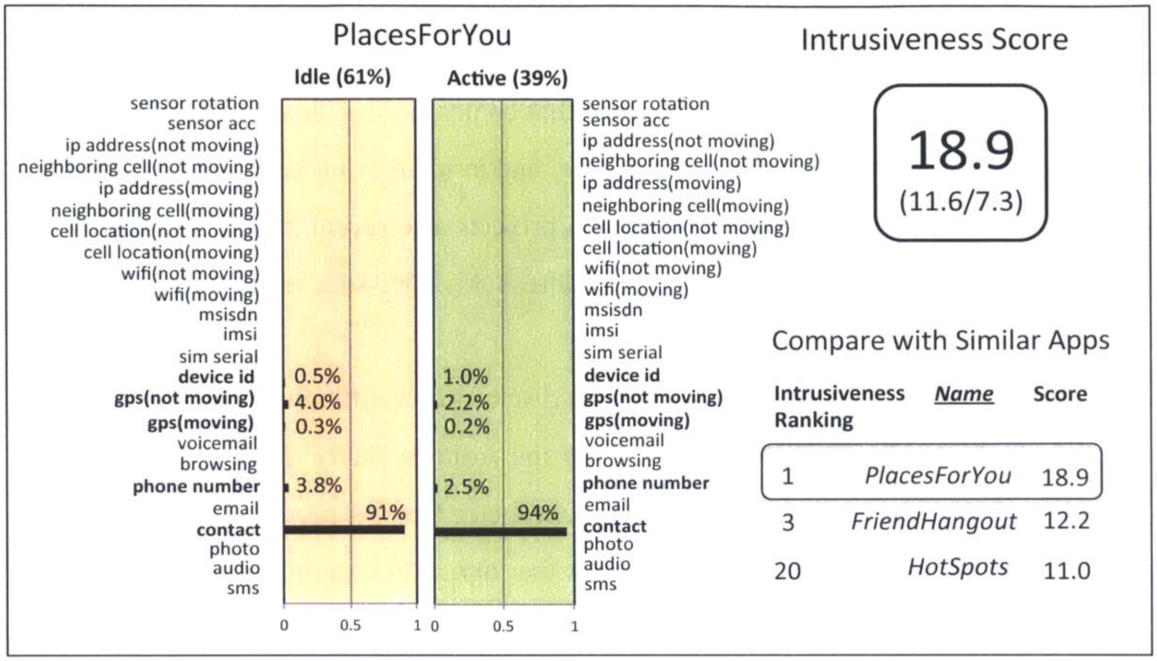


Figure 3-7: Privacy Fingerprint and Privacy Score

their data disclosure to apps.

The intrusiveness study exposed two problems with current data collection mechanisms on mobile platforms. First, norms of the information usage are not respected when apps are collecting personal data. Second, users' privacy expectations that are often inaccurate due to their limited knowledge about data collection and the opaque of the process. Therefore, users are vulnerable to privacy invasions when operating their apps under these false privacy expectations. Developers may want to gain user trust while collecting their data on mobile phones for various justified purposes such as improving apps' user experience or providing personalized services. However, current platforms provide little support for enhancing the transparency around data collection and making the users well informed.

In summary, the intrusiveness study presents a new approach to assess intrusiveness of smartphone apps based on the concept of context. The study also shows the lack of support for app developers to incorporate more transparency into data collection process. The results are a loss of control over personal data and decline in trust for apps from the consumers.

3.3 Challenges of Probing Subjects' Privacy Preferences

The two research projects in this chapter show that context is significant in helping people make privacy choices for information disclosure, and in identifying behaviors in apps that defy privacy expectations. The findings in both projects also reveal the lack of support in current mobile platforms to utilize context to address privacy issues revolving around data collection.

Studying privacy concerns for information disclosure is difficult because it depends on multiple factors. For example, the results of the Attitude Study showed that subjects' trust towards data collectors is another important factor besides personal context. Other factors take effect when subjects interact with the app. For example, the Intrusiveness Study showed that deviations in data access could be used to quantify the intrusiveness of an app.

Further, mobile apps today are more intimately tied to people's daily life, and privacy issues frequently occur when apps access data without subjects' knowledge. As pointed out by Solove [81, p.65], new technologies alter "the extent to which privacy is a dimension of certain activities" and "what do we mean when we speak about certain activities involving privacy". For example, the activity of taking a photo on a smartphone would raise subjects' privacy concerns when they discover that the app quietly uploads photos to a remote server.

One of the challenges for privacy researchers is how to probe subjects' privacy concerns, which are often subjective and context dependent, in a realistic settings and with a systematic approach. For example, the probing needs to *relate to subjects' experiences* of using the app. At the same time the probing should *reveal facts previously unknown to the user* but important to make privacy decisions, such as who is collecting the data and what it will be used for. To that end, some privacy researchers [2][17][58], as described in Chapter 2.2.2, have used the experience sampling method (ESM) to study subjects' behaviors about information disclosure. In the next chapter, we will present a framework that permits experimenters to collect mobile data and to easily conduct ESM studies by creating special-purpose mobile apps that are tailored to particular experiments.

Chapter 4

Probing Privacy Preferences in Context

This chapter describes ContextProbe, a framework that enables researchers to conduct studies using the experience sampling method (ESM). Specifically, the chapter describes how privacy researchers can conduct experiments to explore people's privacy preferences in various contexts. The chapter first summarizes the shortcomings of the survey-based approach in soliciting subjects' privacy preferences, followed by the characteristics of the ContextProbe framework in supports for conducting the experience sampling method on mobile devices. Lastly, the chapters presents a number of use cases about how experimenters can use the framework to create an app tailored to their specific study.

4.1 Moving beyond Surveys

The two studies described in Chapter 3 highlight the role of context in people's expressed privacy preferences. But they also raise several questions:

1. Do people have different privacy preferences with concern to disclosing their data in different contexts? If so, which context factors are more important?

2. Do peoples' preferences differ when they are disclosing their information to apps vs. disclosing to other people? How do they differ?

3. Are people more sensitive to purpose or to context when considering their privacy preferences?

Previous research tried to answer these questions using the survey methodology, which relies on interviews and questionnaires to draw responses within scenario-based studies. For example, interviewees or study participants are presented with a hypothetical story. They are asked to imagine themselves in the situations described in the scenario and answer survey questions based on the context given. Shilton and Martin conducted a scenario-based survey with 979 subjects to rate over 39,000 hypothetical vignettes [79] of how apps collect and use their data. They tested contextual factors including *who* (the data collector), *what* (type of disclosed information), *why* (application purpose), and *how* (use of data by data collector). Tan et al. [84] used online survey showing screenshots of permission requests from real apps and found that purpose strings have impacts on users' behavior for disclosing personal information. A shortcomings of the survey methodology is that it can introduce bias when soliciting people's privacy preferences:

- At the time that subjects are queried, they are not actually operating in the actual context that the survey is describing. Rather, they are asked to merely imagine such a context. So, their responses might not adequately reflect the privacy concerns they would actually have if asked to disclose their data while in the given context.

- The settings in which the survey questions are asked do not reinforce *privacy awareness*, which can be an important factor in triggering privacy concerns. For example, an experimenter might ask if subjects would be willing to disclose their location data to a particular mobile app. If subjects are asked only once in a questionnaire, their answers might differ greatly from a situation where they are reminded multiple times on their device that the app wants to collect their location data.

- The intrusiveness study in Chapter 3.2 demonstrated that some apps behave differently in terms of data collection when the subject's context changes. The survey methodology does not address these situations in which an app's behavior for data collection was triggered by subjects' differing contexts, and how would subjects respond (reject or allow) to data collection if asked.

In other words, people's expressed privacy preferences might be very different when they are asked about them in the context of when their data is actually being requested, as opposed to when they are asked to imagine such a context for the request.

To address these challenges, I created a framework, ContextProbe, that allows experimenters to mitigate the disconnect between the solicitation of data and the context in which the data solicitation actually occurs. The framework lets experimenters create special-purpose mobile apps tailored to their experiments, that can be distributed to subjects.

4.2 Experience Sampling Method with ContextProbe

The ContextProbe framework implements the refined ESM (rESM) [21] for probing subjects' preferences for disclosing information to apps. The refined ESM is an extension to explore issues related to mobile user experience. The refinements include: 1) automation of data collection on mobile devices and 2) triggering of data collection based on user-generated events.

In addition, ContextProbe has the following characteristics when used for conducting the ESM studies:

1. Experimenters would create mobile apps tailored to their ESM study using toolkits provided by ContextProbe.

2. Subjects are aware of what data is collected during the study

3. Subjects retain data ownership, which is stored in their online personal storage space.

4. Experimenters can easily aggregate collected data without additional requirements for hosting web servers or databases.

ContextProbe introduces new concepts to ESM studies by allowing subjects to have the ownership of their data and knowing what is been collected about them. This kind of user-centric architecture could increase trust towards data collection due to the data transparency approach used. On the other hand, it also introduces challenges of data security and authenticity, similar to those challenges [9] raised in systems supporting Personal Health Records [41], which will be discussed more in Chapter 7.

The following sections will present a number of use cases to illustrate how experimenters can use ContextProbe to solicit privacy preferences.

4.3 Building a Study App with ContextProbe

Suppose an experimenter wants to use ContextProbe to study the extent to which people's preferences about disclosing their locations to particular apps are context-specific. For example, the experimenter might wish to determine whether subjects are more willing to disclose their location to the app *WeChat* when they are in one location rather than another, such as being more willing to disclose their location when they are in the office rather than when they are at home. With this in mind, the app tailored for the experiment should have the following capabilities:

1. The app should allow the experimenters to specify different types of questions (multiple choice, yes-no, text input) tailored to their needs.

2. The app should allow the experimenter to collect subjects' responses as well as other data that can represent their context at the moment they answer the survey questions (e.g. subjects' locations and apps used by subjects).

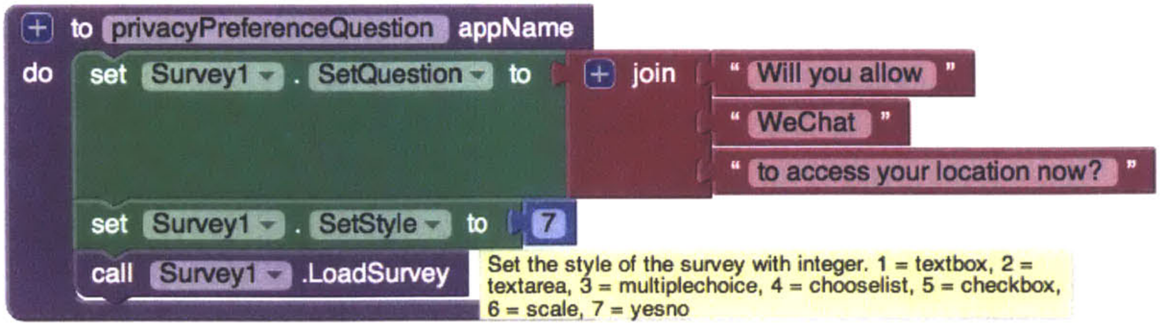
3. The app should be able to trigger survey questions based on different contexts (e.g. moments after subjects use certain apps or being at certain locations)

4. The app should be able to aggregate the data in a centralized server for further analysis

5. The app should be able to receive messages directly sent from the experimenters to allow communication promptly.

4.3.1 Configuring Survey Questions

The experimenter creates the survey content of the study app via drag-n-drop visual blocks. The visual blocks, as shown in Figure 4-1, define the style of the survey question (yes/no



```
to privacyPreferenceQuestion appName
do  set Survey1 . SetQuestion to  join  " Will you allow "
                                        " WeChat "
                                        " to access your location now? "
    set Survey1 . SetStyle to 7
    call Survey1 .LoadSurvey
```

Set the style of the survey with integer. 1 = textbox, 2 = textarea, 3 = multiplechoice, 4 = chooselist, 5 = checkbox, 6 = scale, 7 = yesno

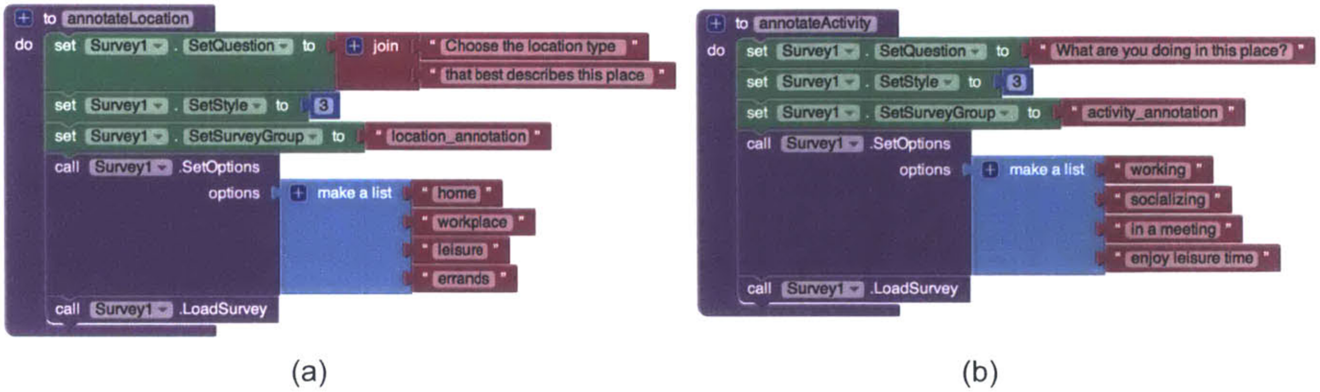
Figure 4-1: Configuring survey question for the study app

questions) in addition to the question itself. As illustrated in Figure 4-1, ContextProbe allows experimenters to configure the style of the survey question by simply giving different integer inputs. Here, the experimenter configures the survey questions for probing subjects' privacy preferences as questions yielding yes/no answers.

Using survey blocks, the experimenter can also configure the study app to ask subjects to annotate locations they have visited, indicating the type of location and the kind of activity performed there. These annotations can be used as auxiliary information to help experimenters understand subjects' contexts through their reporting subjective interpretations of their contexts. These social-cultural aspects of subjects' contexts have significant effects on their privacy preferences [58], as will be illustrated in the results of the experiment described in Chapter 6. As shown in Figure 4-2, the experimenter specifies that both of the questions that to be asked in annotating a subject's current location will be multiple-choice questions. The overall choices for the type of survey answer are: textbox, textarea, multiple-choice, chooselist, checkbox, scale, and yes-no.

4.3.2 Configuring the App to Collect Data for Personal Context

ContextProbe collects two kinds of data: *passively collected data* and *subject-reported data*. The subject-reported data, including context annotations and privacy preferences for disclosing data, are responses obtained by soliciting subjects. The passively collected data are those that the mobile device can collect automatically without explicit interaction with the subjects.



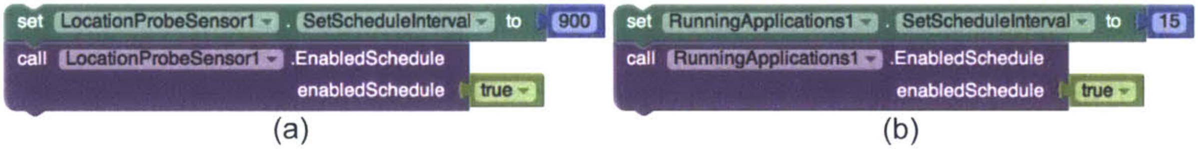
```
to annotateLocation
do  set Survey1 . SetQuestion to  join  " Choose the location type "
                                        " that best describes this place "
    set Survey1 . SetStyle to 3
    set Survey1 . SetSurveyGroup to " location_annotation "
    call Survey1 .SetOptions
                  options  make a list  " home "
                                        " workplace "
                                        " leisure "
                                        " errands "
    call Survey1 .LoadSurvey
```

(a)

```
to annotateActivity
do  set Survey1 . SetQuestion to " What are you doing in this place? "
    set Survey1 . SetStyle to 3
    set Survey1 . SetSurveyGroup to " activity_annotation "
    call Survey1 .SetOptions
                  options  make a list  " working "
                                        " socializing "
                                        " in a meeting "
                                        " enjoy leisure time "
    call Survey1 .LoadSurvey
```

(b)

Figure 4-2: Configuring survey question for soliciting subjects' annotations of their current location, including the type of location and their current activity



```
set LocationProbeSensor1 . SetScheduleInterval to 900
call LocationProbeSensor1 .EnabledSchedule
                          enabledSchedule  true
```

(a)

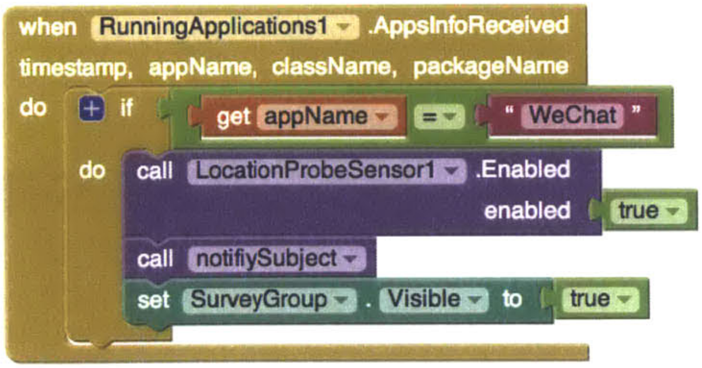
```
set RunningApplications1 . SetScheduleInterval to 15
call RunningApplications1 .EnabledSchedule
                          enabledSchedule  true
```

(b)

Figure 4-3: Configuring the frequencies of data collection

In the *WeChat* example, the experimenter wants to collect information about a subject's locations and about the other apps used by the subject. The experiment uses the block *LocationProbeSensor* and *RunningApplications*, as shown in Figure 4-3, to configure the app to collect location every 15 minutes (900 seconds) and collect information about apps runing on the device every 15 seconds.

ContextProbe can leverage sensors and available personal information on a subject's mobile phone to report about 13 kinds of data that can be useful for representing context. For example, by reading the Wi-Fi sensor, the experimenter can use the data to identify a subject's mobility pattern. Bluetooth sensors can be used to discover other nearby devices. Proximity data collected from the Bluetooth sensors can be used to detect subjects' social circles. More generally, ContextProbe can provide other information such as the subjects' calling records or SMS histories, representing a person's communication patterns.



```
when RunningApplications1 .AppsInfoReceived
timestamp, appName, className, packageName
do  if  get appName = " WeChat "
    do  call LocationProbeSensor1 .Enabled
                                  enabled  true
        call notifiySubject
        set SurveyGroup . Visible to  true
```

Figure 4-4: Configure survey triggers

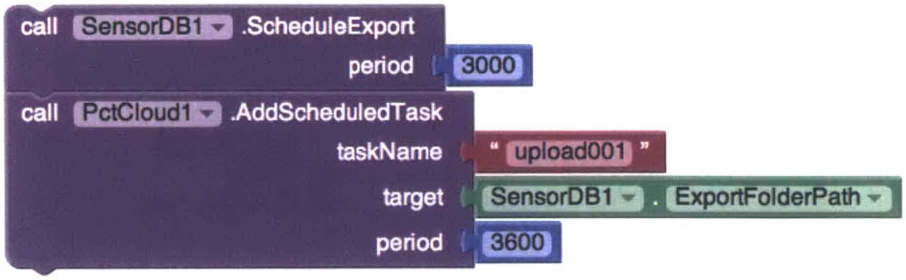
4.3.3 Configuring Triggers for Survey Questions

In the *WeChat* example, the experimenter wants to prompt the subjects every time they use the app. To achieve that, the experimenter can use the visual block that keeps track of which app is currently being used by the subject. As shown in Figure 4-4, the outermost visual block *AppsInfoReceived* indicates an app-use event being triggered. Other blocks define the conditions under which data collecttion should be triggered. The blocks shown here stipulate that if the app currently being used by the subject is WeChat, then the subject will be prompted with the survey questions and the current location will be captured as well.

4.3.4 Uploading Collected Data to the Cloud

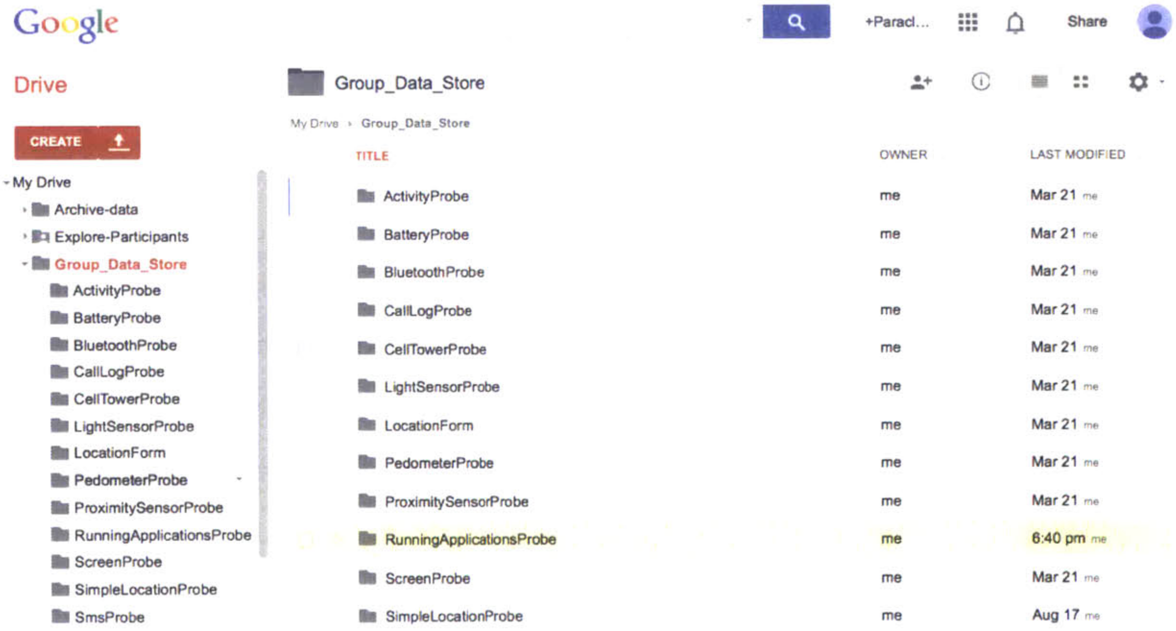
The experimenter next uses *SensorDB* and *PctCloud* blocks to specify the frequency for exporting and uploading the collected data, including passively collected data and the subject-reported data. As shown in Figure 4-5, the study app will upload the collected data to the subjects' Google Drive storage space. The experimenter does not need to specify the URL of the application server for the data upload because ContextProbe uses the Google Drive storage space of the subject to serve as the storage layer for the collected data.

Periodically, the collected data will be sent from subjects' Google Drive to the experimenter's application server and stored as spreadsheets in the experimenter's Google Drive. Figure 4-6 shows what experimenters will see for all data collected from the subjects' mobile phones. The data are aggregated in different folders according to their types.



```
call SensorDB1 .ScheduleExport
                        period 3000
call PctCloud1 .AddScheduledTask
                     taskName " upload001 "
                       target SensorDB1 . ExportFolderPath
                       period 3600
```

Figure 4-5: Configure frequency for uploading collected data

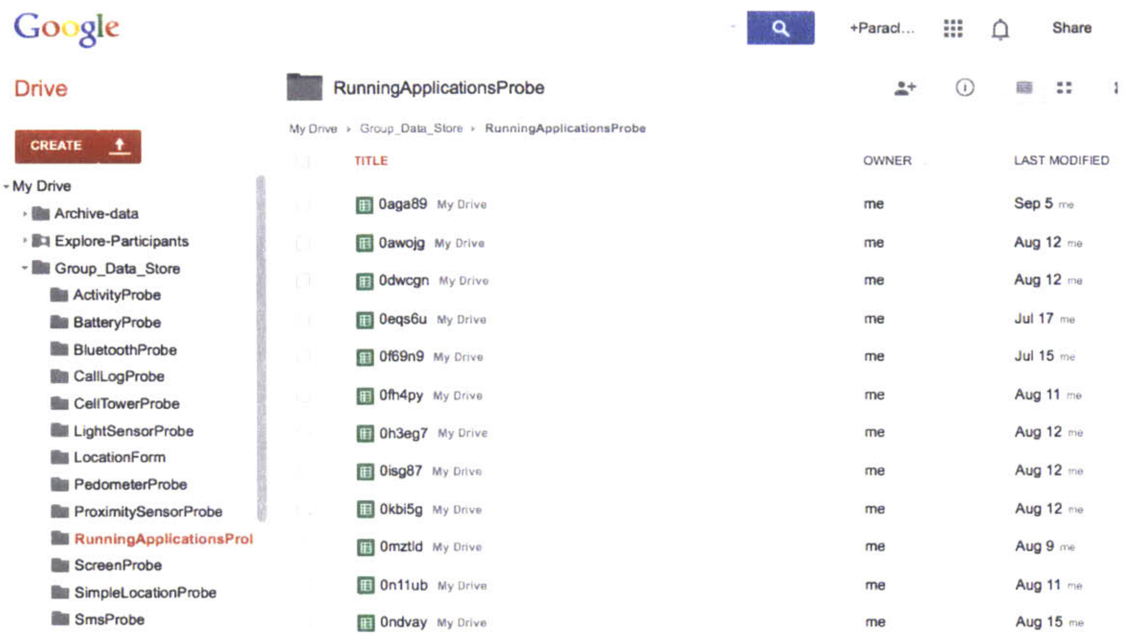


- My Drive
  - Archive-data
  - Explore-Participants
  - Group_Data_Store
    - ActivityProbe
    - BatteryProbe
    - BluetoothProbe
    - CallLogProbe
    - CellTowerProbe
    - LightSensorProbe
    - LocationForm
    - PedometerProbe
    - ProximitySensorProbe
    - RunningApplicationsProbe
    - ScreenProbe
    - SimpleLocationProbe
    - SmsProbe

TITLE	OWNER	LAST MODIFIED
ActivityProbe	me	Mar 21 me
BatteryProbe	me	Mar 21 me
BluetoothProbe	me	Mar 21 me
CallLogProbe	me	Mar 21 me
CellTowerProbe	me	Mar 21 me
LightSensorProbe	me	Mar 21 me
LocationForm	me	Mar 21 me
PedometerProbe	me	Mar 21 me
ProximitySensorProbe	me	Mar 21 me
RunningApplicationsProbe	me	6:40 pm me
ScreenProbe	me	Mar 21 me
SimpleLocationProbe	me	Aug 17 me

Figure 4-6: Folders on experimenters' Google Drive space to store the collected data from study subjects

For example, as shown in Figure 4-7, the *RunningApplicationsProbe* folder contains many spreadsheet files, each representing raw data of a subject's app usage history. Each file is named with a unique identifier that is anonymous, without revealing the subject's identifiable information. Using spreadsheets to store raw data allows experimenters to conduct data analysis with the spreadsheet's built-in functions. Also, the spreadsheet can be easily exported and shared with others.



