Understanding People's Preferences for Disclosing Contextual Information to Smartphone Apps

Fuming Shih and Julia Boortz

Massachusetts Institute of Technology
{fuming,jboortz}@mit.edu

1 Introduction

Smartphones have become the primary and most intimate computing devices that people rely on for their daily tasks. Sensor-based and network technologies have turned smartphones into a "context-aware" information hub and a vehicle for information exchange. These information provide apps and third party with a wealth of sensitive information to mine and profile user behavior. However, the Orwellian implications created by context-awareness technology have caused uneasiness to people when using smartphone applications and reluctance of using them [6]. To mitigate people's privacy concerns, previous research suggests giving controls to people on how their information should be collected, accessed and shared. However, deciding *who* (people or the application) gets to access to *what* (types of information) could be an unattainable task. In order to develop appropriate applications and privacy policies it is important to understand under what circumstances people are willing to disclose information.

In this work, we explore people's willingness to disclose their personal data, especially contextual information collected on smartphones, to different apps for specific purposes. The goal is 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 [11,4] have the limitations in capturing the real causes for people's privacy concerns [2].

We used a hybrid approach of the experience sampling method [10] and the diary study to solicit people's willingness for disclosing information in different contexts. Specifically, we looked at possible contextual factors such as location, time and people's activities at the moment they are asked to disclose the data. Additionally, we 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?

L. Marinos and I. Askoxylakis (Eds.): HAS/HCII 2013, LNCS 8030, pp. 186–196, 2013.

We conducted a three week-long study with 38 participants to collect contextual information and self-reported data using smartphones. In parallel to that, we also solicited people's preference for information disclosure using contextualized questions. The questions specify the type of developers of the app, their purposes for data collection, benefits of sharing, and most importantly the user context. The responses to the questions enable us to build a preference model for each participant that reflects his or her privacy concerns in different contexts. We applied J48 implementation of C4.5 algorithm, a decision tree algorithm, to generate rules that could intuitively represent most relevant contextual factors. For some participants, the resulting models showed strong correlations between their decisions of information disclosure and their context, whereas others had decisions that were strongly biased toward other external factors such as the type of the data requestor or rewarded benefits.

2 Related Work

Research has shown that different types of context can affect smartphone users' decisions to disclose information. Context can include information about the situation users are in such as location, time of day, day of the week, and what users are doing [9,5,1]. It can also include information such as whom users are sharing it with, how the information will be used, what types of information are being shared, the level of detail of the shared information [11,12].

Khalil et al. [9] explored sharing patterns of context information by using the Experience Sampling Method (ESM). Their approach relied on self-reported data to capture user contexts by asking the user to input her location and activity manually every time. This approach, as with other studies that used ESM [1,5], is subject to getting false inputs from the users or missing labels after the users get annoyed because of the frequent prompts from the ESM program. To reduce the bias introduced by human errors, we improved ESM method by automatically detecting frequently visited places and prompting the user with the same label that the user has input earlier for the same place.

Mancini et al. [11] implemented the concept of "memory triggers", a short phrase to remind the participants of the situations when data about their experiences were collected. Using the memory phrase, the interviewer could then carry out a deferred contextual interview in which the participants were brought back in memory to recall a particular experience and the context of previous actions. We used the similar approach with some enhancements at creating the memory triggers. To record and reconstruct an individual's daily contexts, we used a hybrid approach similar to the Day Reconstruction Method (DRM) suggested in [8]. Our approach reconstructs the diary of the previous day automatically using user inputs of locations and activities through the enhanced ESM.

Jedrzejczyk et al. [7] investigated the effectiveness of using contextual information to model user preferences of real-time feedback in social location-tracking system. They built the predictive model by analyzing contextual information from sensor data on the smartphone. While using similar set of contextual information, we focus on exploring the effects of context on people's privacy decisions.

3 Approach

We want to accomplish the following two goals with the study approach: 1) to collect information that describes a participant's daily context, 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 nightly generated and customized for each participant according to their daily contexts collected in the previous day. By the end of the study, qualified participants were called to join in-lab interviews. The interview provides more insights and details about the "contextually ground" reasons of why participants shared or not shared their information under specific contexts.