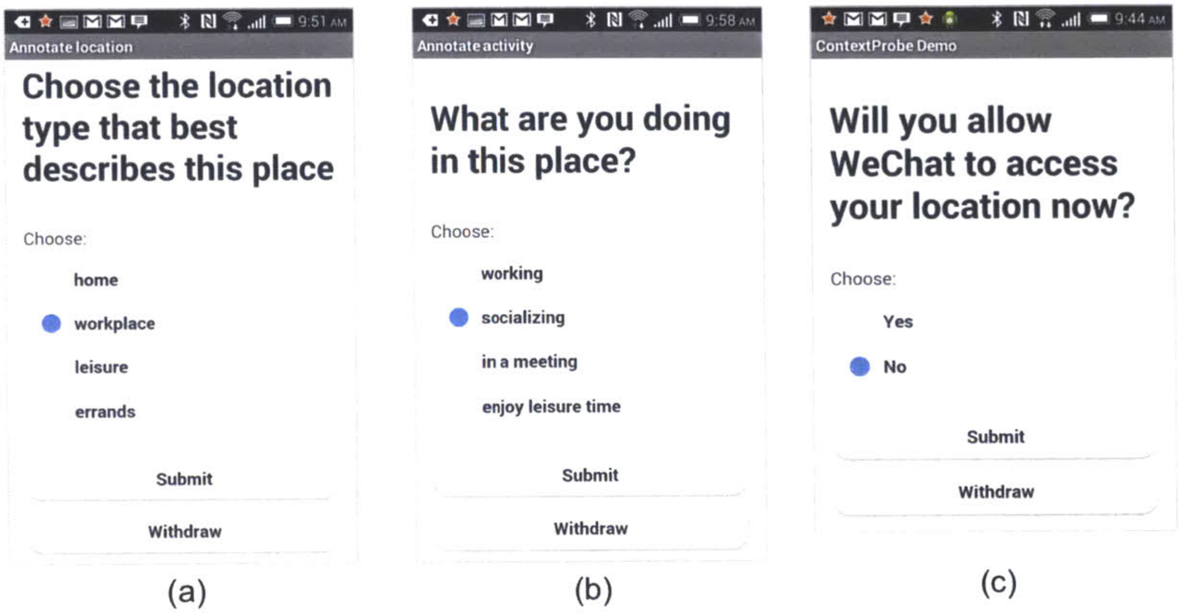
TITLE	OWNER	LAST MODIFIED
0aga89 My Drive	me	Sep 5 me
0awojg My Drive	me	Aug 12 me
0dwcgn My Drive	me	Aug 12 me
0eqs6u My Drive	me	Jul 17 me
0f69n9 My Drive	me	Jul 15 me
0fh4py My Drive	me	Aug 11 me
0h3eg7 My Drive	me	Aug 12 me
0isg87 My Drive	me	Aug 12 me
0kbi5g My Drive	me	Aug 12 me
0mztld My Drive	me	Aug 9 me
0n11ub My Drive	me	Aug 11 me
0ndvay My Drive	me	Aug 15 me

Figure 4-7: Google Drive folder that stores the running application data collected from subjects

4.3.5 ContextProbe App in Action

After the app is configured and built, the experimenter downloads it from the app building platform, distributes it to subjects, and asks them to install the app on their mobile phones. When the study app is opened for the first time, it asks the subject to grant permission to access their Google Drive in order to set up the storage space for the collected data. More details will be discussed in Chapter 5.

During the experiment, whenever subjects use *WeChat* on their mobile phone, the study app will prompt for answers to a set of survey questions, containing both the annotation questions and a privacy preference question as shown in Figure 4-8. The first two questions ((a) and (b)) are for annotating the subjects' current locations and the third question (c) asks for the subjects' privacy preferences for disclosing their current location to *WeChat*. Note how the survey questions' interface matches to the visual blocks used by the experimenter during the app-building process (see Figure 4-2).



Annotate location

Choose the location type that best describes this place

Choose:

home

● workplace

leisure

errands

Submit

Withdraw

(a)

Annotate activity

What are you doing in this place?

Choose:

working

● socializing

in a meeting

enjoy leisure time

Submit

Withdraw

(b)

ContextProbe Demo

Will you allow WeChat to access your location now?

Choose:

Yes

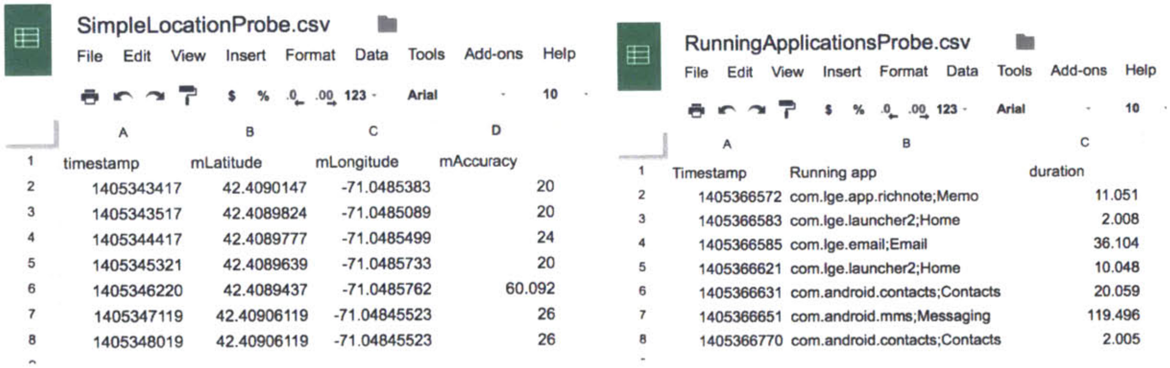
● No

Submit

Withdraw

(c)

Figure 4-8: The survey questions shown on the study app



SimpleLocationProbe.csv

	A	B	C	D
1	timestamp	mLatitude	mLongitude	mAccuracy
2	1405343417	42.4090147	-71.0485383	20
3	1405343517	42.4089824	-71.0485089	20
4	1405344417	42.4089777	-71.0485499	24
5	1405345321	42.4089639	-71.0485733	20
6	1405346220	42.4089437	-71.0485762	60.092
7	1405347119	42.40906119	-71.04845523	26
8	1405348019	42.40906119	-71.04845523	26

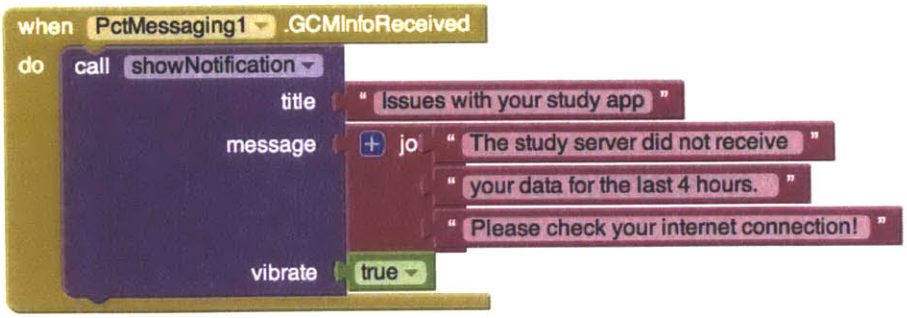
(a) Location data

RunningApplicationsProbe.csv

	A	B	C
1	Timestamp	Running app	duration
2	1405366572	com.lge.app.richnote;Memo	11.051
3	1405366583	com.lge.launcher2;Home	2.008
4	1405366585	com.lge.email;Email	36.104
5	1405366621	com.lge.launcher2;Home	10.048
6	1405366631	com.android.contacts;Contacts	20.059
7	1405366651	com.android.mms;Messaging	119.496
8	1405366770	com.android.contacts;Contacts	2.005

(b) Running applications data

Figure 4-9: Data passively collected by ContextProbe



```
when PctMessaging1 .GCMInfoReceived
do  call showNotification
                 title  " Issues with your study app "
               message  jo " The study server did not receive "
                            " your data for the last 4 hours. "
                            " Please check your internet connection! "
               vibrate  true
```

Figure 4-10: Configure the app to receive pushed notification from the application server

4.3.6 Monitoring the App and Communicating with Subjects

Multiple issues may arise when running such a mobile sensing experiment *in the wild*. Some common issues include missing data uploads because mobile phones are not able to connect to the Internet, or missing too many answers from the subjects. When issues arise, the experimenter may want to notify the subjects with a message directly sent to their mobile phones. In such cases, the experimenter wants to monitor how subjects answer the survey questions as well as the data collection tasks running on the subjects' mobile phones. In addition to the app building platform, ContextProbe's framework includes an application server that the running app will access. ContextProbe provides predefined scripts on the application server for experimenters to check the health status of data collection tasks by analyzing data in the aggregated data store. The script checks the "heartbeats" of the study app on the subjects' mobile phones. If a large portion of data is missing from a particular subject, this indicates that the study app is either not active or there are issues in collecting and uploading data.

To monitor the heartbeats of the study app, the experimenter needs to configure both the app and the application server. In the app-building process, the experimenter uses the block *GCMInfoReceived*, as shown in Figure 4-10, to configure the app for receiving cloud messages pushed from the application server via Google Cloud Messaging technology. The blocks here show that the study app will alert the subject once it receives the pushed messages.

On the application server, the experimenter uses an interface to configure the type of data that should be monitored and the frequency to execute the script for checking the data's health status. In the *WeChat* example, the experimenter configures the script to monitor the *Location* and *RunningApplications* types of data and to be executed every 12 hours on the application server. The script will identify those mobile phones that have issues with collecting and uploading the data and send a predefined message directly to the subjects' mobile phones.

4.4 Summary

In this chapter, we have summarized challenges of soliciting subjects' privacy preferences and the limitations of survey-based methods. We introduced a framework, ContextProbe, that simplifies the task of running ESM studies. Chapter 5 will provide a detailed description of the implementation details of ContextProbe. Chapter 6 describes a specific study that uses ContextProbe to explore privacy preferences in the light of contextual information about user location and behavior.

Chapter 5

Architecture and Implementation of the ContextProbe Framework

This chapter describes the design and implementation of ***ContextProbe***, a framework for for creating experiments that use the Experience Sampling Method (ESM) [53] to solicit subjects' responses in different situations. Specifically, the framework allows experimenters who are not familiar with mobile programming to easily build apps for collecting *in situ* responses as well as other data on mobile phones that can be used for representing personal context. In addition, the framework provides transparency with regards to data collection in a way that can raise the subjects' awareness of ongoing data collection tasks.

The chapter begins with the motivations for building the ContextProbe framework. Next, the chapter gives an overview of ContextProbe to highlight the roles and functions of the different major parts. Lastly, the chapter introduces implementation details of the submodules within each part.

5.1 Motivation

The Experiences Sampling Method (ESM) is an approach that researchers use to solicit and record subjects' responses regarding their daily experiences. With the growing prevalence of smartphones, researchers have adopted ESM on mobile platforms and have built apps that capture subjects' daily experiences while on the go. These mobile apps have common features in order to serve as an experimental testbed for ESM, including, among others, 1) a survey module, 2) a module for collecting sensor data, 3) a scheduling module to create time-based triggers for the survey, and 4) an uploader module to upload collected data to an application server hosted by the experimenter.

Experimenters who wish to conduct ESM studies face several challenges besides designing the experiment. First of all, not all experimenters have strong programming skills to create a mobile app tailored to their ESM study. Aside from creating the app, experimenters have to set up the infrastructure, such as an application server and database to aggregate and store data.

These challenges motivated the creation of ContextProbe to alleviate experimenters' burdens and help them to easily create ESM apps and conduct experiments using these apps.

5.2 Overview

As shown in Figure 5-1, ContextProbe consists of three parts: 1) an ***app building platform*** to create mobile apps that use sensors for data collection, 2) the ***Personal Data Store*** that serves as a data repository for storing the collected data for each subject, and 3) the ***Application Server*** that aggregates all the sensor data collected from subjects' Personal Data Store.

The app building platform contains components that can be used to build apps for an ESM study, including 1) multiple sensor components for mobile sensing and data collection, 2) cloud components for uploading collected data to remote data repositories and receiving pushed notifications from external web servers, and 3) a timer component for scheduling and activating tasks on the mobile phone based on time, and 4) two survey components for creating customized surveys. Using these components, the experimenters could create apps tailored to their experiments and distribute these apps to study subjects.

The Personal Data Store (PDS) serves as a personal data repository to store data uploaded from a subject's mobile phone. The PDS is built on top of the Google Drive service with several built-in scripts executed as web services. Specifically, the PDS is run and

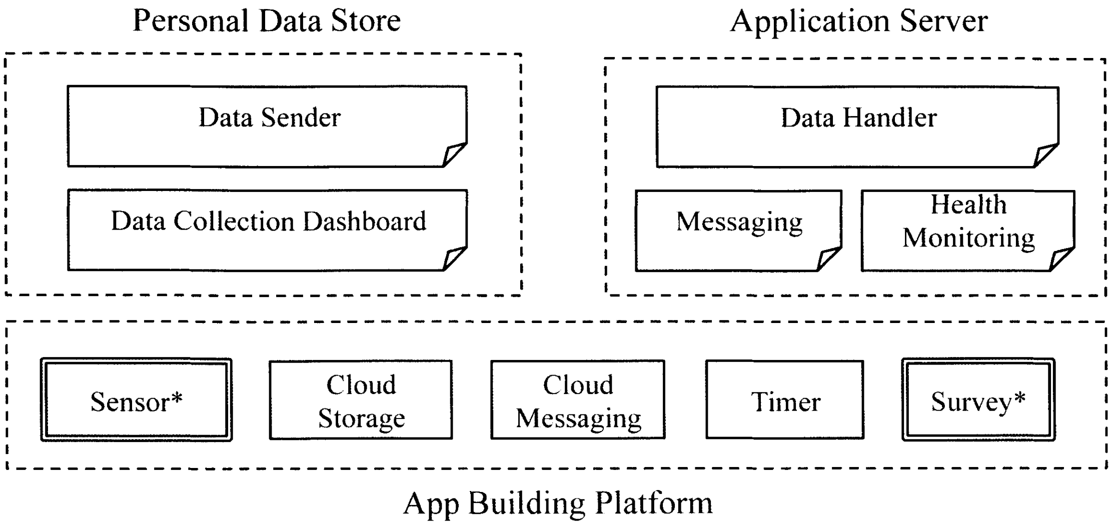


Figure 5-1: Overview of ContextProbe architecture

hosted on the subject's Google Drive space. The ***Data Sender*** script sends data to the application server on a preconfigured time interval. It is possible that some subjects might participate in different experiments and install multiple data collection apps built with ContextProbe. Each app would have its own storage space in the PDS with a Data Sender script preconfigured to send the information to the corresponding application server.

Sometimes the subjects may want to keep track of the data that have been collected by different study apps on their mobile phone. The Personal Data Store has features that allow subjects to monitor the status of ongoing data collection tasks. The ***Data Collection Dashboard*** script presents a summary page about the status of each installed app. This information includes the types of data that each app collects and the timestamp of the latest update. The dashboard also provides some visualization features for certain data types.

The Application Server hosts all the data collected from the subjects' mobile phones. The ***Data Handler*** script aggregates data sent by the Personal Data Store and organizes it by types. The ***Health Monitoring*** script checks for the health of data collection tasks by examining the data periodically. If issues arise, experimenters can use the ***Messaging*** script to send real time notifications to the subjects.

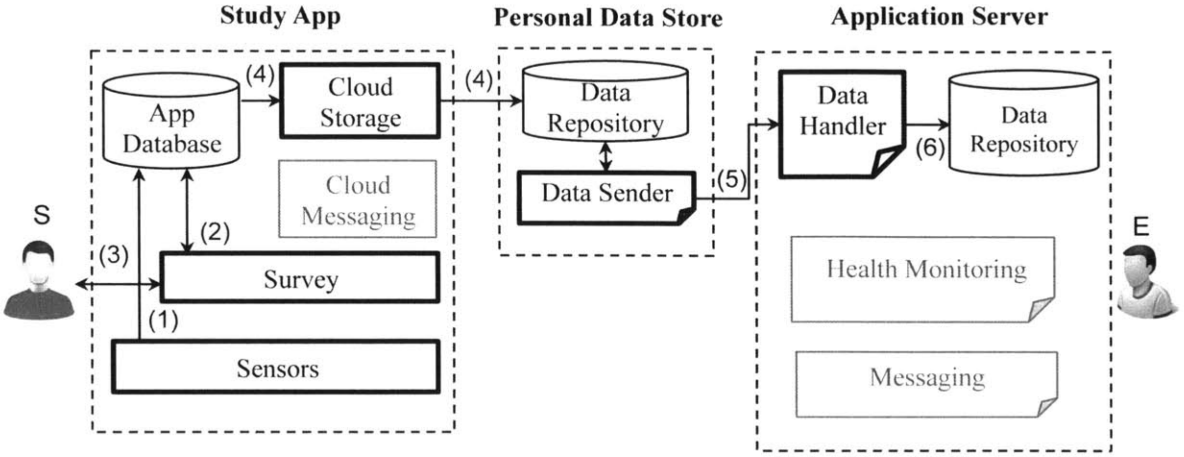


Figure 5-2: Workflow for data collection and aggregation in ContextProbe (S for Subject, E for Experimenter)

5.2.1 Workflow of ContextProbe

This section describes the two main workflows within the ContextProbe framework. One workflow starts with data collection on the study app and ends with data aggregation on the application server. Another workflow describes the steps for experimenters to monitor data collection and address the issues (see Section 5.5.3) in the experiment. Figure 5-2 shows how data is collected and uploaded to the cloud and eventually aggregated on the application server. The study app in Figure 5-2 illustrates a typical ESM app that can be built with the previously mentioned components on the app building platform.

Once the study app is installed on subjects' mobile phone, data collection happens in the following steps:

1. The Sensor components, running on the mobile phone as background services, access sensor data and store it locally in a SQLite database.

2. The Survey component creates survey questions using information from the data stored in the database.

3. The Survey component prompts subjects to answer the survey questions at times that have been preconfigured by the Timer component. After subjects answer the questions, the Survey component saves the answers in the database.

4. The Cloud Storage component periodically exports data from the database and uploads the data to the subject's Personal Data Store hosted on their Google Drive space.

5. The Data Sender script in the Personal Data Store pushes data periodically to the application server at preconfigured time intervals.

6. On the application server, the Data Handler script aggregates data sent from different subjects' PDS and stores it according to data type (see Section 5.5.2).

During the experiment, the experimenter may want to be notified about issues with the data collection (such as missing data uploads) and communicate with subjects. Figure 5-3 shows how this is accomplished:

1. The Health Monitoring script runs periodically to monitor the health status of data collection tasks. The script checks the data for each subject and identifies potential issues with collection or uploading tasks.

2. The Health Monitoring script sends results via email to the experimenter.

3. The Health Monitoring script calls the Messaging script to send messages directly to the mobile phones of those subjects identified as having issues. The Messaging script sends messages through Google Cloud[1] Messaging service.

4. On the study app, the Cloud Messaging component receives the message and notifies the subject.

5.3 App Building Platform

The ContextProbe framework extends App Inventor by adding features to support ESM experiments. App Inventor [94] is an open-source and web-based app development platform that allows its users to create mobile apps for Android devices by dragging and dropping

[1] See Section 5.3.4

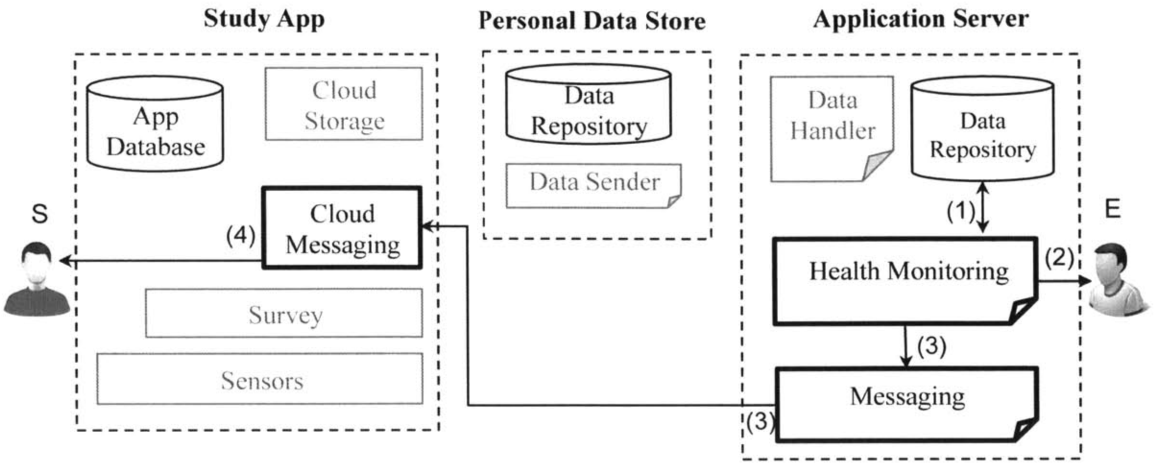
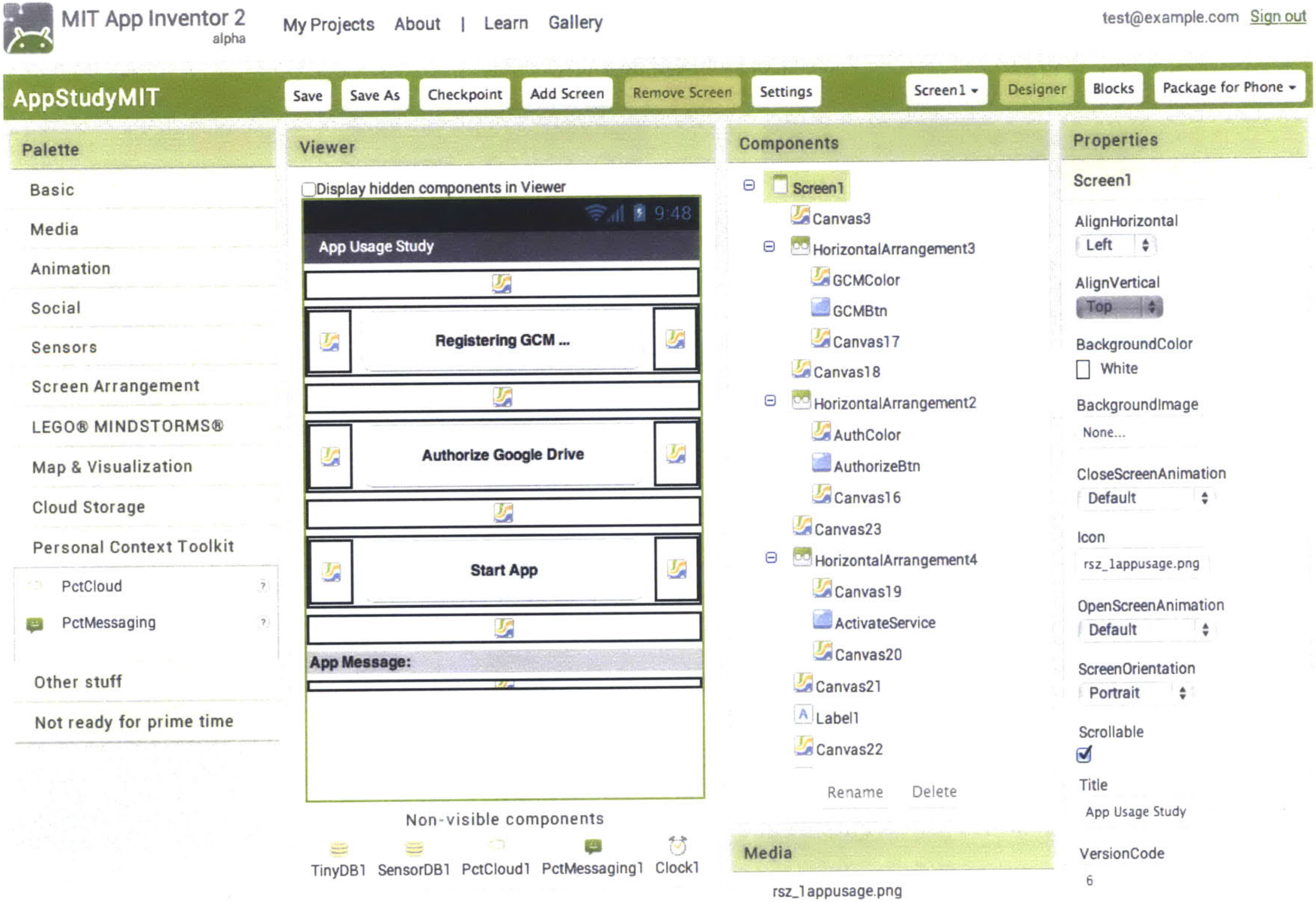


Figure 5-3: Workflow for data collection monitoring and cloud messaging

visual blocks. App Inventor users create mobile apps by piecing together visual blocks in a WYSIWYG environment. The resulting apps are ready to use and can be downloaded to mobile phones. Figure 5-4 shows the interface in App Inventor for designing the layout for an app using UI widgets, while Figure 5-5 shows the interface for implementing the logic of an app by connecting different blocks.



AlignHorizontal: Left

AlignVertical: Top

BackgroundColor: □ White

BackgroundImage: None...

CloseScreenAnimation: Default

Icon: rsz_1appusage.png

OpenScreenAnimation: Default

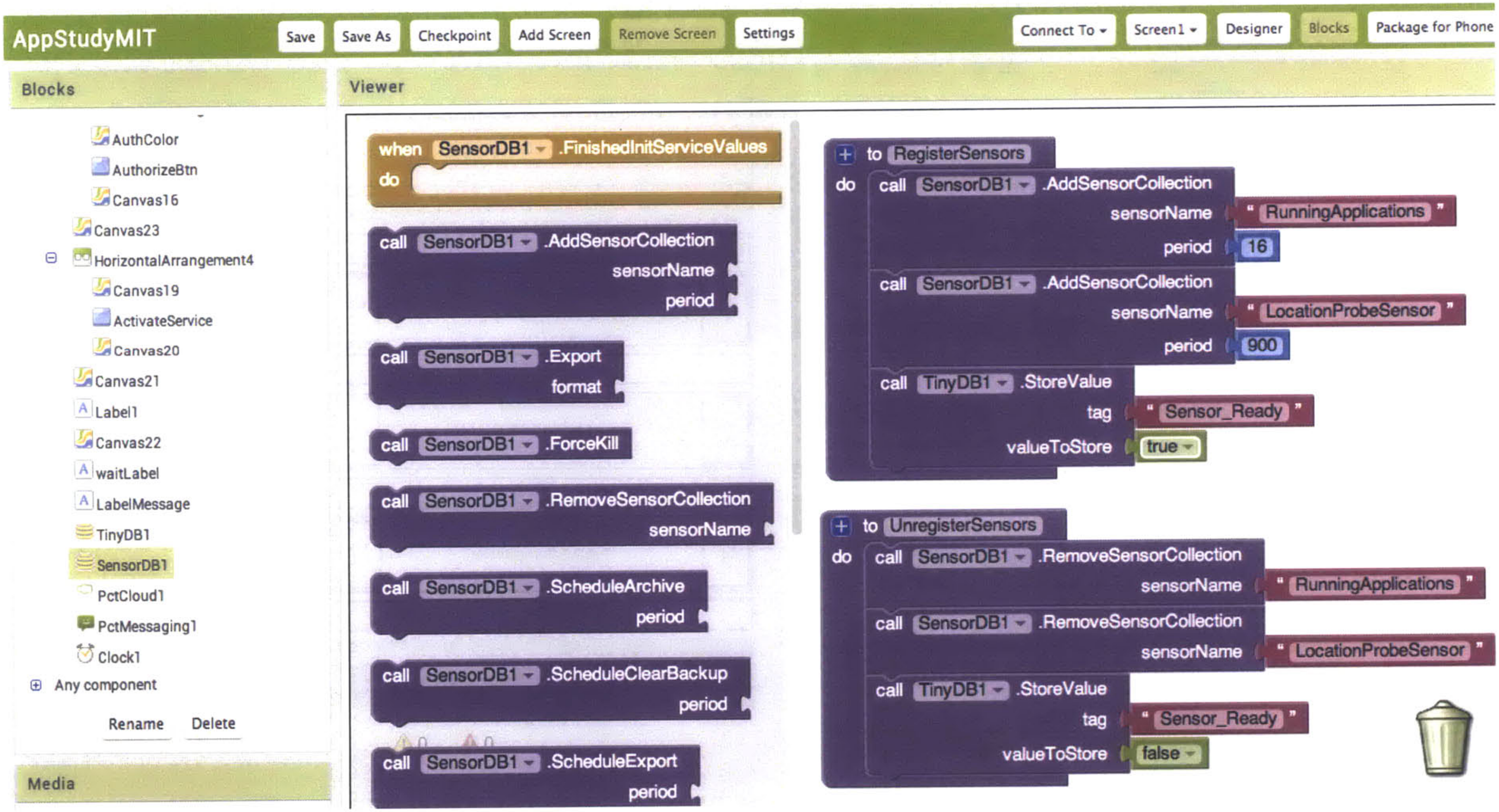
ScreenOrientation: Portrait

Scrollable: ☑

Title: App Usage Study

VersionCode: 6

Figure 5-4: App Inventor designer view



```
when SensorDB1 .FinishedInitServiceValues
do

call SensorDB1 .AddSensorCollection
                     sensorName
                         period

call SensorDB1 .Export
             format

call SensorDB1 .ForceKill

call SensorDB1 .RemoveSensorCollection
                          sensorName

call SensorDB1 .ScheduleArchive
                     period

call SensorDB1 .ScheduleClearBackup
                         period

call SensorDB1 .ScheduleExport
                    period

to RegisterSensors
do  call SensorDB1 .AddSensorCollection
                         sensorName  " RunningApplications "
                             period  16
    call SensorDB1 .AddSensorCollection
                         sensorName  " LocationProbeSensor "
                             period  900
    call TinyDB1 .StoreValue
                        tag  " Sensor_Ready "
               valueToStore  true

to UnregisterSensors
do  call SensorDB1 .RemoveSensorCollection
                            sensorName  " RunningApplications "
    call SensorDB1 .RemoveSensorCollection
                            sensorName  " LocationProbeSensor "
    call TinyDB1 .StoreValue
                        tag  " Sensor_Ready "
               valueToStore  false
```

Figure 5-5: App Inventor block view

Besides typical programming language primitives, App Inventor provides many special components that encapsulate complicated interactions between the app and the underlying Android operating system and also other external resources on the Web. For example, the *Web* component allows app developers to focus on retrieving contents from the Internet without having to write boilerplate code for managing network connections. Following the same principle, ContextProbe includes sensor component extensions that enable app developers to harness the data sensing capability of mobile devices. In addition, other components support cloud storage and cloud messaging capabilities, creating surveys, and scheduling tasks.

5.3.1 Mobile Sensing Components

The implementation of sensor components in App Inventor uses an open source library called *Funf* [5]. The library provides functionality for a collection of a broad range of sensor types. It also provides a unified and configurable interface for developers to access different types of data on Android phones. The integration with App Inventor makes it easy for novices to create sensor-enabled apps by configuring appropriate parameters as inputs to App Inventor program blocks.

The current implementation of App Inventor does not support the creation of Android *Services*, a feature used for executing long-running operations without direct interaction by mobile phone owners, while Funf makes use of Services to implement sensing capabilities.

To bridge this gap, I designed each sensor component to be a "controller" for how the underneath Android Service should conduct sensing tasks. For example, a *WifiSensor* component is used to start, configure, and stop an Android service class in the Funf library that scans and fetches information from nearby wireless access points.

The ContextProbe framework adds 15 different types of sensor components to App Inventor. Here, a broader definition of "sensor data" is used to mean "data on the mobile phone that can be captured in the background without subjects' interaction". Besides reading data from hardware sensors such as accelerometer and GPS, the sensor components can also be used to capture personal information on the mobile phone such as currently used applications, call logs, and contact information. Table 5.1 shows a list of sensor components that are available the ContextProbe framework.

1. ActivityProbe	6. LightSensor	11. ScreenStatus
2. BatterySensor	7. LocationProbe	12. SmsHistory
3. CallLogHistory	8. PedometerSensor	13. SocialProximitySensor
4. CellTowerProbe	9. ProximitySensor	14. TelephonyInfo
5. ContactInfo	10. RunningApplications	15. WifiSensor

Table 5.1: 15 different sensor components for data collection

Below are short descriptions of the kinds of data that each sensor component collects:

ActivityProbe Reads accelerometer data to detect movement

BatterySensor Reports the current battery level of the mobile phone

CallLogHistory Returns all the incoming and outgoing call logs. The logs include the phone number contacted, date, and duration

CellTowerProbe Returns information about nearby cellular towers

ContactInfo Returns information of user contacts such as email addresses, phone numbers, and contacts' first and last names.

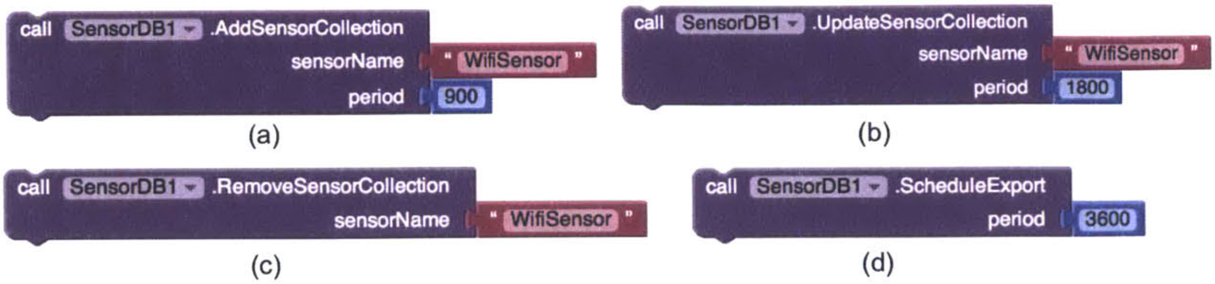
LightSensor Returns the ambient illuminance level

LocationProbe Returns geolocation information obtained from the GPS sensor on the phone. The information includes latitude, longitude and accuracy measure

PedometerSensor Returns step counts by detecting the movement

ProximitySensor Returns a value indicating how close the front of the device is to an object

RunningApplications Returns the information of the application that is currently running in the foreground and being used by the subject



```
call SensorDB1 .AddSensorCollection
                 sensorName  " WifiSensor "
                 period  900
```

(a)

```
call SensorDB1 .UpdateSensorCollection
                 sensorName  " WifiSensor "
                 period  1800
```

(b)

```
call SensorDB1 .RemoveSensorCollection
                 sensorName  " WifiSensor "
```

(c)

```
call SensorDB1 .ScheduleExport
                 period  3600
```

(d)

Figure 5-6: Examples of methods in the SensorDb component

ScreenStatus Returns the screen status (on/off)

SmsHistory Returns all incoming and outgoing SMS messages. The data include message type, message body, and the phone number contacted

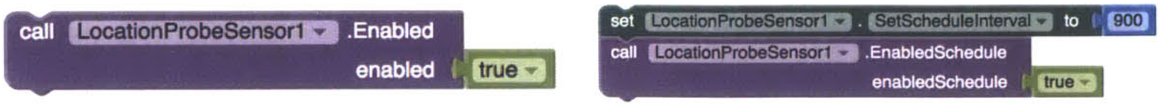
SocialProximity Returns Bluetooth hardware information of other nearby mobile devices (only devices with Bluetooth enabled can be detected).

TelephonyInfo Returns basic phone information including the phone number and the network type

WifiSensor Returns hardware information about the nearby wireless access points

SensorDb Allows the app to export data stored in the database by other sensor components. Also integrates with other sensor components to provide a one-stop data collection interface for all the available sensors.

To simplify data collection tasks, an app can use the *SensorDB* component, an abstraction of a local database for collecting and storing different types of sensor data with a unified interface. As shown in Figure 5-6a, experimenters could specify the name of the sensor that they wish to add for data collection tasks and the time interval for the duty-cycling data request. The SensorDB component also provides methods for data management such as exporting the database as text files in Comma-Separated Values (CSV) format, updating collection intervals, and removing a specific data collection task of a particular sensor.



```
call LocationProbeSensor1 .Enabled
                       enabled true
```

```
set LocationProbeSensor1 . SetScheduleInterval to 900
call LocationProbeSensor1 .EnabledSchedule
                   enabledSchedule true
```

(a) Enable one-time sensing

(b) Schedule duty-cycling sensing

Figure 5-7: The two block methods here represent the two modes for sensing in a sensor component. The app can either perform one-time sensing or schedule a duty-cycling sensing. For example, figure (b) shows that the app is scheduled to read location data every 15 minutes

Configuring sensor components

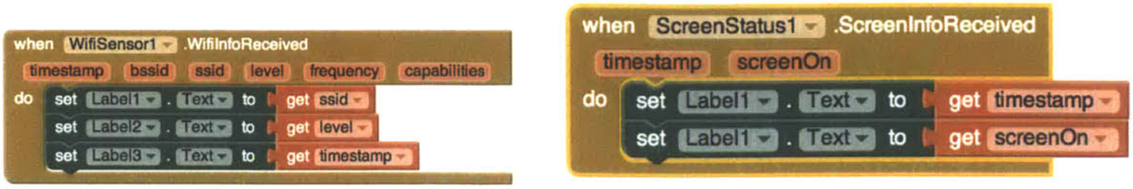
Each sensor component provides two methods for capturing information on the mobile phone: one-time polling and periodic polling. Figure 5-7a shows the block that can be used for the one-time polling request. The periodic polling method, as shown in Figure 5-7b, enables the app to schedule a duty-cycling data request with different intervals. For instance, experimenters might use one-time polling to fetch contact information, and use periodic polling to frequently update the location of a mobile phone.

The periodical polling method is implemented through integration with the Funf library, which uses the *AlarmManager* class of the Android SDK[2] to schedule data fetching tasks. These tasks are triggered by the Android system periodically and return sensor data as requested.

Handling data stream

Data returned in the sensor component are handled by the *event handler* type of method in App Inventor. The method is named as *SensorName.InfoReceived* in the visual block for each component. The implementation for the sensor component registers sensor event listeners with the underneath Android SDK. The event handler is called whenever there is a new datum detected or whenever the value has changed for the registered sensor. Figure 5-8 shows two examples of receiving data from two different sensor components. In Figure 5-8a, the *WifiSensor* component scans all nearby wireless access points and returns the information as a stream of data through the event handler every time when the WiFi

[2] Software development kit



```
when WifiSensor1 .WifiInfoReceived
  timestamp  bssid  ssid  level  frequency  capabilities
do  set Label1 . Text to  get ssid
    set Label2 . Text to  get level
    set Label3 . Text to  get timestamp
```

(a) WifiSensor

```
when ScreenStatus1 .ScreenInfoReceived
  timestamp  screenOn
do  set Label1 . Text to  get timestamp
    set Label1 . Text to  get screenOn
```

(b) Sensor info receive 1

Figure 5-8: Examples of information returned by the *WifiSensor* component and the *ScreenStatus* sensor component

sensor is triggered. On the other hand, the *ScreenStatus* component returns data only when subjects interact with the mobile phone by turning the screen on or off.

5.3.2 Survey Component

The survey component provides different styles of survey questions for experimenters to choose from. These styles of questions include 1) yes-no, 2) checkbox, 3) choose-list, 4) multiple choice, 5) scale, 6) textbox, and 7) text area. The implementation of the survey component uses default HTML templates for presenting survey questions and an Android WebView class for displaying the survey content as web pages. The HTML templates are preloaded as Android asset files on the App Inventor platform. At runtime, the app would determine which template to load to display the survey questions as specified by the experimenter in the code blocks (see the example given in Figure 4-1).

Advanced survey component

The advanced survey component gives experimenters the capability to create more complicated surveys by providing customized HTML files and additional Javascript files. The implementation of the advanced survey component uses an Android WebView to show survey content and the *JavascriptInterface* to pass data objects (survey answers) from Javascript code back to application code. For example, an app can use customized HTML templates to display survey questions and implement the logic in Javascript code for how to present the survey questions. Once answers are received, they can be stored using method calls through the *JavascriptInterface*.

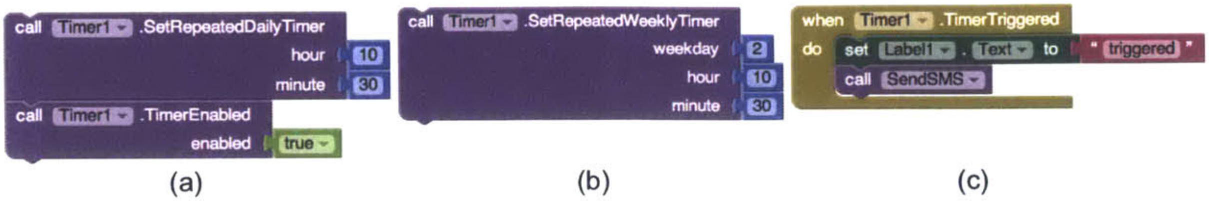

call Timer1 .SetRepeatedDailyTimer
hour 10
minute 30
call Timer1 .TimerEnabled
enabled true
(a)
call Timer1 .SetRepeatedWeeklyTimer
weekday 2
hour 10
minute 30
(b)
when Timer1 .TimerTriggered
do set Label1 . Text to " triggered "
call SendSMS
(c)

Figure 5-9: (a) Configure a daily alarm to be triggered at 10:30 am (b) Configure a weekly alarm to be triggered on Monday 10:30 am (b) Receive an event when the timer is triggered

5.3.3 Timer Component

The **Timer** component provides the capability for the experimenter to create an “alarm” that wakes up the app to perform some action at a specified time. The implementation of the Timer component communicates with a long-running background process to create alarms with Android *AlarmManager*. Figure 5-9 shows examples of using the component to create daily and weekly alarms to receive events when the timer is triggered.

5.3.4 Cloud Messaging Component

The cloud messaging component (*PctMessaging*) allows the app to receive real-time messages from other web servers via Google Cloud Messaging (GCM) [45]. GCM is a popular web service that “pushes" data onto a registered device. The service is independent of the device’s current location and network connection availability and it will try to deliver the message until successfully sent.

To use the GCM service, Android app developers need to write code, both for the client and server sides, following the steps shown in Figure 5-10:

1. The app needs to register with GCM servers
2. The GCM server returns a registration ID to the app
3. The app sends the registration ID to the application server and registers with the application server
4. The application server sends acknowledge signal indicating that registration is complete

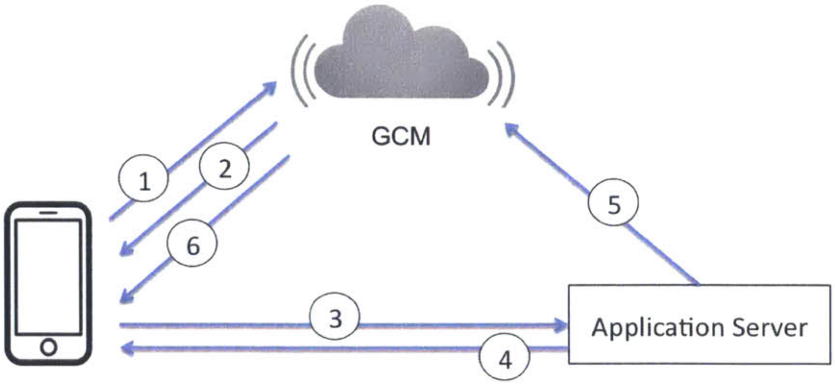
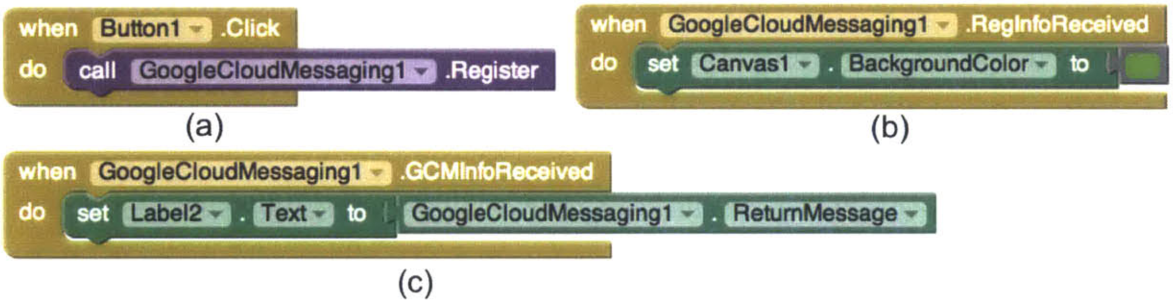


Figure 5-10: Protocol for utilizing GCM service



when Button1 .Click
do call GoogleCloudMessaging1 .Register

(a)

when GoogleCloudMessaging1 .RegInfoReceived
do set Canvas1 . BackgroundColor to 🟩

(b)

when GoogleCloudMessaging1 .GCMInfoReceived
do set Label2 . Text to GoogleCloudMessaging1 . ReturnMessage

(c)

Figure 5-11: Examples of methods provided in the cloud messaging component *PctMessaging*

5. The application server calls the GCM service to send messages to the device

6. The GCM server takes messages from the application server and sends them to the device

ContextProbe simplifies cloud messaging for app developers by providing an implementation of both the client and server parts that are required when using the GCM service. On the client side, the implementation of the cloud messaging component (see Figure 5-11) hides most of the details in Figure 5-10 for App Inventor users.

Unlike other App Inventor components, the cloud messaging component will receive pushed messages while the app is in the background or when it is not in use. The implementation of this component registers a listener for a GCM event on the Android system and wakes up the device only when a message is received.

5.3.5 Cloud Storage Component

The cloud storage component (*PcTCloud*) in the ContextProbe framework leverages free cloud storage space of each subject's Google Drive to store data collected on the mobile phone. Using Google Drive for storing collected data has the following advantages:

1. Subjects can keep a copy of the collected data in their Google Drive before sending it to the application server on the experimenter side.
2. Google Drive can execute programs written in *Google Apps Script* (GAS), a script language based on Javascript that can be used for task automation. This provides the foundation for the Personal Data Store that can store, monitor, and send collected data to the application server.

In the Personal Data Store, each app has a specific folder to store different types of data that are collected on the subjects' phones (see Figure 5-15). When performing the task, *PctCloud* will automatically upload data exported by the *SensorDb* component to the app folder in subjects' PDS and convert the data to Google Spreadsheet files.

5.3.6 Bootstrapping the ContextProbe Data Service

One advantage of using apps built by ContextProbe for conducting experiments about data collection is that the framework simplifies the process of registering subjects for the study and the process of bootstrapping the services that send data to the application server. The PctCloud component integrates with preconfigured GAS scripts to automate the entire process.

The implementation of the cloud storage component uses the Google Drive SDK to upload collected data. Google Drive requires the app to be authorized by subjects before it can upload any data. For this reason, the authorization process needs to be integrated into the initialization of the PctCloud component.

Figure 5-12 shows the workflow of an app that uses the *PctCloud* and *PctMessaging* to register subjects, request authorization, and bootstrap the data service with the integration of the application server. The boxes shown in the workflow with dashed lines are steps

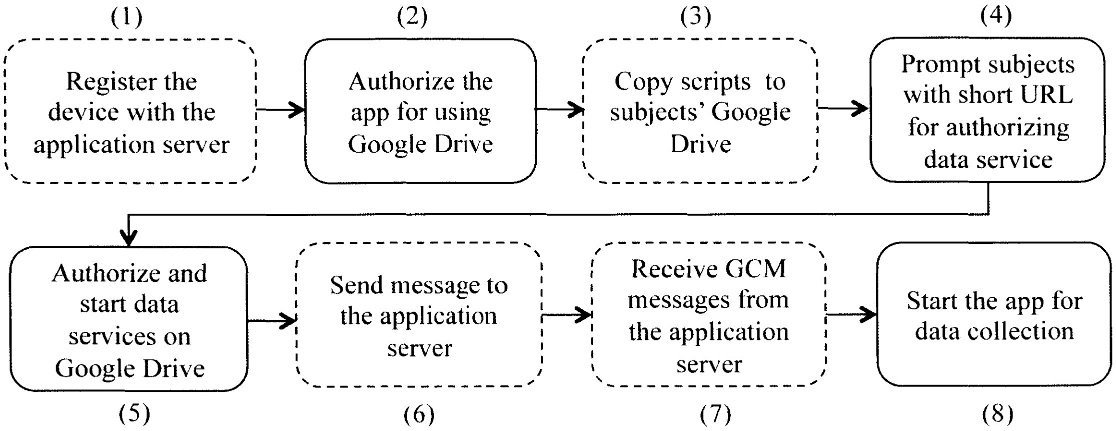

(1)
Register the device with the application server
(2)
Authorize the app for using Google Drive
(3)
Copy scripts to subjects' Google Drive
(4)
Prompt subjects with short URL for authorizing data service
Authorize and start data services on Google Drive
(5)
Send message to the application server
(6)
Receive GCM messages from the application server
(7)
Start the app for data collection
(8)

Figure 5-12: Workflow for starting the app and services for Personal Data Store on the subject side

executed by the app in the background; the rest are steps that involve subjects' interaction. Below are details of the execution for each step in the workflow:

1. **Register the device** When a subject opens the app for the first time, the *PctMessaging* component will first register the device with the GCM service. Then it will register the device with the application server using the registration ID obtained from the GCM service. The application server will store the registration ID and use it as an identifier linked to the subject throughout the study.

2. **Authorize the app** The app will trigger the authorization process and prompt the subject to grant permissions for the app to access the subject's Google Drive. Figure 5-13a shows a snapshot of what subjects will see.

3. **Copy Data Sender script and configuration file** Once the app has obtained permission from the subject, it will copy a Google Spreadsheet document, *Client-App-Controller*, to Google Drive. The spreadsheet (see Figure 5-14a) contains UI widgets and the Data Sender scripts that are responsible for sending collected data from the Personal Data store to the application server. In addition, a configuration file is created and uploaded to the app folder. The file has information about the application server and a unique identifier that is used later to register the device with the application server.

4. **Notify the subject** Once the spreadsheet is copied, *PctCloud* will retrieve its URL and convert it to a short URL. As shown in Figure 5-13b, the app will immediately show a popup window to notify the subject that the Personal Data Store is ready and awaits authorization to start the service.

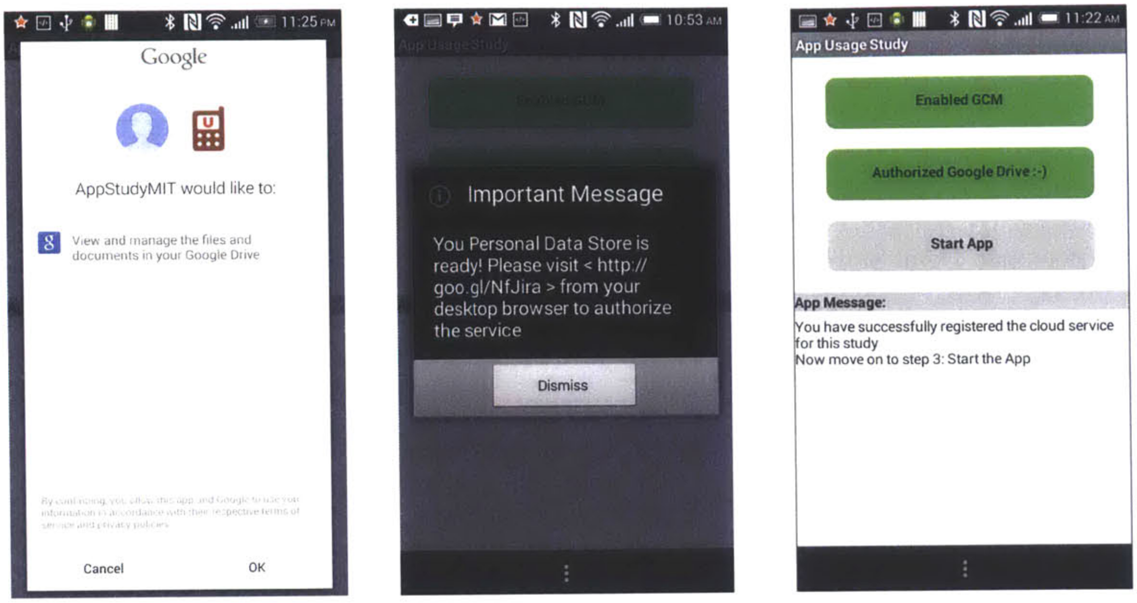
5. **Authorize and start the service** To execute a program on Google Drive, subjects need to grant permissions for reading and managing files. Subjects have to visit the spreadsheet using the short URL. The spreadsheet has two buttons with a short description for authorizing and starting the data service (see Figure 5-13a). When the "Authorize" button is clicked, a window will pop up to display the permissions that the app will be granted, as shown in Figure 5-13c.

6. **Notify the application server to complete the registration process** After subjects authorize and start the scripts, the PDS will send a message to the application server signaling that it is now ready to receive data. Optionally, the application server can send a message through the GCM service to notify that the registration process has been completed.

7. **Start Data Collection** Once the app receives the GCM message, it can notify the subject that data collection tasks are ready to start (see Figure 5-13c), depending on the app's implementation.

5.4 Personal Data Store

The Personal Data Store (PDS) is a data repository that stores all collected data uploaded by the apps created with the ContextProbe framework. Conceptually, the PDS consists of many **app-specific submodules** that store data and communicate with the corresponding application servers. The implementation of the PDS uses Google Drive [44], a popular and free cloud storage solution. Figure 5-15 shows the structure of how data is organized in the PDS. Each app-specific submodule is a Google Drive folder that contains spreadsheet files of the collected data uploaded by each app. The submodule also has a **Data Sender** script



Google

AppStudyMIT would like to:

View and manage the files and documents in your Google Drive

By continuing, you allow this app and Google to use your information in accordance with their respective terms of service and privacy policies

Important Message

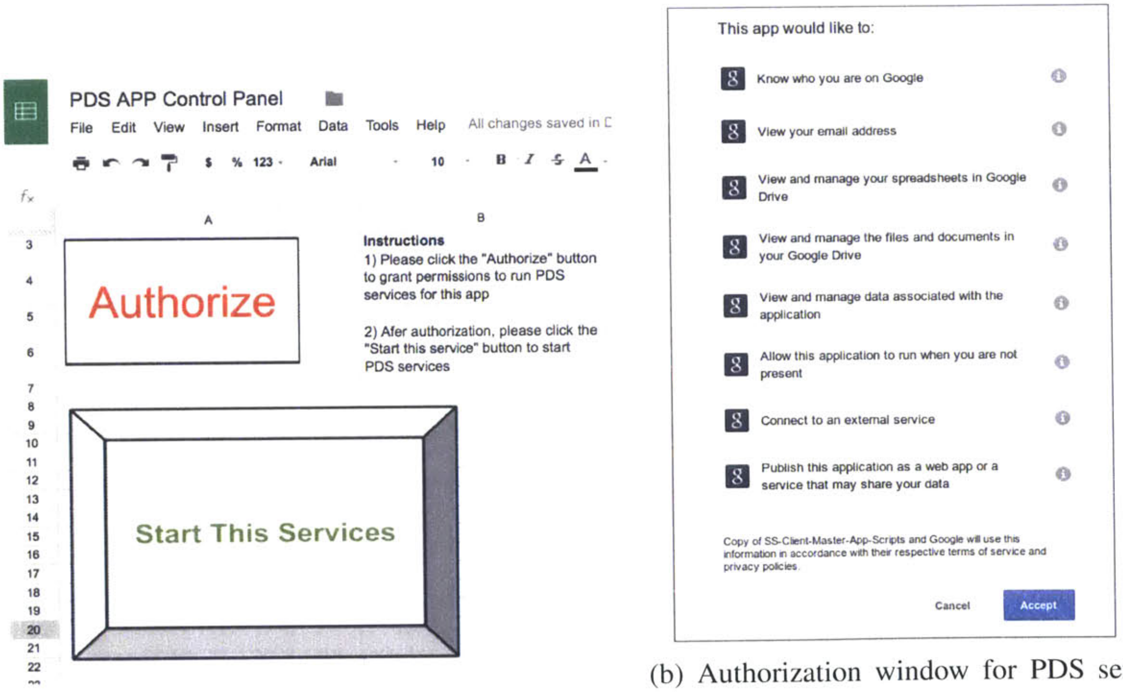
You Personal Data Store is ready! Please visit < http://goo.gl/NfJira > from your desktop browser to authorize the service

App Usage Study

App Message:
You have successfully registered the cloud service for this study
Now move on to step 3: Start the App

(a) Authorize Google Drive (b) Notify the subject (c) Receive pushed message

Figure 5-13: (a) Subjects need to grant permissions for the app to copy scripts to the PDS (b) A pop up window that notifies the subjects to visit the file on the PDS and to grant permissions for the service to be started (c) The app receives pushed messages from the application server when the process is complete



PDS APP Control Panel

File Edit View Insert Format Data Tools Help All changes saved in
Arial 10
A
B

Authorize

Instructions
1) Please click the "Authorize" button to grant permissions to run PDS services for this app

2) Afer authorization, please click the "Start this service" button to start PDS services

Start This Services

This app would like to:

- Know who you are on Google
- View your email address
- View and manage your spreadsheets in Google Drive
- View and manage the files and documents in your Google Drive
- View and manage data associated with the application
- Allow this application to run when you are not present
- Connect to an external service
- Publish this application as a web app or a service that may share your data

Copy of SS-Client-Master-App-Scripts and Google will use this information in accordance with their respective terms of service and privacy policies.

(a) Control panel of app-specific submodule

(b) Authorization window for PDS service

Figure 5-14: Subjects would click the authorization button and follow the instructions to start the data service on PDS

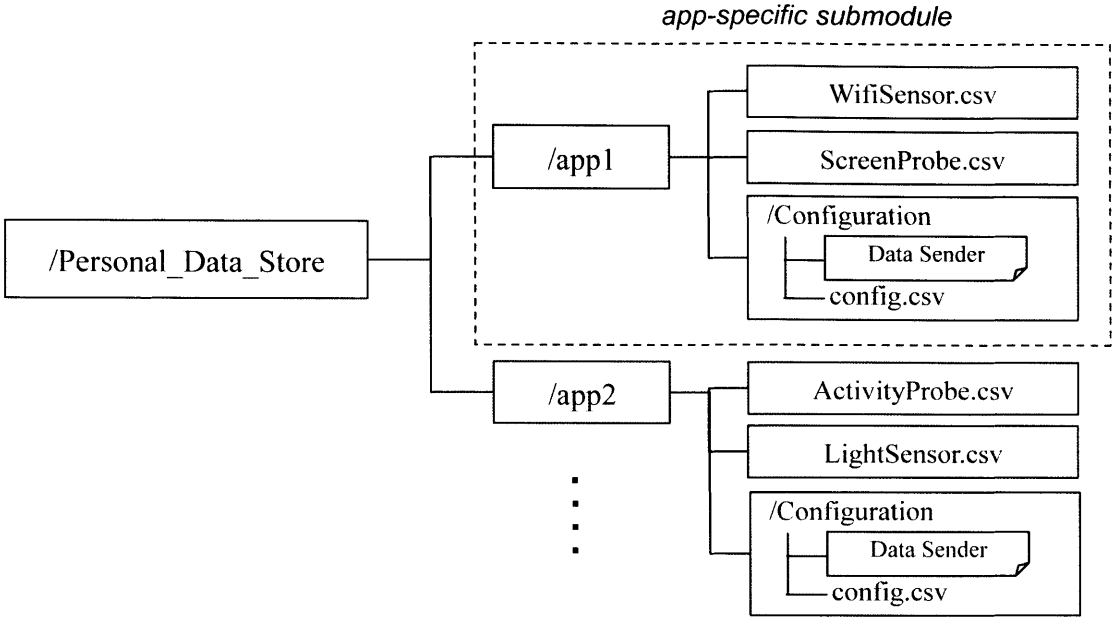


Figure 5-15: The structure of how collected data are organized in the Personal Data Store

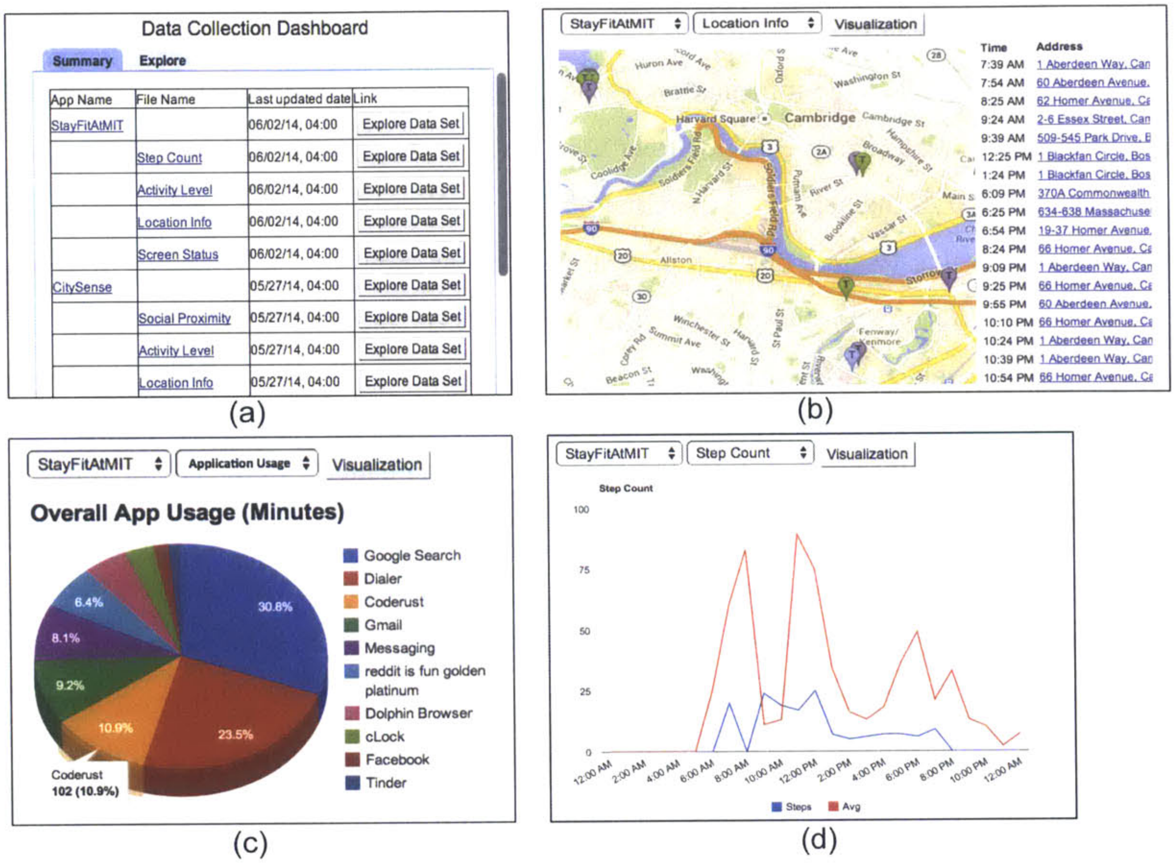
that is responsible for sending data to the corresponding application server. As mentioned in section 5.3.5, the *PctCloud* component automatically creates folders for the app in Google Drive and copies the default Data Sender scripts during the initialization process.

The Data Sender script is triggered every hour to send only the newly available data to the application server. The script reads information stored in the configuration file to find out the application server's URL and the device's ID when sending the data.

5.4.1 Data Collection Dashboard

The Personal Data Store provides an alternative to the current process of data collection on mobile phones, in which data is often collected in an opaque way and sent directly to experimenters' servers without the subjects' knowledge. With the PDS, data is uploaded to the subjects' personal space before being sent to the experimenters, so the subjects now retain the ownership of their own data. Thus, adopting the PDS architecture in ContextProbe affords a foundation for providing transparency and control over personal data.

The **Data Collection Dashboard** provides subjects with a summary of the sensor data collected by different apps. Subjects might join several experiments and have apps collecting different kinds of data in the background on their phones simultaneously. The dashboard on the PDS offers subjects a quick overview of the types of data that have been collected, as well as the timestamp of when the data was last uploaded to the PDS. To enable the dashboard service, subjects need to manually copy a script to their Google Drive space and start it as a web app.



App Name	File Name	Last updated date	Link
StayFitAtMIT		06/02/14, 04:00	Explore Data Set
	Step Count	06/02/14, 04:00	Explore Data Set
	Activity Level	06/02/14, 04:00	Explore Data Set
	Location Info	06/02/14, 04:00	Explore Data Set
	Screen Status	06/02/14, 04:00	Explore Data Set
CitySense		05/27/14, 04:00	Explore Data Set
	Social Proximity	05/27/14, 04:00	Explore Data Set
	Activity Level	05/27/14, 04:00	Explore Data Set
	Location Info	05/27/14, 04:00	Explore Data Set

Time	Address
7:39 AM	1 Aberdeen Way, Can
7:54 AM	60 Aberdeen Avenue,
8:25 AM	62 Homer Avenue, Ca
9:24 AM	2-6 Essex Street, Can
9:39 AM	509-545 Park Drive, B
12:25 PM	1 Blackfan Circle, Bos
1:24 PM	1 Blackfan Circle, Bos
6:09 PM	370A Commonwealth
6:25 PM	634-638 Massachuse
6:54 PM	19-37 Homer Avenue,
8:24 PM	66 Homer Avenue, Ca
9:09 PM	1 Aberdeen Way, Can
9:25 PM	66 Homer Avenue, Ca
9:55 PM	60 Aberdeen Avenue,
10:10 PM	66 Homer Avenue, Ca
10:24 PM	1 Aberdeen Way, Can
10:39 PM	1 Aberdeen Way, Can
10:54 PM	66 Homer Avenue, Ca

Figure 5-16: (a) Data collection dashboard which shows all the apps that are currently collecting data (b) Visualization of a subject's location (c) Visualization of the subject's app usage information (c) Visualization of pedometer data

As shown in Figure 5-16(a), there are two different interfaces in the dashboard for improving data transparency and raising subjects' awareness of the ongoing data collection:

1. **Data Summary** The summary page encapsulates all information about the ongoing data collection tasks by different apps. It shows the name of the app, links to the files that store the sensor data, last update time, and links to the visualization of the sensor data. Figure 5-16(a) shows that the subject has installed two apps, *StayFitAtMIT*

and *CitySense* on the phone. The app *StayFitAtMIT* is collecting social proximity, activity, and location data. The app *CitySense* is collecting social proximity, activity, location data, and the screen status.

2. **Data Exploration** The exploration page provides several visualizations to help subjects interpret their data. Figure 5-16 shows examples of visualizations that are currently available in PDS. Box (b) displays the locations of a subject on a map with the timestamp of when the information was captured. It also shows the address of each location that links to the Google Street View service when clicked. Box (c) shows the visualization of the app usage information that is computed using the data collected by the *RunningApplications* component. The graph shows the percentages of how apps are used on subjects' phone. Box (d) shows the step count and average step count of a subject during the day.

Data visualization is a powerful tool to reveal trends and patterns in the captured data that can raise subjects' privacy awareness. For example, as shown in Figure 5-16(c), the subject might discover that her activities of preparing for a new job interview were revealed by the data. The visualization shows that the app "CodeRust", was the third most used app (102 minutes) last week which indicates that the subject spent quite a lot of time preparing programming interview questions. This realization of "what the app (the data collector) knows about me" can help subjects understand the privacy implications of disclosing their data to third parties. Subjects might then consider not sharing data that they deem sensitive.

If data transparency raises subjects' privacy awareness, then control over data disclosure becomes an important topic to the subjects. Chapter 7 includes a more elaborate discussion on the challenges of "selective disclosure" of personal data and its implications to both data collectors and subjects.

5.5 Application Server

Data collected from subjects is sent to the Application server and ultimately aggregated. The Application server has three major parts: 1) the ***Data Hanlder*** script that receives data sent from subjects' PDS and organizes it according to data types; 2) the ***Health Monitoring*** script that checks the health status of data collection tasks, and 3) the ***Messaging*** script that allows experimenters to send messages to subjects directly through the GCM service.

Similar to the implementation of Personal Data Store, the Application Server is built on top of storage space on Google Drive. Experimenters who use ContextProbe are not required to set up a web server and database on their own. They simply copy the deployment script to their Google Drive and deploy the script as a web application. The deployment script will create folders corresponding to the available sensor components on App Inventor. As shown in Figure 4-6 and Figure 4-7, raw sensor data is saved as Google Spreadsheet files and organized in different Google Drive folders according to types.

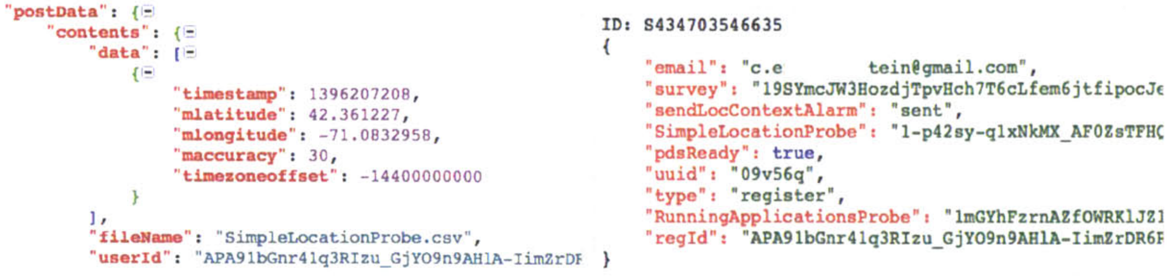
5.5.1 Registering Study Subjects

ContextProbe simplifies the registration process for both the subjects and the experimenters when starting the experiment. Without entering any personal information on the Application Server, subjects are guided to complete the registration on their mobile phone when they open the study app for the first time. As described in Section 5.3.6, the Application Server receives a GCM registration ID sent from the device and uses the ID as the identifier for subjects. The registration process is finished when the Application server receives the same ID after the activation of the data service in the PDS.

Subjects' information is stored on the Application as a JSON document and is accessible by the experimenter through a web query interface. As shown in Figure 5-17(b), this information includes email, registration ID, and identifiers of spreadsheets that store subjects' data.

5.5.2 Aggregating Collected Data

Newly available data is sent from the PDS as collection continues throughout the experiment. As shown in Figure 5-17(a), the data is sent in JSON format through a HTTPS POST method and it contains the types of data and the registration ID to identify the data source. When receiving the data, the Data Handler scripts process the content and use the



```
"postData": {
    "contents": {
        "data": [
            {
                "timestamp": 1396207208,
                "mlatitude": 42.361227,
                "mlongitude": -71.0832958,
                "maccuracy": 30,
                "timezoneoffset": -14400000000
            }
        ],
        "fileName": "SimpleLocationProbe.csv",
        "userId": "APA91bGnr41q3RIzu_GjYO9n9AHlA-IimZrDF
```

(a)

```
ID: S434703546635
{
    "email": "c.e        tein@gmail.com",
    "survey": "19SYmcJW3HozdjTpvHch7T6cLfem6jtfipocJe
    "sendLocContextAlarm": "sent",
    "SimpleLocationProbe": "1-p42sy-q1xNkMX_AF0ZsTFHC
    "pdsReady": true,
    "uuid": "09v56q",
    "type": "register",
    "RunningApplicationsProbe": "1mGYhFzrnAZfOWRKlJZ1
    "regId": "APA91bGnr41q3RIzu_GjYO9n9AHlA-IimZrDR6F
}
```

(b)

Figure 5-17: (a) Data sent from the PDS in JSON format (b) Example of a subject's information stored in the Application Server's database

combination of the registration ID and the type of data to determine the which spreadsheets will save which content.

For example, as shown in Figure 5-17(a), when new location information is sent from the PDS, the Data Handler script first queries the database using the identifier *userId* to find the corresponding spreadsheet file that matches the data type (*SimpleLocationProbe*), and saves the content to that file. As shown in Figure 5-18, a unique 5-character alphanumeric sequence is used to name all sensor data collected from a particular subject. The files are organized and saved in different folders according to their types in the application server.

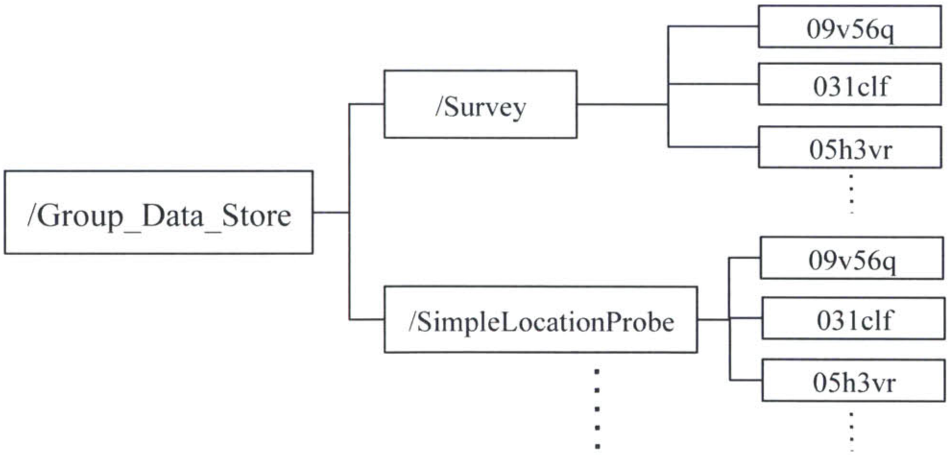
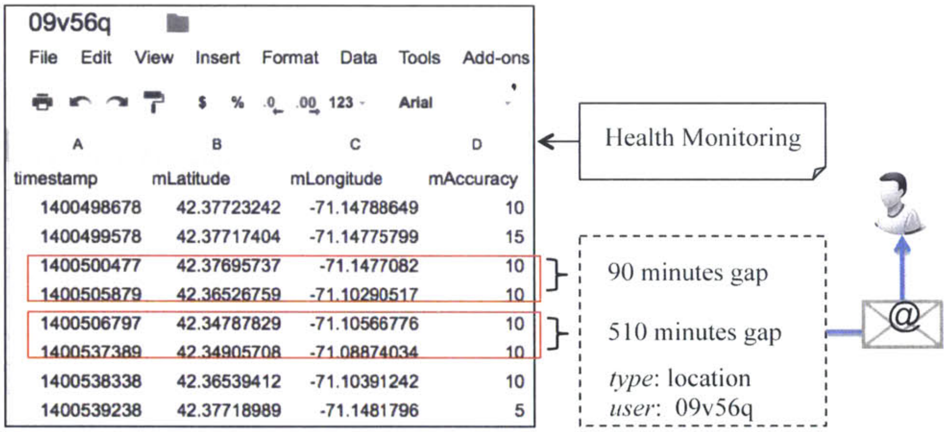


Figure 5-18: The structure of how collected data is organized in the Application Server

5.5.3 Monitoring Data Collection

Experimenters can configure the *Health Montoring* script through a web interface to monitor the status of all data collection tasks. Multiple issues can arise during the experiment. For example, mobile phones can be out of battery or unable to connect to the Internet to upload data. The Health Monitoring script checks the files to identify if there are any "gaps" within the collected data. For example, Figure 5-19 shows that there are two gaps within a subject's location data. In this example, the app would normally collect location data every 15 minutes; therefore the difference between two consecutive timestamps should be 900 (seconds). The result shows that the subject's mobile phone failed to collect or upload location data twice during the day (90 minutes and 510 minutes). These results are sent to the experimenter through email periodically.



A	B	C	D
timestamp	mLatitude	mLongitude	mAccuracy
1400498678	42.37723242	-71.14788649	10
1400499578	42.37717404	-71.14775799	15
1400500477	42.37695737	-71.1477082	10
1400505879	42.36526759	-71.10290517	10
1400506797	42.34787829	-71.10566776	10
1400537389	42.34905708	-71.08874034	10
1400538338	42.36539412	-71.10391242	10
1400539238	42.37718989	-71.1481796	5

Figure 5-19: Examples of monitoring data collection and finding 2 gaps in subject *09v56q*'s location data

5.5.4 Sending Pushed Messages

When issues of data collection arise, the experimenter can choose to email the subjects or use the *Messaging* script to directly send messages to the mobile phone. The Application Server provides a web interface for the experimenter to deliver GCM messages to the study app. The GCM service can do more than just deliver real time communication between experimenters and study subjects. Experimenters can use pushed messages to change the app's behavior. For example, the experimenter can change the duty cycle for data collection of one particular sensor component by pushing a new value of the time interval. And on the device's end, the *PctMessaging* component can receive the new value and reconfigure the interval. Another example of using the GCM service is the Contextual Privacy Preference study that will be discussed in Chapter 6. The GCM service is used to update the survey content on the study app. A list of app names computed from subjects' app usage data are pushed to the study app to update the survey questions for each participant.

5.6 Conclusion

The ContextProbe framework lowers the technical barriers for experimenters to collect sensor data and use existing cloud-based services to conduct ESM study. By extending the app building platform App Inventor, ContextProbe mitigates the experimenters' burden of developing mobile apps and setting up the infrastructure for the experiment. Subjects who participate in different experiments supported by the ContextProbe framework are given transparency tools to explore their own data. The architecture of PDS for data collection provides a foundation for implementing access control over disclosure of personal data in order to respect subjects' preferences. We will discuss the possible future direction in Chapter 7.

Experimenters can build context-aware apps to explore subjects' preferences that are affected by various contextual factors. Likewise, the experiment described in Chapter 6 uses ContextProbe to conduct a user study that uses the ESM approach to explore how the expressed purpose of data collection interacts with context in influencing people's preferences about disclosure of personal information.

Chapter 6

Purpose & Personal Context's effects on Privacy preferences

This chapter describes a user study, the Contextual Privacy Study, that was conducted using the ContextProbe framework. The aim of the study is to understand how the expressed purpose of data collection interacts with context in influencing people's privacy preferences. One surprising result of the study is that people chose to disclose more when purpose of data collection was omitted than when a vague purpose was presented. Using the ContextProbe framework, the experimenters created a mobile app (which is referred as "the study app" in the following sections) to collect subjects' *in situ* responses and generate personalized survey questions based on the collected data.

The chapter first presents the motivations for the Contextual Privacy Study, followed by the summarization of the study and its key findings. Then the chapter presents the design and implementation of the study with details of how it was executed. Next, the chapter reports results and discusses the implications of those results for the understanding of consumers' privacy preferences with regards to disclosing personal information. The study is a joint work with Ilaria Liccardi and it has been accepted to the Computer-Human Interaction (CHI 2015) conference.

6.1 Contextual Privacy Study

6.1.1 Motivation

As mentioned in Chapter 1, consumers today have to disclose different kinds of personal information when using mobile apps on their phone. However, when users opt in for services that require collecting personal information, they should not be expected to consent to data collection for every foreseeable purpose. For example, research by Liccardi [55] found that many mobile apps collect a variety of information, but the purposes for collection are often missing or incomplete in the disclosed privacy policy.

People's level of concern rises when they find data collection happening in a context that is unexpected. Such an experience leads to a sense of "creepiness" and results in loss of trust [80]. To ease tensions revolving around data collection, app developers are advised to follow the principle of "respect for context" [92] when harvesting user data. Perhaps to really ease people's privacy concerns, consumers should have control of their data in terms of how services can use it, rather than simply granting carte blanche access.

We are interested in how different circumstances might affect users' choices of disclosing their contextual information, specifically the location they are at and the activity they are performing when they are asked by the data collector. We wanted to investigate the effects on privacy preferences by varying two factors within real contexts: data collectors, and the purpose of data collection. In particular were are interested in these questions:

1. Do different type of purposes affects subjects' willingness to disclose their data?

2. Are people more likely to disclose their personal data with apps that they frequently use?

3. Does the type of the location context (e.g. home versus workplace) affect the willingness to disclose their data?

4. Does the type of the activity context (e.g. working versus socializing) affect people's willingness to disclose their data?

6.1.2 Summary and Key Findings

We applied the Experience Sampling Method (ESM) [21] to elicit subjects' responses to disclose their personal information in the form of their location and activities throughout the day. We use the *ContextProbe* framework to build a mobile app that presents survey questions customized to a subject's app usage and personal context.

We conducted a 4-week *in the wild* experiment, prompting subjects hourly from 9am to 10pm each day. Subjects were asked to describe their personal context, followed by questions asking about their willingness to disclose context information (their whereabouts or activities) to different apps installed on their devices, for different types of purposes. The apps were chosen based on each subject's frequency of use. The types of purpose include: 1) purpose unspecified, 2) for capturing information, 3) for system testing, 4) for improving user experience, 5) for advertisement, 6) for profiling users, and 7) for revenue needs.

The main findings of the study are:

1. Participants become **less willing** to disclose personal information when **more specific** information is provided. This information includes disclosing to which specific app and for what specific purpose will the data be used. Our results showed that when presented with **no information** about purpose and data collector, the average willingness to disclose is the highest.

2. The type of purpose affected a significant fraction (15/34) of participants' behavior about disclosing information. Participants were more willing to disclose for purposes that are beneficial solely to them.

3. The type of app, but not the frequency of use, affected participants' behavior about disclosing information.

4. Participants were least willing to disclose their information when their location context is *home*.

6.2 Approach

We conducted a user study over a four-week period to understand subjects' preferences towards disclosing their personal context to different apps for different purposes.

We collected subjects' most recent location and enquired about their activities, while also collecting information about the apps running on their smartphones. We asked subjects to answer a set of *personalized* questions on an hourly basis, based on their current location and activity. Each set of questions contained two phases, with different types of questions presented within each of them. The two phases were designed to:

1. **Gather users' personal context** in order to first understand each participant's current state. For example, where they were (home, work etc.), who they were with (friends, family, colleagues etc.), what were they doing (working, leisure etc.) and who were they willing to share this information with (friends, family, everybody etc.)

2. **Measure users' privacy preferences** in order to understand if the app, the type of purpose, and the type of location (for example whether they were home or at work) or activity (whether they were working, going somewhere etc.) affected subjects' willingness to disclose their personal context.

6.2.1 Study Design

The study was designed to sample subjects' responses for their disclosure preferences while preserving the ecological validity of the measurements by using the experience sampling method. The factors tested in the ESM study were the purpose for disclosure, and the actual app that would be receiving the information.

In other words, we want to capture subjects' experiences (making disclosure decisions) in their natural, everyday environments. To that end, the study adopted time-contingent sampling [23] to solicit subject-reported location annotations and answers about privacy preferences. Time-contingent sampling is a widely adopted protocol in ESM studies in which subjects' answers are solicited at specific times of the day.

Gathering subjects' personal context

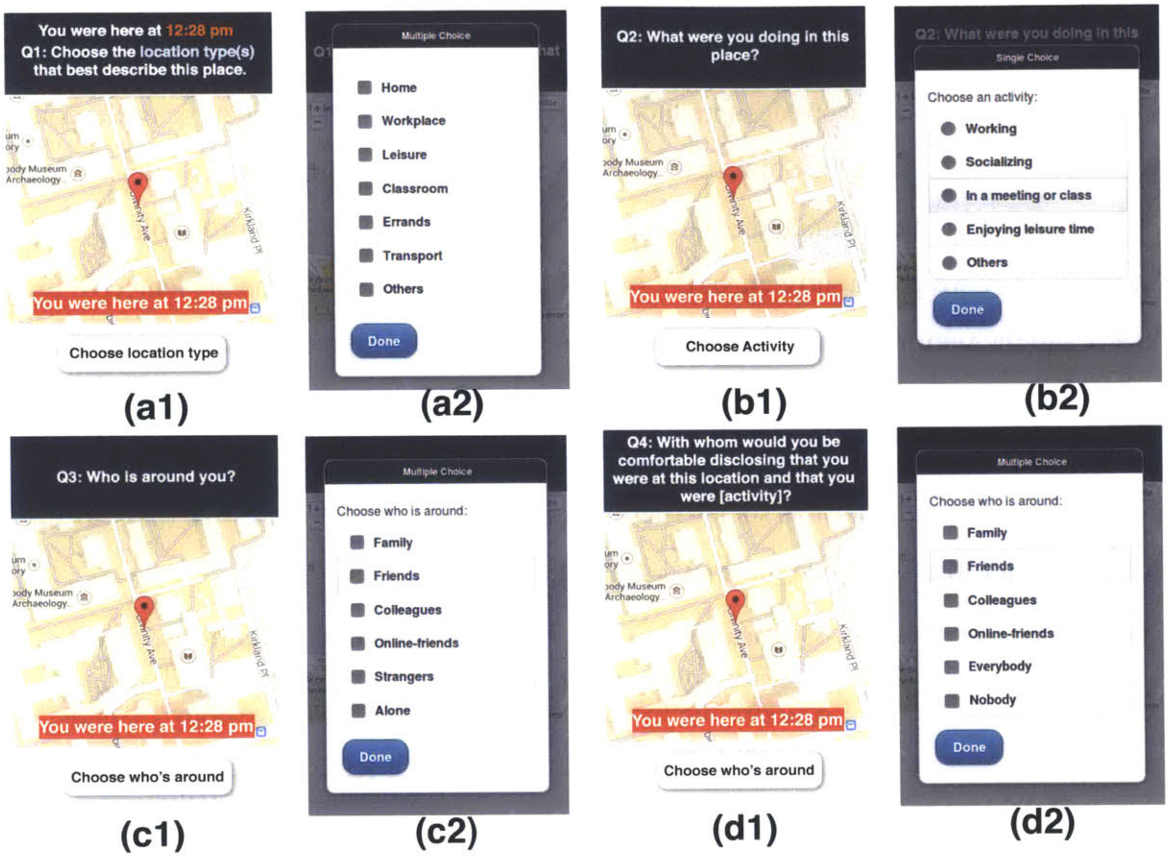
In phase 1, the questions were designed to obtain information about subjects' current context in the form of (a) the type of place they were located, (b) the activity they were engaged in, (c) who was around them, and (d) to whom they would have been comfortable sharing the activity they were currently engaging in. Table 6.1 lists the questions presented to the subjects and the options for annotations for each context type. Figure 6-1 shows snapshots of each qusetion that was presented to the subjects in phase 1.

The study app uses the answers provided by subjects to the activity type question (Figure 6-1 (c)) in formulating the social sharing attitude questions (Figure 6-1 (d)), so the activity type questions precede the social sharing questions.

Context type	Question	Options for annotations
Location context	You were here at [time]. Choose the location type that best describes this place	Home, Workplace, Transport, Leisure, Classroom, Errands, Others
Activity context	What were you doing in this place?	Working, Socializing, In a meeting/class, Enjoying leisure time, Others
Social proximity context	Who is around you?	Family, friends, colleagues, online friends, strangers, alone
Social sharing attitude	With who would you be comfortable disclosing that you were at this location and that you were [activity]?	Family, friends, colleagues, online friends, everybody, nobody

Table 6.1: Questions presented to subjects in phase 1

Another goal of the questions presented in phase 1 is to "prime" subjects to think about context before answering the privacy preference questions in phase 2. Previous research has shown that people's willingness to disclose information about personal context (e.g. whereabouts and activities) is affected by the privacy comfort level associated with a particular type of context information. For example, Khalil and Connelly [50] found that participants were least likely to disclose their location in context-aware telephony applications, in which callers are provided with context information about the receivers in order


You were here at 12:28 pm

Q1: Choose the location type(s) that best describe this place.

Q2: What were you doing in this place?

Q3: Who is around you?

Q4: With whom would you be comfortable disclosing that you were at this location and that you were [activity]?

Figure 6-1: User study questions designed to gather users' personal context in the form of (a) location type, (b) activity type, (c) social, and (d) social sharing attitudes

not to interrupt them. We wanted to use the questions in phase 1 to prepare the subjects to consider their privacy comfort level of disclosing their personal context before other factors that might affect their decision were introduced in phase 2.

Figure 6-1 shows how contextual cues such as time and locations are presented in the questions. The marker displayed on the map and the time label shown to the subjects are used as *reminders* to help subjects recall their most recent context. In addition, questions presented in phase 1 encourage subjects to interpret their location in terms of its social-cultural aspects that affect their answers for social sharing attitudes (see Table 6.1).

Measuring users' privacy preferences

Questions in phase 2 were designed to explore users' privacy preferences. In particular we were interested in testing familiar apps as well as purposes that have been commonly cited as reasons for collecting personal data. In order to measure subjects' preferences, we posed questions in which two factors (app name and purpose) were randomly and evenly selected. Figure 6-2 shows how questions were presented to subjects and where each value condition would appear.

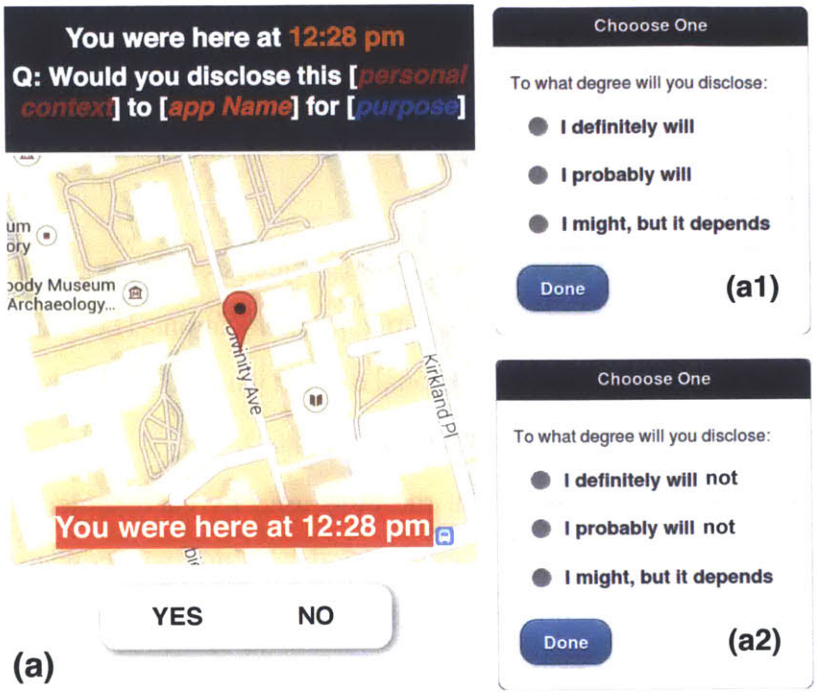
Personal context condition is set either to the subject's current location or the type of activity that the subject is currently performing at this location. The value (location or activity) is used to test whether subjects' personal context would affect their willingness to disclose the information (which is also context information).

App name condition is gathered from the apps used by each subject. Three apps were selected by computing an subject's app usage data which was collected in week 1 of the study. These three apps represented three types of app usage (low, medium, and high usage). We assume that the more a subject used an app, the more familiar they were with the app. One of three apps would be randomly selected and shown in the privacy preference questions to represent the entity that subjects' information would be disclosed to. Questions were also posed with the app name omitted, for a total of four variations. The value of app name condition is used to test whether app's familiarity would affect subjects' preferences.

For this particular study, we wrote a script to process the app usage data that was aggregated in the Application Server 5.5 at the end of week 1. The script selected the three candidate apps based on the level of app usage. Then we used the Messaging script on the Application Server to push the list of three app names to the study app for *app name condition* in the privacy preference questions.

Purpose condition is randomly selected from six different types of purpose commonly cited for accessing users' personal data. A question is also asked when the purpose is omitted; hence this field is blank (for a total of seven conditions). For example, if the purpose "improving experience" is selected, then the question as shown in Figure 6-2 would be:

> *"Would you disclose this [context information] to [specific app]* ***to improve the app experience for you****?"*


You were here at 12:28 pm

Q: Would you disclose this [personal context] to [app Name] for [purpose]

Chooose One

To what degree will you disclose:

● I definitely will

● I probably will

● I might, but it depends

Chooose One

To what degree will you disclose:

● I definitely will not

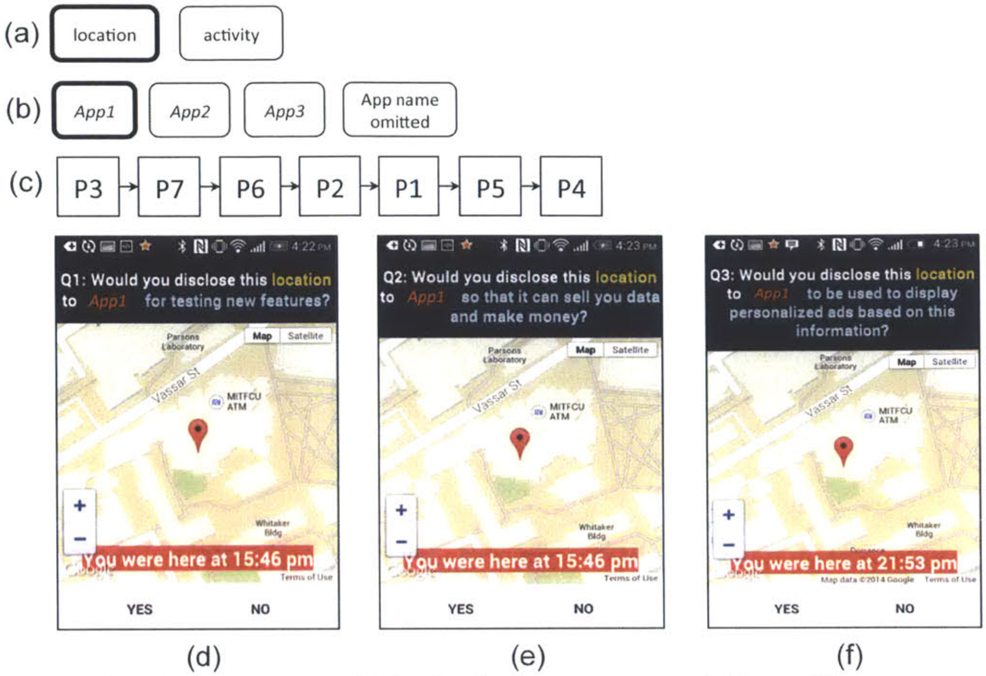
● I probably will not

● I might, but it depends

Figure 6-2: Privacy preference questions designed to gather subjects' willingness (or unwillingness) to disclose their personal context. Each subject had to choose YES or NO. Afterwards they had to give a rating of positive (a1) or negative (a2) feelings

The types of purposes used in the questions are:

- **Vague Purposes**:
  - Purpose omitted: nothing is displayed;
  - Captures information: "so that the app has your information";
- **User-focused purposes**:
  - Testing needs: "for testing new features";
  - Improving experience: "to improve the app experience for you";
- **User- and developer-focused purposes**:
  - Advertising: "to be used to display personalized ads relevant to you";
- **Developer-focused purposes**:



Q1: Would you disclose this location to App1 for testing new features?

You were here at 15:46 pm

YES NO

Q2: Would you disclose this location to App1 so that it can sell you data and make money?

You were here at 15:46 pm

YES NO

Q3: Would you disclose this location to App1 to be used to display personalized ads based on this information?

You were here at 21:53 pm

YES NO

(a) Randomly chosen context; (b) Randomly chosen app name; (c) A possible sequence of question, with P meaning purpose; (d-f) Samples of screenshots of the first three questions for this example

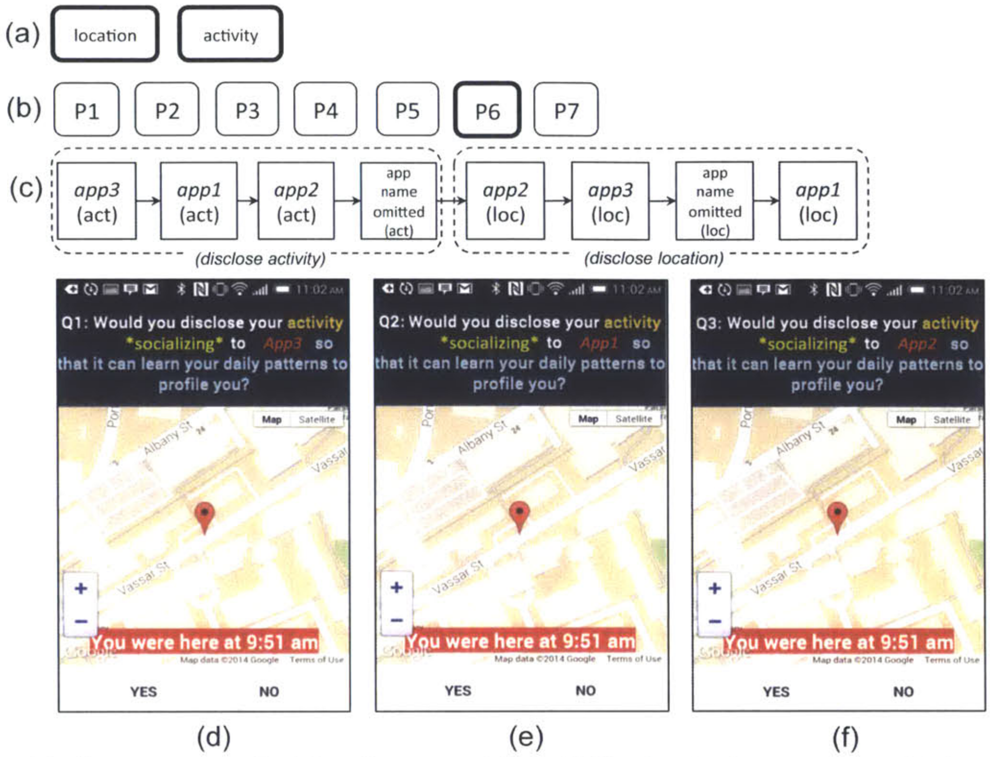
Figure 6-3: A sample of one possible permutation for randomization for app-specific condition

- Profiles: "so that the app can learn your daily patterns, to profile you for market research";
- Revenue needs: "so that it can sell this information and make money".

Randomization of privacy preference questions

In order to ensure that both *app name* and *purpose* factors were introduced in an even manner during phase 2, we introduced a scheme to randomize the appearance of different values for each factor. The privacy preference part of the questions was randomized between two conditions: the ***app-specific*** and the ***purpose-specific*** conditions.

When the app-specific condition was chosen, questions for all seven pre-defined purposes were created and presented randomly by the study app. As shown in Figure 6-3, the study app randomly would first choose a personal context to be used, either a location or an activity, from the information collected in phase 1. Then one app name was randomly selected from the list of app names. This app name was shown in all seven questions. The



Q1: Would you disclose your activity *socializing* to App3 so that it can learn your daily patterns to profile you?

Q2: Would you disclose your activity *socializing* to App1 so that it can learn your daily patterns to profile you?

Q3: Would you disclose your activity *socializing* to App2 so that it can learn your daily patterns to profile you?

(a) Chosen contexts (both location and activity); (b) Randomly chosen purpose; (c) A possible sequence of question, first set being questions for disclosing activity and second set for disclosing location; (d-f) Samples of screenshots of the first three questions for this example

Figure 6-4: A sample of one possible permutation for randomization for purpose-specific condition

purposes appeared in random order from question to question. Figure 6-3(d) to (e) show samples of screenshots of the first three questions that a subject would see in app-specific condition. For example, if *App1* was chosen and the context to be disclosed was *location* information, the seven questions would be similar to:

*Would you disclose this [**location**] to [**App1**] for [**Purpose**]?*

When the purpose-specific condition was chosen, 8 questions for choices of app names were created (4 for activity, and 4 for purpose). As shown in Figure 6-4, the study app first randomly selected one purpose from the seven-purpose list, then showed four questions for app names for location and four questions for app names for activity (the order was randomized between each set of four questions). The app names appeared in random order in each set. Figure 6-3(d) to (e) show the snapshots of the first three questions that a subject would see in the purpose-specific condition.

Even though the number of questions was different for each condition, the two factors (app name and purpose) were evenly distributed across questions for both conditions. During week 1, information about app usage was collected, so during this period the app name was omitted, and when purpose-specific condition was chosen, only two questions were displayed to subjects.

6.2.2 Experimental Procedure

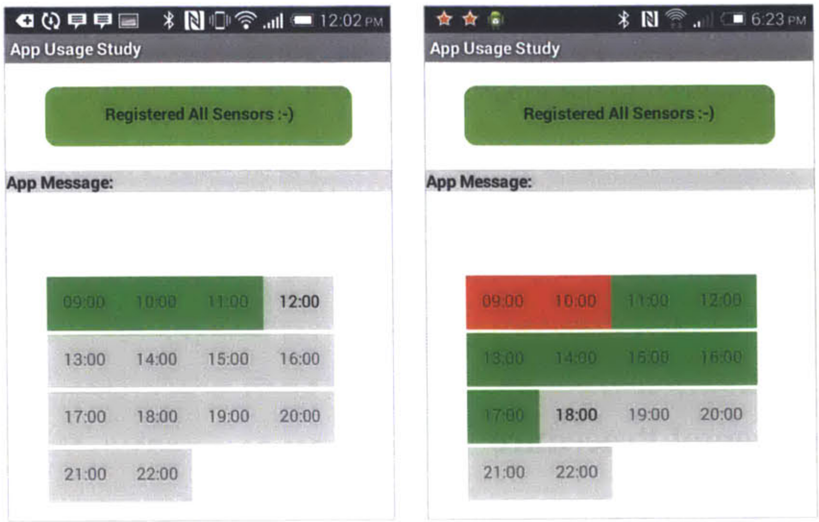
One-hour information session

Each subject was required to attend a one-hour information session prior to the start of the four-week study. This session was used to explain how the study app worked, to install the app on their smartphone, to allow them to sign the consent form, and to familiarize themselves with the study requirements and remuneration scheme.

User Study Requirements

Once subjects installed the study app and began the study, they had to answer a minimum of 10 sets of questions per day. Each set of questions contained two phases: 1) four initial questions, and 2) 7-8 questions depending on how the questions were randomized (see previous section). The first set was shown at 9:00 am, the last set at 10:00 pm, for a total of 14 sets a day. A notification was used to alert subjects that a new set of question needed to be answered. Subjects had up to three hours after the set was first shown to answer the questions for that particular time-window. We chose to set the limitation for answering a set of questions up to three hours, because we wanted to make sure subjects clearly recall their context.

As shown in Figure 6-5, subjects could view the study app's home page to see the sets that needed to be answered (marked in green), and sets that had expired (marked in red). The home page also showed greyed-out buttons indicating the next sets that were not yet available to be answered. If subjects did not respond to a minimum of ten sets per day, a warning email was sent to them by the researchers, to inform them that they were not



12:02 PM

App Usage Study

Registered All Sensors :-)

App Message:

🟩 09:00	🟩 10:00	🟩 11:00	12:00
13:00	14:00	15:00	16:00
17:00	18:00	19:00	20:00
21:00	22:00		

(a)

6:23 PM

App Usage Study

Registered All Sensors :-)

App Message:

🟥 09:00	🟥 10:00	🟩 11:00	🟩 12:00
🟩 13:00	🟩 14:00	🟩 15:00	🟩 16:00
🟩 17:00	18:00	19:00	20:00
21:00	22:00		

(b)

Figure 6-5: Examples of study app's home page (a) Subjects had answered all available sets (b) Subjects missed the opportunity to answer the first two sets of questions

complying with the requirements and that they could be removed from the study. After one warning, the subjects who failed to comply again were removed.

Subjects were asked to answer the questions as honestly as possible. They were notified in the info session that if random clicking was detected either during or after the study, they would not get paid.

Semi-Structured Survey

Once the data was analyzed, we created semi-structured questions for each participant. The questions were created using an online survey service. In the survey, we ask subjects to rate the importance of different factors (location context, activity context, time, purpose of data use, nature of the app) in making decisions for information disclosure. For after each factor, we also enquired about their motivations and reasons behind disclosing or not disclosing their context information based on this factor. For privacy-conscious, privacy-indifferent subjects, and trust-based subjects, we were particularly interested in their uniformed answers or their concerns toward specific apps. So we sent another round of questions through email to ask for their reasons behind privacy decisions. Below are the examples of questions we sent to these three groups of people:

1. **Trust-based subjects** For subjects whose responses showed strong preferences for disclosing, or not disclosing, information to particular apps. We included the mosaic plot similar to Figure 6-7 (c)) that summarizes subjects' answers and ask the following question.

   > *Could you tell us the reason why you would [never, sometimes, always] want to disclose your information to [App name]? Your data showed that no matter what purpose was presented to you, you always replied [Answer]. Could you give us your reasons behind this decision? Please do explain in detail. We are attaching a picture that shows your answers in the last few weeks in order to help you remember. The picture includes answers you gave for all the apps [App Names], but we are mainly interested in [App name].*

2. **Privacy-conscious subjects** For subjects whose responses were always "no". We included the mosaic plot similar to Figure 6-7 (a) that summarizes subjects' answers and asked the following question.

   > *Could you tell us the reason why you* ***would never*** *want to disclose your information in any case? Your data showed that no matter what purpose was presented to you, you always replied* ***NO****. Could you give us your reasons behind this decision? Please do explain in detail.*

3. **Privacy-indifferent subjects** For subjects whose responses were always "yes". We included the mosaic plot similar to Figure 6-7 (d) that summarizes subjects' answers and asked the following question.

   > *Could you tell us the reason why you* ***would always*** *disclose your information in all cases? Your data showed that no matter what purpose was presented to you, you always replied* ***YES****. Could you give us your reasons behind this decision? Please do explain in detail.*

6.3 Results

6.3.1 Participants

We solicited participation in our study using internal MIT staff and dorm mailing lists and Craigslist announcements. Participation in the study was voluntary and each participant received $120 after completing the whole study. 61 participants attended the one -hour information session. Three participants who came to the information session did not have an Android device capable of running the app. Seven additional participants who attended the information session never started the study. A total of fifty-one people started the study.

Fourteen participants were removed from the study. Twelve participants were asked to leave (5 participants in week one; 4 participants in week two and 3 participants in week 3) because their responses fell below the allotted threshold of 10 sets a day. Two participants were asked to stop the study in week one because we could not receive any new GPS location data even though they reported being in different locations. Thirty-seven participants participated in the entire study.

At the end of the study, three participants' responses were removed because due to their type of Android device, information about the apps running on their smartphone could not be collected for the study (they received payment); in addition, one had problems receiving questions due to a poor Internet connection around his neighborhood.

Thirty-four people successfully completed the study. Of these, 18 were male (avg. age = 25) and 16 female (avg. age = 29). Level of educations varied from having completed high school (4), two-year college degree (7), being an undergraduate student (2), completed four-year college degree (10), completed master degree (4), being a graduate student (3) to advanced graduate work or completed Ph.D. (4).

6.3.2 Responses

Over the period of four weeks we gathered 74,713 total responses from the 34 participants (Table 6.2). Questions were evenly distributed across the factors (Table 6.2). The 28 variations for the two factors (app name and purpose) can be grouped into four different

Table 6.2: Number of subjects' responses grouped by app frequency types for each purpose condition

TYPES OF PURPOSES	*App Name Omitted	Least Used	Medium Used	Most Used	Total responses
*Purpose Omitted	4,212	2,138	2,069	2,145	**10,564**
Captures Information	4,234	2,196	2,132	2,198	**10,760**
Testing Needs	4,181	2,131	2,059	2,135	**10,506**
Improving Experience	4,278	2,184	2,114	2,186	**10,762**
Ads Needs	4,198	2,110	2,048	2,121	**10,477**
Profiles	4,298	2,224	2,154	2,223	**10,899**
Revenue Needs	4,263	2,182	2,118	2,182	**10,745**
Total responses	**29,664**	**15,165**	**14,694**	**15,190**	**74,713**

Number of subjects' responses grouped by app frequency types (no app name condition is included) for each purpose condition (no purpose condition is included). The total response for each and across conditions is also shown.

conditions:

1. Purpose unspecified and app unspecified (4,212 responses)

2. Purpose specified and app unspecified (25,452 responses, excluding the *purpose omitted* condition)

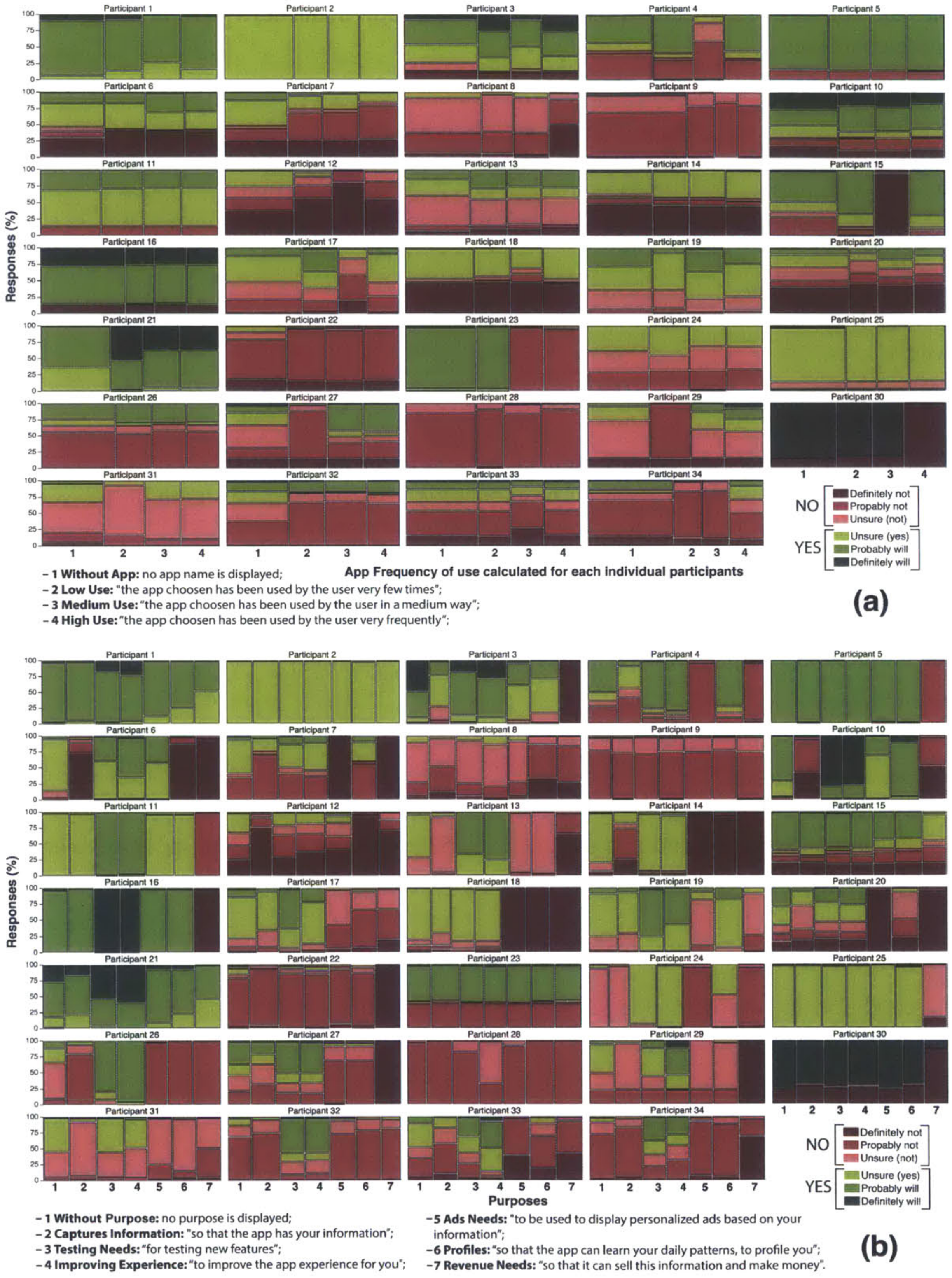
3. Purpose unspecified and app specified (6,352 responses, excluding the *app omitted* condition)

4. Purpose specified and app specified (38,697 responses, excluding the *app omitted* condition and excluding the *purpose omitted* condition)

6.3.3 Participant Privacy Profiles

We found that subjects disclosed the personal context according to different factors: usage purpose, trust in the app itself, and location type. From Figure 6-6 (a) & (b) we can see that six participants (P4[1], P15, P23, P27[1], P29[1], P30) disclosed mostly according to the

[1] These participants did not disclose their personal information when the purposes displayed within the question did not show a conceivable gain to them. Otherwise they based their decision on the app. Even when the purposes displayed within the questions were tailored to have a clear and conceivable gain for them, they still would not disclose their personal information with certain apps.

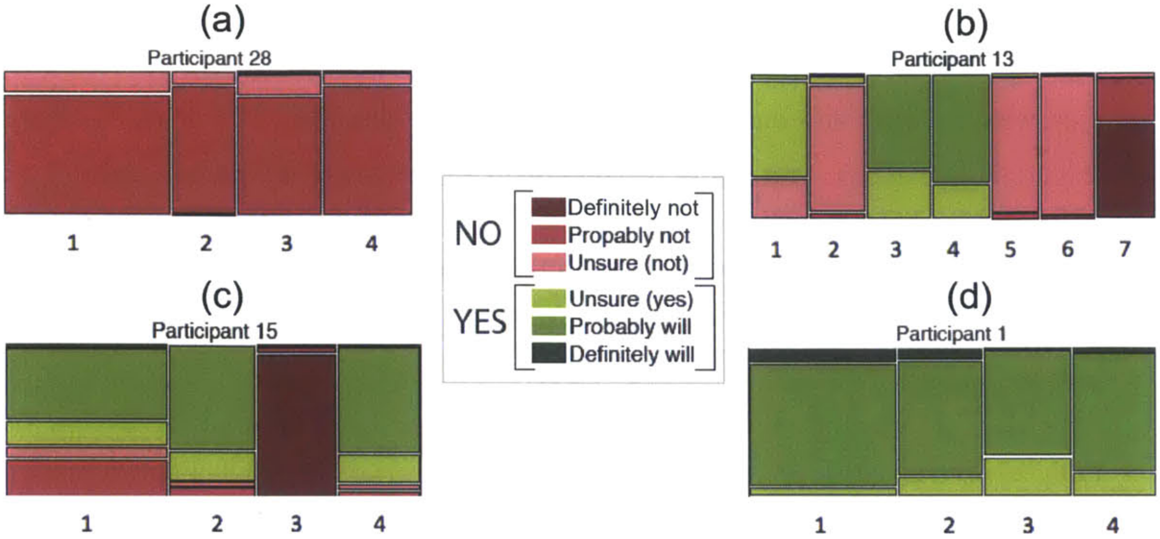


- **1 Without App:** no app name is displayed;
- **2 Low Use:** "the app choosen has been used by the user very few times";
- **3 Medium Use:** "the app choosen has been used by the user in a medium way";
- **4 High Use:** "the app choosen has been used by the user very frequently";

- **1 Without Purpose:** no purpose is displayed;
- **2 Captures Information:** "so that the app has your information";
- **3 Testing Needs:** "for testing new features";
- **4 Improving Experience:** "to improve the app experience for you";
- **5 Ads Needs:** "to be used to display personalized ads based on your information";
- **6 Profiles:** "so that the app can learn your daily patterns, to profile you";
- **7 Revenue Needs:** "so that it can sell this information and make money".

Overview of subjects' responses as to willingness or unwillingness to disclose their personal information with apps, in the form of either their location or their activities, for six different types of purposes as well purpose omitted.

Figure 6-6: Subjects' responses for the willingness or unwillingness to disclose their personal information



The mosaic plots are examples taken from Figure 6-6 for subjects' reponses of the willingness to disclose personal information (a) Privacy-conscious subjects, participant 28 (b) Purpose-driven subjects, participant 13 (c) Trust-based subjects, participant 15 (d) Privacy-indifferent subjects, participant 1

Figure 6-7: Examples of mosaic plots for subjects' responses of the willingness to disclose personal information

app that they were using (with the exception of never disclosing with any app only when specific purposes were specified). For example, as shown in Figure 6-7 (c), P15 would never disclose information to the medium used app.

The majority of participants (fifteen participants: P3, P5, P6, P10, P11, P13, P14, P16, P17, P18, P19, P24, P25, P26, P33) disclosed according to the type of purpose (Figure 6-6 (b)) displayed in each question. In other words, the type of purpose had an impact on the willingness to disclose.

These latter participants did not disclose their personal information when the purpose specified did not present any conceivable gain to them. For example, as shown in Figure 6-7 (b), P13 would tend to disclose when the purpose was for "testing needs" or for "improving experience". However, when the purpose was for profiling or for revenue needs, P13 would tend not to disclose.

For the remainder of the participants, eight participants demonstrated behavior of either always being willing to disclose their data (P1, P2, P21) or being extremely privacy-sensitive and never sharing (P8, P9, P12, P22, P28). The remaining five participants (P7, P20, P31, P32, P34) showed no effect according to the type of purpose or the app; a closer look at their data showed that they tended not to disclose their personal information when at particular locations. Among these participants, four participants tended not to disclose when they were at *home* and one participant tended not to disclose when at *work*. Our results highlight five distinct patterns of behavior with respect to privacy preferences:

- **Privacy-conscious subjects (3)**: These subjects did not tend to disclose their personal data. They were very conservative about disclosing information and they seldom disclosed.
- **Purpose-driven subjects (15)**: These subjects disclosed their personal data according to different purposes. They tended to disclose data when the claimed purpose includes a benefit to them.
- **Trust-based subjects (6)**: These subjects trusted certain apps and, no matter the activity or the purpose, would disclose the data. On the other hand, for apps that they didn't trust, they would not disclose the data.
- **Privacy-indifferent subjects (5)**: These subjects disclosed their personal data because they saw no problem or harm in doing so.
- **Location-sensitive subjects (5)**: These subjects disclosed their personal data when they were not in a particular location. When they were at a specific location, they tended to never disclose.

6.3.4 Usage and Collection Effects on Preferences for Information Disclosure

Table 6.3 shows the general trend in subjects' answers for the willingness to disclose personal information. Subjects tended to disclose more ($\mu = 3.12$; $\mu_{rank} = 3.04$) when there was no information on the purpose for data collection or which app was collecting it (combination 1) (Table 6.3). When details were given, such as name of the app requesting the information and the purpose for which the data would be used, subjects tended to disclose the least ($\mu = 2.71$; $\mu_{rank} = 1.60$; combination 4). When either the purpose or the name of the app was shown, subjects tended to disclose less when data purpose was specified ($\mu = 2.86$; $\mu_{rank} = 2.50$; combination 3) than when the information of the app was specified ($\mu = 3.02$; $\mu_{rank} = 2.85$; combination 2).

Table 6.3: Descriptive statistics and Friedman test for each factor combination of absence and/or presence of purpose and app name combinations

	COMBINATIONS		**Number**			**Mean**
	APP		**of**	**Mean**	**Std. Dev**	**Rank**
	NAME	PURPOSE	**Responses**	(μ)	(σ)	(μ_{rank})
1	✗	✗	4,212	3.12	0.61	3.04
2	✓	✗	6,352	3.02	0.65	2.85
3	✗	✓	25,452	2.86	0.69	2.50
4	✓	✓	38,697	2.71	0.69	1.60

[a] The number of responses for each combination is shown. The mean is calculated using a 5-point Likert-scale (with 5-Will Definitely disclose to 1-Definitely NOT disclose). The Friedman test is significant Chi-Square (3, N of Participants = 34) = 25.295, $p < .0001$. Kendall's W is 0.248, indicating fairly strong differences among the four groups.

[b] Post hoc analysis with Wilcoxon conducted with a Bonferroni correction with $p < 0.0083 (p = 0.6/6)$ shows a statistically significant in users' sharing preferences in G1 vs. G3 (Z=-3.198; $p < .001$), G1 vs. G4 (Z= 3.676; $p < .000$), G3 vs. G4 (Z = -3.258; $p < .001$), and G2 vs. G4 (Z = 3.771; $p < .000$)

App Effect: Does the Frequency or the Type of App Matter?

Six participants based their decision to disclose or not disclose their personal information according to the app they were using.

The frequency of use of the app does not affect willingness to disclose data with that app, with some participants choosing to disclose more with less frequently used apps, rather than the most frequently used apps. P15 and P4 indicated not wanting to disclose personal information with a medium usage app. For these participants, the reasons for not wanting to disclose depended on the type of app. P4 stated being afraid that personal sensitive information could be used for other purposes by the app (web browser app):

> P4:*"I tend to type in a lot of personal stuff via [browser's name] that I don't want used. That is why I denied its access to my information."*

P15 showed to be reluctant in disclosing personal information with an (dating) app that has already collected and stored a lot of personal information.

> P15:*"I was just getting a little aggravated with the site [Name], it already had a ton of my data, so no more!"*

P4 tended not to disclose personal information (with any type of app, including when the app name was omitted) when purpose 6 (target advertising) and purpose 8 (revenue needs for company) were shown.

P27 and P29 tended not to disclose their personal information with their least-used app, however the reasons for these choices are different. For P27, frequency of use and hence familiarity with the app (diet app) was an important factor. P27 explained that once a relationship with an app has been established, the subject becomes more willing to disclose personal data, for certain types of purposes that are beneficial.

> P(27)*"[...] I didn't use the app, I was not able to establish a relationship with it"*

P29 explained not wanting to disclose based on distrust in the app (banking app) itself:

> P29*"I feel it's an extreme invasion of privacy letting my bank know where I am during the day. My phone is for leisure, for fun [...]. The less the bank knows about me, the better."*

Similarly to P4, P27 and P29 also tended not to disclose with any type of app (including when app name was omitted), when purposes did not show any conceivable gain to them (purpose 2, 5, 6, 7).

P23 and P30 instead were unwilling to disclose their personal data with their mostly frequently used apps. P30's most frequently used app is a social network app and similarly to P15, the reason behind not wanting to disclose was not to give additional personal data to the data already collected and stored by the app. P23 explained concerns that a former employer (a bank) would be able to use the personal data captured:

> P23*"I used to work for [Company Name] and I would just not want them to have any of my information."*

Participants did not base their decision to disclose or not disclose with specific apps on the frequency, but on the nature of app itself.

We found that participants denied access to their location to apps which actually requested the location permission on installation. This means that these apps can already collect data on the user's location. These participants were not aware that this could be happening, underlining the need for clearer and more specific regulation of data purpose rather than data access. Such conflicts are, however, not being addressed enough today by mobile platforms.

Purpose Effect: Does the Type of Purpose Matter?

Data usage and collection matter to subjects when deciding to disclose their information (Table 6.3). As we can see from Table 6.4, Figure 6-8 and Figure 6-6 (b), the type of purpose affects subjects' decisions to disclose or not disclose their personal context.

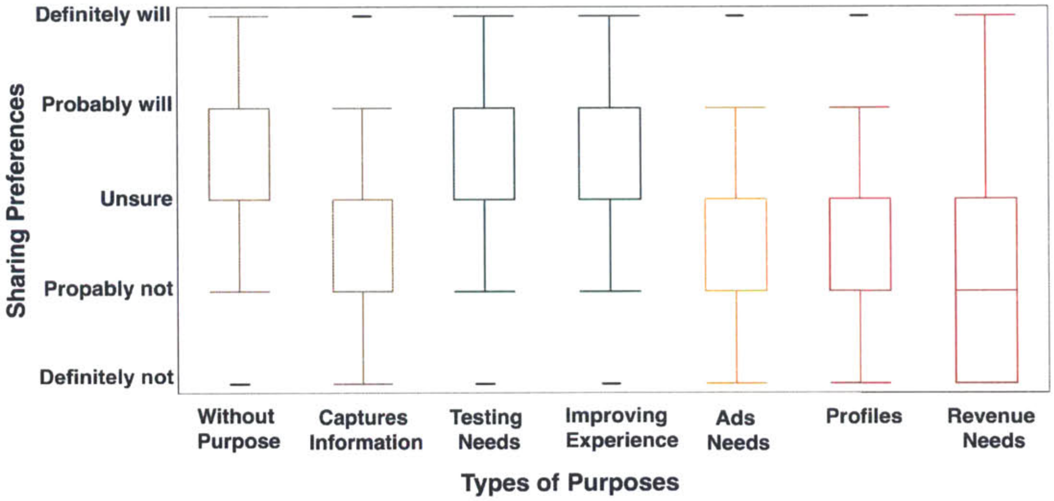


Figure 6-8: Box Plot of participants' responses of disclosing their personal context (location and activity) grouped by different types of purposes

Participants showed willingness to disclose their personal context when there was a conceivable gain to them. When participants were prompted to disclose their personal data for purposes of *testing needs* and *improving app experiences*, their disclosing preferences had a $\mu = 3.324$ and $\mu = 3.36$ respectively (Table 6.4). When the purposes specified were more beneficial for developers (i.e. companies), the willingness of participants to disclose dropped, with $\mu = 2.60$ (*ads needs*), $\mu = 2.66$ (*gathering profiles*) and $\mu = 1.96$ (*revenue need*).

Table 6.4: Descriptive statistics and Friedman test in willingness or unwillingness to disclose for different types of purposes

			Mean	**Std. Dev**	**Mean Rank**
	TYPES OF PURPOSES	**N**	(μ)	(σ)	(μ_{rank})
1	Purpose omitted	34	3.06	0.59	4.94
	Captures Information	34	2.73	0.71	3.28
2	Testing Needs	34	3.33	0.71	5.9
	Improving Experience	34	3.37	0.71	6.06
3	Ads Needs	34	2.61	0.88	3.29
4	Profiles	34	2.66	0.88	3.09
	Revenue Needs	34	1.97	0.92	1.44

[a] Purposes vary from seven predefined purposes: purposes in group 1 were vague, or without purpose; in group 2 were beneficial to subjects; in group 3 were beneficial to both subjects (ads relevant to him/her) and developers; in group 4 were solely beneficial for the developers.

[b] The Friedman test is significant Chi-Square (6, N = 34) = 126.794, $p < .0001$. Kendall's W is 0.622, indicating fairly strong differences among the seven purposes

When the purpose specified was non-explanatory and non-existent, participants' willingness to disclose their data differed (Figure 6-8), even though these two purposes describe the same circumstances [52]. When the purpose was non-explanatory and vague, $\mu = 2.74$, while when the purpose was missing, $\mu = 3.05$. This underlines the fact that participants are more willing to disclose data when they are not aware of what their data is used for.

When participants were asked to make decision to disclose their personal context information, they were more conservative towards the app than people. Figure 6-9 shows that even when participants were willing to disclose their personal information with a large group of people or with everyone, they were still not willing to disclose it for purposes which are mostly beneficial to developers (Figure 6-9).

The appearance of a vague and non-explanatory purpose caused participants to disclose

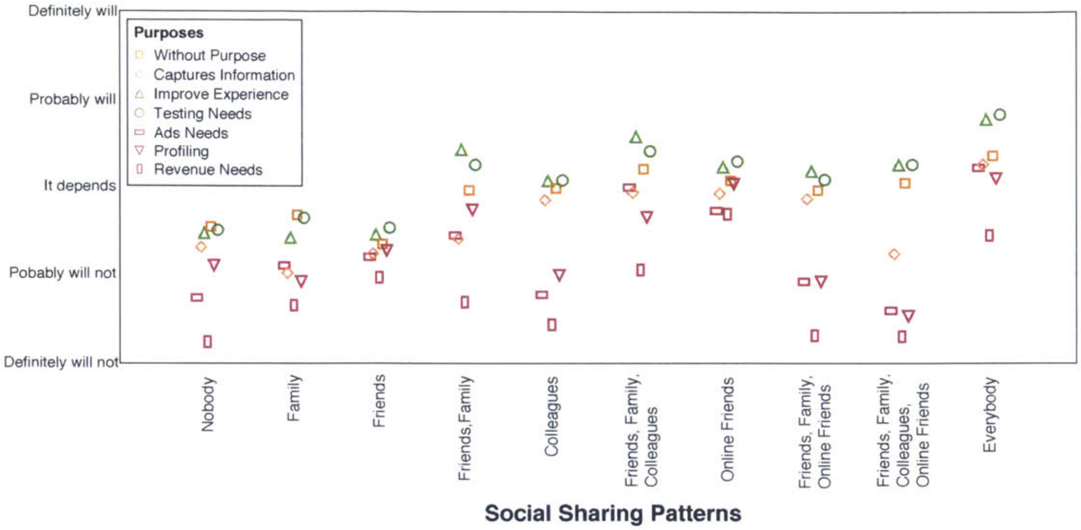


Figure 6-9: Participants' responses of disclosing their personal information with their social connections and with apps grouped by different types of purposes

less compared to when the purpose was missing. Participants might have been alerted by the vague purpose shown, indicating the intention to keep participants' data for other unknown uses (*so that it (the app) can have your information*). P10 underlines the importance of purpose information.

> P10*"The purpose of the data collection is very important to me. If it is just collecting it to store, I would not be comfortable because I wouldn't know what it is doing with the data"*.

Similarly, P7 and P29 both had concerns about unknown uses of their personal data.

> P29*"I am more sensitive to disclosing data that may have personal information that can be intercepted or used without my permission."*

> P7*"I am not about volunteering information to unknown sources, [...] Just because an app is particularly useful doesn't mean I would grant it a blank check to record and sell my personal data."*

6.3.5 Context's Effects on Disclosing Preferences

Most of the participants were affected by the type of purpose and the type of apps shown to them. However, for location-sensitive subjects, we found that the probability of not disclosing personal information different significantly across different location contexts. Participants showed different levels of (un)willingness to disclose their data which depended on their current location. As shown in Table 6.5, some participants were much less willing to disclose the location *home* than other places, but some were the opposite. Out of the five location-sensitive subjects, 3 subjects (P20, P31, P34) showed that when their location was *Home*, the odds of not disclosing was about 2 times greater than when in other locations. The other two subjects (P7, P32) showed that when their current location was *workplace*, *transport*, or *leisure* the odds of not disclosing were between $1.2 \sim 3$ times larger than *Home*.

Table 6.5: Repeated measure logistic regression with *location* context as the predictor variable and participants' disclosure preference as binary variable ($0 =$ NOT DISCLOSE; $1 =$ DISCLOSE).

PART. ID	INTERCEPT	*Home*	*Work*	*Leisure*	*Transport*
20	0.46***	0.54***	1.40*	1.34	1.05
31	0.51***	0.49***	2.78***	0.66*	1.11
34	0.22***	0.43***	0.94	0.96	0.98
7	0.66***	0.81+	N/A	0.22***	0.53***
32	0.35***	1.04	0.82	0.76	0.62*

[a] The odds ratios for each context is displayed. The computation is done per individual participant.

[b] The intercept value represents the bias of each participant to respond positively or negatively to disclose their information. The higher the value, the more likely they are to respond (irrespective of context).

[c] (***) $p < 0.001$; (**) $p < 0.01$; (*) $p < 0.05$; (+) $p < 0.1$.

Participants' remarks from the semi-structured interview revealed how they considered these factors together. P10 thought it was fine to disclose his locations if used appropriately:

> P10*"[...], I would be fine with an app knowing that I am relaxing in [coffee shop] or at home if it actually put that information to use"*.

P7 and P20 noted that the sensitivity of disclosed information was related to personal context, but also emphasized that purpose is the main factor for disclosing information:

> P7*"[...] whether or not to release my location depended on the other factors. [...] it's easy to say "I'm enjoying leisure time" sure, so are 1 billion other people, but having my location is a *lot* more specific, and so I'm less inclined to share that data [...] Purpose of use and nature of app are both extremely important factors in deciding whether or not to disclose the information.*

> P(20)*I don't like having my home location or home activities available to any app, [...] while I don't mind anyone knowing if I'm running errands or at work. [...] Purpose of use was the most important thing - if it was just for the app developers' benefit [...] (making ads, getting personal data) I didn't want to do it. But if I could get something out of it - improved experience - I was more willing to do it.".*

This observation showed that while users' locations and activities might be helpful in governing the social norms of disclosing information with other people [65], these factors are outweighed by purpose when considering disclosing information to apps.

In Android apps, users grant permission prior to installation, allowing apps to use personal information for a variety of unknown purposes. In this study, we have shown that participants based their decisions to disclose their personal information on the types of purpose. The all-or-nothing decision that the Android OS currently uses, should therefore be improved to allow finer-grained control.

6.4 Discussion

The contribution of this study is to examine smartphone users' behavior of disclosing context information in contextually-situated settings. Using the ContextProbe framework, we were able to create an app for ESM that collects subjects' *in situ* responses and generates personalized survey questions based on the collected data.

Overall, we found that the **more specific** the information presented to subjects for data collection, the **less willing** subjects were to disclose their information. Surprisingly, subjects were most willing to disclose their information when **no information** about who is collecting and for what purpose.

While previous research has shown that context plays an important role in shaping people's privacy concerns (see Chapter 2), our results found that subjects *were not aware of context* in thinking about disclosure. In contrast, when subjects were presented with purposes of data collection, the decisions for information disclosure were much more affected by whether the purpose was perceived as beneficial to them.

6.4.1 Privacy Awareness: Vague vs. Explicit

Our study highlighted the importance of showing *specific information* regarding information about data collectors (*app name*) and data usage (*purpose*). As seen in Table 6.3 participants are more willing to disclose their personal context when none of the information (purpose and app name) is displayed. When the appearance of this information is alternated, participants disclosed less than when information about data usage (*purpose*) was displayed.

The appearance of vague and non-explanatory purposes caused participants to disclose less, compared to when a purpose is missing. Tan et. al [84] reported that users were more willing to disclose their personal information when a purpose was shown in the permission request. This disparity in results might be due to different approaches we took in probing privacy preferences. Tan et. al [84] used an online survey, showing screenshots of permission requests from real apps with hypothetical question about personal data. Whereas our study was conducted in the wild with privacy preference questions about participants' real, personal and current context information tailored to specific apps that were used by each participant. Subjects in Tan et. al's study [84] might not have been familiar with the apps shown in the survey, and each subject was given the survey questions once.

In our study, each participant's answers were collected using repeated measures designed to cover all conditions for different purposes. The main goal of conducting the study in the wild with repeated measures was to compare subjects' responses under different conditions, including their physical context (their locations and activities) and different purpose presented to them. The aim was to trigger privacy concerns that are more subjective and sample responses from real life situations.

One plausible explanation for the impact of showing a non-explanatory purpose is that privacy awareness was increased. Participants were reminded of the trade-offs [40] between the unpredictable costs (privacy risks) and benefits (functionalities) brought by apps, therefore becoming less willing to disclose personal information. This finding highlights the importance of specificity when describing purpose of data access since a vague purpose can alert users to potential privacy risks and discourage them from disclosing.

Developers are able to collect personal data about users, because to date, mobile platforms lack support for fine-grained control over data collection with specified purposes. Our study suggests that when any explanations, if vague, are provided for the purpose of data collection, people do become alerted to privacy concerns and make different decisions compared to when no information is displayed.

6.4.2 Purpose Matters: Give Me a Reason to Disclose

Differently from previous research [88][50][86] which found that users' preferences for disclosing their locations are contextual, our study demonstrated that users' decisions about whether to disclose context information are affected not only by the sensitivity of the disclosed information itself, but by the purposes for collecting the data.

In fact, in our study, users' contextual information (current location, activity and social surrounding) did not have a significant impact on the decision to disclose personal information, when compared with the impact of knowing the purpose for which information would be used.

Our results have shown that when control is given to participants, they tend to make more specific choices regarding their perceived benefits. For some participants, preferences for disclosing can be strictly app-specific or location specific. However, even in these cases, participants were unwilling to disclose their information for some purpose types such as purposes for *profiling users* and for *revenue needs*.

Our findings confirmed that when purpose was presented to participants, they were alerted that the data might be used for additional purposes. Participants tried to interpret the potential use of the data based on the purpose string, and from there they would justify whether such data collection is reasonable or not.

We analyzed all 34 participants' apps and found that many (85%) of the apps that participants denied access to their location, actually requested the location permission on installation, meaning that these apps collect subjects' locations anyway. This underlines the drawback and failings of current practices of allowing any access to users' personal data without being able to specify and restrict the usage. Such conflicts are, however, not being addressed enough today by mobile platforms and should be addressed with regulations and legislation aimed at helping users safeguard and protect their right to privacy.

6.5 Related Research

Research has shown that many smartphone users lack the knowledge to perform changes in privacy control settings and mistakenly trust the app will protect their data privacy [68]. Felt et al. [33] interviewed users to evaluate their understanding of Android permissions. Their results revealed that only a few users actually read and understand the implications of permissions requested by apps. On the other hand, users may still hold unrealistic belief about how their data should be treated. Urban et al. [87] studied people's privacy attitudes towards data collections by apps and showed that many people consider information on their phone to be private and overwhelmingly reject several types of data collection. Given the disparity between users' privacy expectations and non-transparent practices of data collectors, researchers sought to understand and address the privacy issues in the context of consumers' app using experience, such as app-selection decision [49], privacy expectation [56] and data leakage [10].

One way to address users' privacy concerns is to model their privacy expectations while using the apps. Shilton et. al conducted a scenario-based survey with 979 subjects to rate over 39,000 hypothetical vignettes [79] of how apps collect and use their data. They tested contextual factors including *who* (the data collector), *what* (type of disclosed information), *why* (application purpose), and *how* (use of data by data collector). Their results showed that scenarios of data harvesting of apps in different contexts often do not meet users' privacy expectations. In fact, researchers found that apps transmit sensitive data that users intend to use on-device only to third parties [28]. The misappropriation of user data violates the contextual integrity as defined by Helen Nissenbaum [65], disregarding user personal context and expected information flows. On knowing how apps violate their privacy, users developed a sense of discomfort and lose their trust on the app [80].

Besides privacy concerns, past research also explored the motivations of why people disclose their information. Acquisti and Grossklags described the privacy decision-making process as the trade off of long-term privacy for short-term benefits [3]. Their experiment showed that people's willingness to disclose their personal information increase when they perceive beneficial gains such as monetary rewards [4]. Others showed that users disclose more when they find justifications to do so. Hurwitz found that high relevance of the service increases users' acceptance of data collection [43].

Disclosing decisions also relates to the subjective evaluation of personal information [1]. Carrascal et al. used an auction game with refined ESM [21] to study how much they value personal information when browsing online. The result showed that users value their offline identities three times more than their browsing behavior. Staiano et al. [83] studied users' valuation of their mobile data using Day Reconstruction Method [47] and showed that users value location data the most. The evidence from prior research suggests that values or even concerns of disclosing personal information are context specific. Thus, it is important to "sample" through users' experience and draw users' responses in different situations.

In the context of smartphone permission request, messages presented to users can actually affect users' decisions for disclosing personal information. A recently study done by Tan et al. [84] found that purpose strings shown in the permission dialog have impacts on users behavior. Participants were more likely to approve data requests when purpose string was displayed whereas when purpose string was empty they approved less. Our work has a similar interest in exploring the effect of purpose but targets a wider range of factors (context, purpose and app) with a different methodological approach.

An alternative approach for modeling users' privacy preferences is through crowdsourced method. Toch implemented Super-Ego [85] that used crowdsourced method to predict each user's privacy preferences for location sharing. Lin [56] used crowdsourced user surveys to measure the unexpectedness of certain data accesses by the apps. The study let users rate the app by comparing their perceived app functionality with the actual permissions that are requested by the app. While crowdsourcing makes their approach more scalable, crowd opinions only represent users' *a priori* preferences of how an app should work. The results might not reflect users' practical privacy concerns when they actively engage in making "privacy vs. benefit" trade-offs [40]. Our work introduces purposes of data collection to obtain user responses that serve as indicators of how much users value privacy over benefits they obtain from disclosing data to app services.

6.6 Limitations of the Experiment

When using the experience sampling method to probe participants' privacy preferences, we were aware of the natural bias introduced by the time-based triggers as discussed in [54]. For example, *home* and *workplace* were the predominant places for location context. Therefore we carefully reported the results with appropriate statistical indicators such as odds ratios to show the effects of location context.

Our understanding for subject location and activity context depends on self-reported data from participants. It is possible that some errors were introduced when annotating the locations. Our study required the participants to respond to questions fairly frequently (once an hour). It is possible there could be a fatigue effect that caused decays in response rate or lesser quality of data as a result. Due to the monetary incentive and weekly removal of disqualified subjects, we found only a slight decrease in response rate in the last week (5%). They study was conducted with Android smartphone users and might not be completely representative for other smartphone operating systems.

Chapter 7

Conclusion and Future Work

This chapter summarizes the important contributions of this thesis and presents future research directions.

7.1 Thesis Summary

Viewed as the most disruptive technology in the 21 century [60], mobile Internet has changed almost all aspects of our daily life. In this "always on" and "always connected" culture, consumers are starting to appreciate that the cost of economic convenience is the erosion of privacy. To address consumers' rising concerns over data collection on mobile phones, one dominant approach is to offer people more control over disclosure of personal information on their mobile phones through user-preference mechanisms. However, modeling people's privacy preferences is hard, if not impossible, due to the malleability of people's privacy preferences, which can easily change in with changing situations. Thus, simply looking at one-time answers from survey results can at best provide only a snapshot of people's privacy attitudes, but nothing more. Privacy research needs a *better approach to accurately and realistically probe people's preferences*, by taking into account the ubiquitous nature of the mobile environment. The contribution of this thesis lies in filling this gap, by identifying the crucial elements for soliciting people's privacy preferences on mobile apps and introducing a framework that enables experimenters to conveniently conduct this kind of study. More specifically, this thesis has made significant contributions to the current body of privacy research in the following aspects.

First, this thesis demonstrates the importance of presenting *contextual* information to study subjects when soliciting people's privacy preferences. Based on Nissenbaum's theory of contextual integrity, Chapter 3 shows how revealing apps' data collection behavior within the app usage context (idle vs. active) could raise people's privacy awareness and realign their privacy expectations. Through the lens of the app usage context, similar apps can be distinguished by looking at their data access behavior. Apps that access data while their users are not using the app (out of context) would be given a higher intrusiveness score.

Second, this thesis introduces a framework, ContextProbe, that lets experimenters capture *user context* at the time of information disclosure and solicit privacy preferences *in situ*. ContextProbe allows experimenters to conveniently conduct ESM studies for investigating privacy issues in the mobile environment. By leveraging data captured from on-board sensors on mobile phones and self-reported input, ContextProbe adopts a limited notion of context to help experimenters investigate the contextually grounded reasons behind information disclosure. Different than other frameworks, ContextProbe allows experimenters to create mobile apps tailored to their privacy studies and remove the overheads of setting up backend web servers and databases.

Third, this thesis provides evidence that asking privacy questions *in situ* produces more realistic and unbiased answers than the survey-based approach. Using ContextProbe, the Contextual Privacy study in Chapter 6 probed subjects' privacy preferences arising from their real experiences of using mobile apps. Surprisingly, results drawn from *in situ* responses are the exact opposite to previous survey-based approaches on the effect of apps' showing their purpose strings when requesting personal information: showing less information seems to result in greater willingness to disclose. More importantly, the study showed that personal context is not the only factor that affects people's behavior regarding information disclosure. For a significant fraction of participants, external factors such as *purpose of data use* and *trust in the app* outweigh *user context* when considering information disclosure.

7.2 Messages for policymakers and platform owners

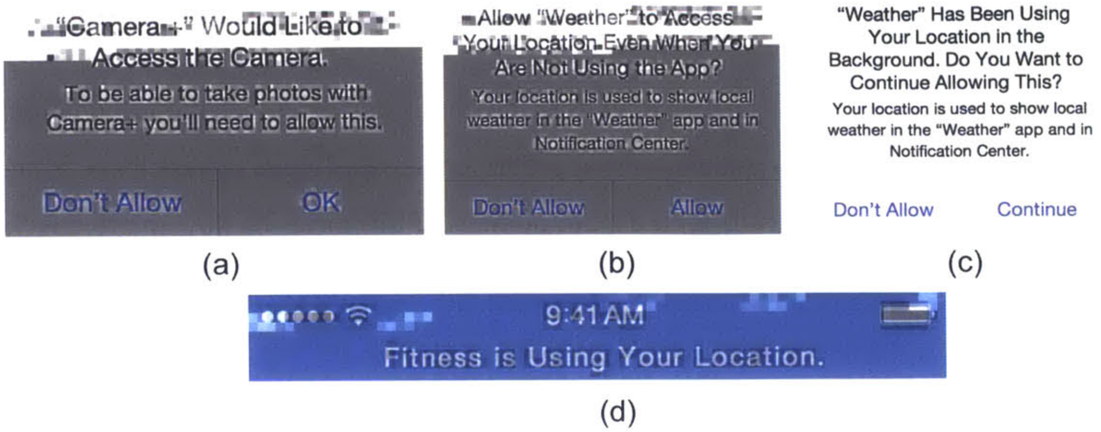
This section revisits the key research findings of the thesis and discusses the implications for policymakers, platform providers, and app developers. Different stakeholders in the mobile app ecosystem can learn from the main findings of this thesis on how to effectively improve the privacy of consumers while they are using mobile apps.

7.2.1 For Policymakers: Pay Closer Attention to Purpose

Research in this thesis shows that although user context[1] plays a significant role in affecting people's privacy expectations, the external factors *purpose of use* and *trustworthiness* of the data collector often outweigh context when people are making their privacy decisions. Findings in this thesis show that subjects were not aware of context in thinking about disclosure when purpose of data use was presented together with context. Further, people are more willing to disclose their information to apps when *no information* is provided in the privacy questions for disclosure than when *vague information* about purpose is presented. Essentially, the privacy implications of such results are twofold. First, purpose is a critical element for people to evaluate the appropriateness of data collection when making privacy choices in real-time. Second, omission of purpose or other information about data collectors can be harmful to people's privacy because it causes people to be more permissive than otherwise when disclosing their personal information. Policies need to discourage value statements of purpose that are vague because the results in this thesis show that people find these vague statements even less satisfying than no purpose at all. Future research can focus at finding what might be a good purpose rather than a vague one to address this issue.

Overall, purpose plays the role in raising people's privacy awareness, which in turn affect their privacy decisions about disclosing information to apps. For example, results from the Contextual Privacy Study showed that merely presenting a vague purpose would be sufficient enough to alert subjects and decrease their willingness to disclose information. Given the importance of purpose, policymakers should pay closer attention to the mechanism for permissions request for data access on mobile platforms. The industry's

[1] For the definition of user context, please refer to Chapter 2.1.4



"Camera+" Would Like to Access the Camera.
To be able to take photos with Camera+ you'll need to allow this.

Allow "Weather" to Access Your Location Even When You Are Not Using the App?
Your location is used to show local weather in the "Weather" app and in Notification Center.

"Weather" Has Been Using Your Location in the Background. Do You Want to Continue Allowing This?
Your location is used to show local weather in the "Weather" app and in Notification Center.

Fitness is Using Your Location.

Figure 7-1: Permission request window in iOS apps that contain app usage context information

self-regulatory efforts to address privacy concerns by showing purpose of data collection should be re-examined in determining whether the expressed purpose is helpful or whether it is deceiving users in making privacy choices.

7.2.2 For Platform Providers and App Developers: Utilize Context to Gain Trust

The Intrusiveness Study in Chapter 3.2 shows that context is significant in helping people make privacy choices. For example, by showing how apps access mobile data within users' app usage context (idle vs. active), AppWindow can help users identify the extent to which apps behave intrusively. To better solicit people's privacy preferences, the findings in this thesis highlight the importance of showing privacy- relevant information in the privacy inquiries as well as presenting them *in situ*. Similar to principles outlined in the FTC report [29] "Building Trust Through Transparency", this thesis recommends platform providers to utilize context information for "just-in-time" disclosures to help consumers make informed decisions about the information they actually disclose.

Platform owners can actually play a key role in improving overall trust in the mobile app ecosystem by providing consumers enhanced transparency for when they are making privacy decisions. For example, new privacy features introduced by Apple's iOS 8 are coincidentally similar to concepts presented in the Intrusiveness Study. As shown in Figure 7-1, iOS 8 gives consumers control over which apps have access to information and shows the **context** in which data access is occurring, even in the background. The permission model in iOS applies users' app usage context (app in use or not in use) to raise users awareness of the apps tracking their locations, and to allow them to make decisions "just in time". The approach of iOS privacy framework shows the significance of context information in protecting user privacy from malign apps.

In addition, results in the Contextual Privacy Study (see Chapter 6.3.3) showed that some people disclosed their information solely based on their trust in the app itself. App developers are advised to request data in a more appropriate context in order to appear trustful to users. For example, in Apple's iOS 8, warning messages as shown in Figure 7-1(b) and (c) would encourage the app to switch to the "While in use" type of request if the app needs data only when interacting directly with users.

7.3 Future Work

There are several directions to pursue suggested by the research presented in this thesis. This section relates what has been done here about probing privacy in context to future extensions of the framework.

7.3.1 Advanced Data Processing and Visualization

The ContextProbe framework focuses mainly on providing high-level components for building context-triggered ESM apps and simplifying the process of bootstrapping an ESM study. Extending the capability to analyze and visualize tabular data in both the Personal Data Store and the Application Server in ContextProbe will help distill the potentially huge amounts of collected data into salient results for more insights. As demonstrated in Chapter 5.4.1, different kinds of charts and graphs can be used to visualize different types of sensor data in the Personal Data Store. Also, the application server currently only monitors the status of ongoing data collection by identifying gaps existing in the data. By adding more analytic power to ContextProbe, experimenters can construct more advanced queries to identify an event of interest, which can be represented by a particular pattern in the aggregated data. For example, one could have a query that detects the number of subjects who would not disclose their location information in the previous 3 hours with the location context being "at work". With this capability, experimenters can design more elaborate privacy studies to investigate subjects' behavior in real-time, or even change privacy questions on mobile phones through the cloud messaging service on the application server (as described in Chapter 5.5.4).

The modular design of the ContextProbe framework also makes it easy to integrate other existing frameworks for data analysis and visualization. Experimenters can use the app building platform in ContextProbe to create a standalone app for data collection and upload the data to other ESM platforms such as *ohmage* [42]. Alternatively, the Personal Data Store (described in Chapter 5.4) can be extended so that it can send data to other web servers besides the dedicated application server in ContextProbe.

7.3.2 Learning Personal Privacy Preferences

As mentioned in section 5.4.1, the Personal Data Store (PDS) provides an alternative to the current process of data collection on mobile phones, in which data is often collected in an opaque way and sent directly to experimenters' servers without the subjects' knowledge. The Personal Data Store improves the transparency of data collection by showing subjects what information about them has been collected. Furthermore, subjects who join ESM studies supported by ContextProbe can retain the ownership of their own data because the collected data is saved to the subjects' personal space before being sent to the application server. One future extension will be adding access control to the Personal Data Store for subjects to decide themselves what data they are comfortable of disclosing to the experimenters (data collectors). More generally, adding the capability of access control over collected data would make the ContextProbe framework suitable for use cases of a user-centric data consumption model as mentioned in section 1.2.

Using data collected in the Personal Data Store, the same predictive analytics and machine learning algorithms used to understand and manage preferences for context-aware services to improve user experience can be applied to learning privacy preferences. For example, by treating people's disclosure preferences as the dependent variable and other contextual factors as the independent variables, we can use techniques such as regression analysis to train a model for predicting people's future disclosure behavior. An alternative approach is to use a decision tree classifier to generate decision rules for the classification results in terms of the context variables. For example, a person might have a rule saying that he or she will never disclose location information to social apps with the location context being *home* and the activity context being *working alone*.

By learning from user's previous history of disclosures, we can implement a privacy assistant that automates privacy decisions when disclosing (or not disclosing) data to apps. Furthermore, we can design a *policy learner* that uses active learning techniques [77] and the Experience Sampling Method to iteratively build up a predictive model for data disclosure. While it is impractical to sample an individual's responses in all possible contexts, actively learning algorithms can be applied to reduce the times of probing for building the predictive user model. Active learning algorithms are employed to make decisions about the next unlabeled data points that should receive labels from subjects in order to improve the classification results. For example, when using a binary classifier to predict people's willingness to disclose their location data under specific contexts, the active learner will query labels (willingness to disclose) for those data points on which the classifier is least certain of their classification.

7.3.3 Possible Issues for Future Study

Results from the Contextual Privacy Study indicate that people's privacy preferences can be affected by how privacy questions for information disclosure are presented to them. Future studies in privacy preference should be aware of the differences between individuals and factor in people's prior privacy attitudes when analyzing their responses about information disclosure. Researchers can conduct interview or in-depth survey (see Chapter 6.2.2) to investigate what are the contextually grounded reasons behind subjects' behavior of information disclosure.

Further, some privacy questions might result in less disclosure solely due to human's psychological phenomena such as the "uncanny valley" effect [69] that triggers negative emotional responses. For example, the contrasting results between *no purpose* and *vague purpose* in the Contextual Privacy Study can be due to negative (creepy) feelings incurred by the purpose string "so it can have your information". It is important to factor in these psychological effects in the future studies for privacy preferences.

Lastly, it would be interesting to investigate the "cumulative effect" in the case of privacy preference studies that use the Experience Sampling Method. Using the ESM approach, study subjects might become more aware of their daily context when answering privacy questions. It is possible that subjects' answers given in the later stage of the study might be different than those provided in the early because of the cumulative effect of privacy awareness on information disclosure. For example, subjects might be willing to disclose locations of their visits to a physician when asked for the first few times. However, subjects might later find out the sensitivity of such activities and become less willing if asked in the same situation again.

7.4 Conclusion

As pointed out by Solove [81, p.65], new technologies alter "the extent to which privacy is a dimension of certain activities" and "what do we mean when we speak about certain activities involving privacy". Entering the brave new world of smart devices with ubiquitous connectivity, privacy researchers need to upgrade their approach, the theories their studies are based on, and the tools they use in order to fully understand people's privacy-relevant behavior that even they are not aware of. We all want control of our personal data as our right to privacy. However, identifying potential contexts where privacy conflicts might occur and raising peoples privacy awareness within those contexts are what this thesis recommend as the prerequisites for the design and implementation of any meaningful privacy control.

This thesis highlights the need for probing privacy preferences *in context* and *in situ*. It provides a framework and tools for experimenters to probe privacy in context and gain insights from both individual and systemic perspectives. This research has shown that the recognition of privacy-related context changes users' privacy expectations and helps identify intrusive behaviors of mobile apps. In the end, this thesis has shown evidence that people ignored context in thinking about disclosure when purpose of data use was presented together with context. When disclosing data to apps, people look for reasons to justify the appropriateness of information flow instead of fearing the sensitivity of the disclosed contents. As a result of my study, further research in fields such as behavioral economics that require real-time monitoring of user context, data collection, and in-situ responses might well be conducted using the ContextProbe framework.

Bibliography

[1] Privacy evaluation: what empirical research on users valuation of personal data tells us. *Internet Policy Review*, 2013.

[2] Fehmi Ben Abdesslem, Iain Parris, and Tristan Henderson. Mobile experience sampling: Reaching the parts of facebook other methods cannot reach. In *In Privacy and Usability Methods Powwow*, 2010.

[3] A Acquisti and J. Grossklags. Privacy and rationality in individual decision making. *Security Privacy, IEEE*, pages 26–33, 2005.

[4] A Acquisti and J. Grossklags. When 25 cents is too much: An experiment on willingness-to-sell and willingness-to-protect personal information. In *WEIS*, page 12, 2007.

[5] Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. Social fmri: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, 7(6):643–659, 2011.

[6] Shane Ahern, Dean Eckles, Nathaniel S Good, Simon King, Mor Naaman, and Rahul Nair. Over-exposed?: privacy patterns and considerations in online and mobile photo sharing. In *Proc. ACM CHI*, pages 357–366, 2007.

[7] Irwin Altman. *The Environment and Social Behavior: Privacy, Personal Space, Territory, and Crowding*. Brooks/Cole Publishing Company, 1975.

[8] Yaw Anokwa, Carl Hartung, Waylon Brunette, Gaetano Borriello, and Adam Lerer. Open source data collection in the developing world. *Computer*, 42(10):97–99, 2009.

[9] Sasikanth Avancha, Amit Baxi, and David Kotz. Privacy in mobile technology for personal healthcare. *ACM Computing Surveys (CSUR)*, 45(1):3, 2012.

[10] Rebecca Balebako, Jaeyeon Jung, Wei Lu, Lorrie Faith Cranor, and Carolyn Nguyen. “Little Brothers Watching You”: Raising Awareness of Data Leaks on Smartphones. In *Proc. of ACM SOUPS*, page 12, 2013.

[11] Gaurav Bansal, Fatemeh Mariam Zahedi, and David Gefen. The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online. *Decision Support Systems*, 49(2):138–150, 2010.

[12] Louise Barkhuus. The mismeasurement of privacy: using contextual integrity to reconsider privacy in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 367–376. ACM, 2012.

[13] Adam Barth, Anupam Datta, John C Mitchell, and Helen Nissenbaum. Privacy and contextual integrity: Framework and applications. In *Security and Privacy, 2006 IEEE Symposium on*, pages 15–pp. IEEE, 2006.

[14] M. Blum, A. Pentland, and G. Troster. Insense: Interest-based life logging. *MultiMedia, IEEE*, 13(4):40 –48, oct.-dec. 2006.

[15] danah boyd and Alice E Marwick. Social privacy in networked publics: teens attitudes, practices, and strategies. 2011.

[16] E.B. Boyd. *Personal.com creates an online valut to manage all your data*, 2014 (accessed November 19, 2014).

[17] Juan Pablo Carrascal, Christopher Riederer, Vijay Erramilli, Mauro Cherubini, and Rodrigo de Oliveira. Your browsing behavior for a big mac: Economics of personal information online. In *Proc. ACM WWW*, pages 189–200, 2013.

[18] JA Castañeda and FJ Montoro. The effect of Internet general privacy concern on customer behavior. *Electronic Commerce Research*, 2007.

[19] Centers for Medicare & Medicaid Services. The Health Insurance Portability and Accountability Act of 1996 (HIPAA). Online at https://www.cms.gov/Regulations-and-Guidance/HIPAA-Administrative-Simplification/HIPAAGenInfo/index.html, 1996.

[20] Ramnath K Chellappa and Raymond G Sin. Personalization versus privacy: An empirical examination of the online consumers dilemma. *Information Technology and Management*, 6(2-3):181–202, 2005.

[21] Mauro Cherubini and Nuria Oliver. A refined experience sampling method to capture mobile user experience. *Proc. of ACM CHI Workshop of Mobile User Experience Research*, pages 1–6, 2009.

[22] Erika Chin, Adrienne Porter Felt, Vyas Sekar, and David Wagner. Measuring user confidence in smartphone security and privacy. In *ACM SOUPS '12*, 2012.

[23] Tamlin S Conner, Matthias R Mehl, Mihaly Csikszentmihalyi, Harry T Reis, Norbert Schwarz, Ellen Hamaker, Peter Wilhelm, Meinrad Perrez, and Kurt Pawlik. *Handbook of research methods for studying daily life.*

[24] Sunny Consolvo and Miriam Walker. Using the experience sampling method to evaluate ubicomp applications. *IEEE Pervasive Computing*, 2(2):24–31, 2003.

[25] Lorrie Faith Cranor. Necessary but not sufficient: Standardized mechanisms for privacy notice and choice. *J. on Telecomm. & High Tech. L.*, 10:273, 2012.

[26] Yves-Alexandre de Montjoye, Erez Shmueli, Samuel S. Wang, and Alex Sandy Pentland. openpds: Protecting the privacy of metadata through safeanswers. *PLoS ONE*, 2014.

[27] William Enck, Peter Gilbert, Byung-Gon Chun, Landon P. Cox, Jaeyeon Jung, Patrick McDaniel, and Anmol N. Sheth. Taintdroid: an information-flow tracking system for realtime privacy monitoring on smartphones. OSDI'10. USENIX Association, 2010.

[28] William Enck, Peter Gilbert, Byung-Gon Chun, Landon P. Cox, Jaeyeon Jung, Patrick McDaniel, and Anmol N. Sheth. Taintdroid: An information-flow tracking system for realtime privacy monitoring on smartphones. In *Proc. OSDI'10*, pages 99–106, 2010.

[29] Federal Trade Commision. "Mobile Privacy Disclosures: Building Trust Through Transparency", 2013.

[30] Federal Trade Commission. COPPA: Children's Online Personal Privacy Act, 1999.

[31] Federal Trade Commission. Protecting Consumer Privacy in an Era of Rapid Change, March 2012.

[32] Adrienne Porter Felt, Kate Greenwood, and David Wagner. The effectiveness of application permissions. WebApps'11, Berkeley, CA, USA, 2011. USENIX Association.

[33] Adrienne Porter Felt, Elizabeth Ha, Serge Egelman, Ariel Haney, Erika Chin, and David Wagner. Android permissions: User attention, comprehension, and behavior. In *Proc. ACM SOUP*, page 3, 2012.

[34] Adrienne Porter Felt, Elizabeth Ha, Serge Egelman, Ariel Haney, Erika Chin, and David Wagner. Android permissions: User attention, comprehension, and behavior. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, SOUPS '12, pages 3:1–3:14, New York, NY, USA, 2012. ACM.

[35] Klint Finley. *Two Frenchmen Help You Quantify Yourself (Without Selling Your Soul)*, 2014 (accessed November 19, 2014).

[36] Jon Froehlich, Mike Y Chen, Sunny Consolvo, Beverly Harrison, and James A Landay. Myexperience: a system for in situ tracing and capturing of user feedback on mobile phones. In *Proceedings of the 5th international conference on Mobile systems, applications and services*, pages 57–70. ACM, 2007.

[37] R Gavison. Privacy and the Limits of Law. *Yale LJ*, 1979.

[38] Michelle R Grech, Andrew Neal, Gillian Yeo, Michael Humphreys, and Simon Smith. An examination of the relationship between workload and fatigue within and across consecutive days of work: is the relationship static or dynamic? *Journal of occupational health psychology*, 14(3):231, 2009.

[39] Larry Greenemeier. *3 Projects Prove Privacy Is Not Dead*, 2014 (accessed November 19, 2014).

[40] Il-Horn Hann, Kai-Lung Hui, Tom Lee, and I Png. Online information privacy: Measuring the cost-benefit trade-off. *Proc. of ICIS*, page 1, 2002.

[41] Kerm Henriksen, James B Battles, Margaret A Keyes, Mary L Grady, James R Fricton, Diane Davies, et al. Personal health records to improve health information exchange and patient safety. 2008.

[42] John Hicks, Nithya Ramanathan, Hossein Falaki, Brent Longstaff, Kannan Parameswaran, Mohamad Monibi, Donnie H Kim, Joshua Selsky, John Jenkins, Hongsuda Tangmunarunkit, et al. ohmage: An open mobile system for activity and experience sampling. Technical report.

[43] JoshuaB. Hurwitz. User choice, privacy sensitivity, and acceptance of personal information collection. In Serge Gutwirth, Ronald Leenes, Paul de Hert, and Yves Poullet, editors, *European Data Protection: Coming of Age*, pages 295–312. Springer Netherlands, 2013.

[44] Google Inc. Get started with Google Drive, 2014. [Online; accessed 6-October-2014].

[45] Google Inc. Google cloud messaging for android | android developers, 2014. [Online; accessed 6-October-2014].

[46] Keith Irwin and Ting Yu. An identifiability-based access control model for privacy protection in open systems. In *Proceedings of the 2004 ACM workshop on Privacy in the electronic society*. ACM, October 2004.

[47] Daniel Kahneman, Alan B Krueger, David A Schkade, Norbert Schwarz, and Arthur A Stone. A survey method for characterizing daily life experience: The day reconstruction method. *Science*, 306(5702):1776–1780, 2004.

[48] Patrick Gage Kelley, Lucian Cesca, Joanna Bresee, and Lorrie Faith Cranor. Standardizing privacy notices: an online study of the nutrition label approach. In *Proc. of ACM CHI*, pages 1573–1582, 2010.

[49] Patrick Gage Kelley, Lorrie Faith Cranor, and Norman Sadeh. Privacy as part of the app decision-making process. In *Proc. ACM CHI*, pages 3393–3402, 2013.

[50] Ashraf Khalil and Kay Connelly. Context-aware telephony: privacy preferences and sharing patterns. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, CSCW '06, pages 469–478. ACM, 2006.

[51] Bart P. Knijnenburg and Alfred Kobsa. Helping users with information disclosure decisions: Potential for adaptation. In *Proc. ACM IUI*, pages 407–416, 2013.

[52] E.J. Langer. Minding matters. *In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 22). New York, Academic Press.*, 1989.

[53] Reed Larson and Mihaly Csikszentmihalyi. The experience sampling method. *New Directions for Methodology of Social and Behavioral Science*, 15:41–56, 1983.

[54] Neal Lathia, Kiran K Rachuri, Cecilia Mascolo, and Peter J Rentfrow. Contextual dissonance: Design bias in sensor-based experience sampling methods. In *Proc. of ACM UBICOMP*, pages 183–192, 2013.

[55] Ilaria Liccardi, Joseph Pato, and Daniel J. Weitzner. Improving Mobile App selection through Transparency and Better Permission Analysis. *Journal of Privacy and Confidentiality: Vol. 5: Iss. 2, Article 1.*, pages 1–55, 2014.

[56] Jialiu Lin, Shahriyar Amini, Jason I. Hong, Norman Sadeh, Janne Lindqvist, and Joy Zhang. Expectation and purpose: Understanding users' mental models of mobile app privacy through crowdsourcing. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 501–510. ACM, 2012.

[57] Heather Richter Lipford, Gordon Hull, Celine Latulipe, Andrew Besmer, and Jason Watson. Visible flows: Contextual integrity and the design of privacy mechanisms on social network sites. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 985–989. IEEE, 2009.

[58] Clara Mancini, Keerthi Thomas, Yvonne Rogers, Blaine A Price, Lukazs Jedrzejczyk, Arosha K Bandara, Adam N Joinson, and Bashar Nuseibeh. From spaces to places: emerging contexts in mobile privacy. In *Proc. ACM UBICOMP*, pages 1–10, 2009.

[59] Erika McCallister, Tim Grance, and Karen Scarfone. Guide to Protecting the Confidentiality of Personally Identifiable Information (PII) -Recommendations of the National Institute of Standards and Technology. Technical report, NIST, April 2010.

[60] McKinsey Global Institute. Disruptive technologies: Advances that will transform life, business, and the global economy. May 2013.

[61] Min Mun, Shuai Hao, Nilesh Mishra, Katie Shilton, Jeff Burke, Deborah Estrin, Mark Hansen, and Ramesh Govindan. Personal data vaults: a locus of control for personal data streams. In *Proceedings of the 6th International Conference*, page 17. ACM, 2010.

[62] Min Y. Mun, Donnie H. Kim, Katie Shilton, Deborah Estrin, Mark Hansen, and Ramesh Govindan. Pdvloc: A personal data vault for controlled location data sharing. *ACM Trans. Sen. Netw.*, 10(4), June 2014.

[63] Ian Murphy. *Respect Networks launches private cloud*, 2014 (accessed November 19, 2014).

[64] Arvind Narayanan, Vincent Toubiana, Solon Barocas, Helen Nissenbaum, and Dan Boneh. A critical look at decentralized personal data architectures. *arXiv preprint arXiv:1202.4503*, 2012.

[65] Helen Nissenbaum. Privacy as Contextual Integrity. *Washington Law Review*, 79, 2004.

[66] Helen Nissenbaum. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford Law Books, 2009.

[67] Leysia Palen and Paul Dourish. Unpacking privacy for a networked world. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 129–136. ACM, 2003.

[68] Yong Jin Park and S. Mo Jang. Understanding privacy knowledge and skill in mobile communication. *Computers in Human Behavior*, 38:296 – 303, 2014.

[69] Frank E Pollick. In search of the uncanny valley. In *User centric media*, pages 69–78. Springer, 2010.

[70] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[71] Kiran K Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J Rentfrow, Chris Longworth, and Andrius Aucinas. Emotionsense: a mobile phones based adaptive platform for experimental social psychology research. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 281–290. ACM, 2010.

[72] N. Ramanathan, F. Alquaddoomi, H. Falaki, D. George, C. Hsieh, J. Jenkins, C. Ketcham, B. Longstaff, J. Ooms, J. Selsky, H. Tangmunarunkit, and D. Estrin. ohmage: An open mobile system for activity and experience sampling. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, 2012.

[73] Jeff John Roberts. *"Brightest Flashlight" Android app disclosed location of 50 million people, but FTC imposes no fine*, 2014 (accessed November 20, 2014).

[74] Franziska Roesner, Tamara Denning, Bryce Clayton Newell, Tadayoshi Kohno, and Ryan Calo. Augmented reality: Hard problems of law and policy. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, UbiComp '14 Adjunct, pages 1283–1288, New York, NY, USA, 2014. ACM.

[75] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[76] Paul Schwartz and Daniel Solove. The PII Problem: Privacy and a New Concept of Personally Identifiable Information. *New York University Law Review*, August 2011.

[77] Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52:55–66.

[78] Fuming Shih and Julia Boortz. Understanding people's preferences for disclosing contextual information to smartphone apps. In Louis Marinos and Ioannis Askoxylakis, editors, *Human Aspects of Information Security, Privacy, and Trust*, volume 8030 of *Lecture Notes in Computer Science*, pages 186–196. Springer Berlin Heidelberg, 2013.

[79] Katie Shilton and Kirsten E. Martin. Mobile privacy expectations in context. In *Proc. of Communication, Information and Internet Policy*, 2013.

[80] Irina Shklovski, Scott D. Mainwaring, Halla Hrund Skúladóttir, and Höskuldur Borgthorsson. Leakiness and creepiness in app space: Perceptions of privacy and mobile app use. In *Proc. ACM CHI*, pages 2347–2356, 2014.

[81] Daniel Solove. *Understanding Privacy*. Harvard University Press, 2010.

[82] Sarah Spiekermann, Jens Grossklags, and Bettina Berendt. E-privacy in 2nd generation e-commerce: privacy preferences versus actual behavior. In *Proceedings of the 3rd ACM conference on Electronic Commerce*, pages 38–47. ACM, 2001.

[83] Jacopo Staiano, Nuria Oliver, Bruno Lepri, Rodrigo de Oliveira, Michele Caraviello, and Nicu Sebe. Money walks: A human-centric study on the economics of personal mobile data. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2014.

[84] Joshua Tan, Khanh Nguyen, Michael Theodorides, Heidi Negrón-Arroyo, Christopher Thompson, Serge Egelman, and David Wagner. The effect of developer-specified explanations for permission requests on smartphone user behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 91–100, New York, NY, USA, 2014.

[85] Eran Toch. Crowdsourcing privacy preferences in context-aware applications. *Personal and Ubiquitous Computing*, pages 129–141, 2014.

[86] Eran Toch, Justin Cranshaw, Paul Hankes Drielsma, Janice Y Tsai, Patrick Gage Kelley, James Springfield, Lorrie Cranor, Jason Hong, and Norman Sadeh. Empirical models of privacy in location sharing. In *Proc. ACM UBICOMP*, pages 129–138, 2010.

[87] J Urban, C Hoofnagle, and Su Li. Mobile phones and privacy. In *UC Berkeley Public Law Research Paper*, 2012.

[88] Jayant Venkatanathan, Denzil Ferreira, Michael Benisch, Jialiu Lin, Evangelos Karapanos, Vassilis Kostakos, Norman Sadeh, and Eran Toch. Improving users consistency when recalling location sharing preferences. In *Proc, of INTERACT (Springer)*, pages 380–387. 2011.

[89] Daniel J Weitzner, Harold Abelson, Tim Berners-Lee, Joan Feigenbaum, James Hendler, and Gerald Jay Sussman. Information accountability. *Communications of the ACM*, 51(6):82–87, 2008.

[90] A. Westin, Louis Harris, and Associates. *Equifax-Harris Consumer Privacy Survey*. Equifax, Atlanta, Georgia, 1991.

[91] Alan F Westin. Privacy and freedom. *Washington and Lee Law Review*, 25(1):166, 1968.

[92] White House. Consumer data privacy in a networked world: A framework for protecting privacy and promoting innovation in the global digital economy. February 2012.

[93] White House. Big data and privacy: A technological perspective. pages 40–41, 2014.

[94] David Wolber. App inventor and real-world motivation. In *Proceedings of the 42Nd ACM Technical Symposium on Computer Science Education*, pages 601–606. ACM, 2011.

[95] World Economic Forum. Unlocking the value of personal data: From collection to usage. 2013.

[96] Fan Zhang. Assessing intrusiveness of smartphone apps. Master's thesis, Massachusetts Institute of Technology, 2012.

[97] Frances Zhang, Fuming Shih, and Daniel Weitzner. No surprises: measuring intrusiveness of smartphone applications by detecting objective context deviations. In *Proc. ACM Workshop on privacy in the electronic society*, pages 291–296, 2013.