3.1 Recruitment and Demographics

We recruited 38 participants from the campus through email-lists and flyers posted on bulletin boards. Twenty-eight participants were students (19 undergraduates and 9 graduate students) and twenty were female. The participants were screened for their English proficiency and use of the Android smartphone as their primary mobile device. About half of the participants lived outside of the campus; they possessed different lifestyle and composition of daily context (e.g. commuting between work places and homes) than that of the students. Participants were compensated based on their level of participation in the study, including 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 were 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.

3.2 Pre-experiment Survey

The participants were asked to fill a pre-experiment survey before the study to capture their familiarity of 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 1 summarizes the questions and the statistics of the answers in the survey. We 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

Table 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)

survey results showed that more than half of the participants are heavy users of smartphone apps and Internet web services.

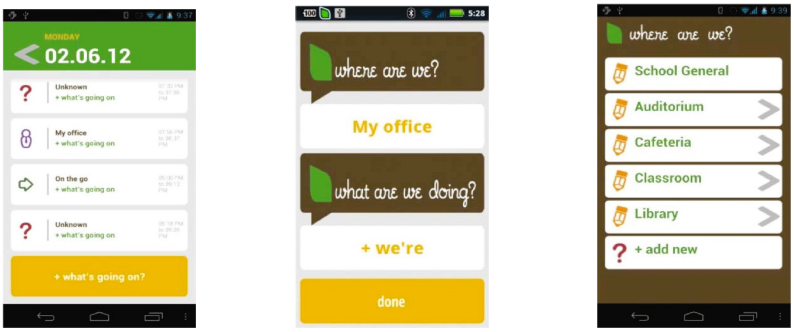
3.3 Data Collection: Recording a History of Daily Contexts

We used a hybrid approach of combining the experience sampling method with the diary method for acquiring in situ answers from the participants. We call the experience sampling method the *context recording* part and the diary method the *experience reconstruction* part 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 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, we were able to generate a history of contexts for different events that a participant encountered during the day.

3.4 Diary Study

For the "experience reconstruction" part, we sent a customized survey to each participant everyday with questions generated from the annotations of locations and activities each participant gave 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 2. Each survey contained 4 to 10 question groups, depending on how many



(a) Context history (b) Prompt asking for annotation (c) List of options for annotation

Fig. 1. Screenshots of the data logging application

annotations the participant provided for that day. Although a participant might provide several annotations of her locations and activities within an hour, we only sampled at most one annotation from that set. We chose one-hour window because people tend to regiment their life according to work-related schedule as described in previous research of life-logging applications [3].

For each question group, we presented three questions to collect the preferences for disclosing different contents: location data, situation data, and proximity data (bluetooth scanning of the nearby devices). We ask the participants "Would you have disclosed..." to clearly indicate that we want them 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. The participant was asked questions about her 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 developer type is 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, we used multiple different organizations for each category of requestors. For the category academic entities, we used *MIT*, *Media Lab* and *Harvard Medical*. For the category local companies, we used *banking*, *retail store*, and *grocery store*. The specific grocery store, retailer, or banking company is customized to participants based on the pre-experiment survey indicating which companies they normally use. We anticipate that this customization will make users responses more representative of their actual disclosure preferences, since it brings the experiment closer in-line with their everyday life. Finally, we used *Google*, *Amazon*, *Facebook* to represent well-know large companies with web services.

For each category of requestors, we included the benefit or the purpose for collection the information. For academic requestors, the survey questions tell

users that the data is being collected for research purposes. When the requestor is a company, users are asked if they would disclose the information in return for a $2 coupon. Finally, when well-known large companies with web services are asking for information, users are told that the purpose is for improving personal service. We expect that these purposes will help eliminate hesitancy to share by showing users that the information disclosed will be useful for the requestor.

4 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. Those participants started but quitted the study early, their results were not taken into consideration. The overall participant rate of the study, counting those who finished both the data collection and diary survey, was 71%.

In this section, we report the main findings of our study, including both quantative data collected from the study and qualitative interview data. We start by

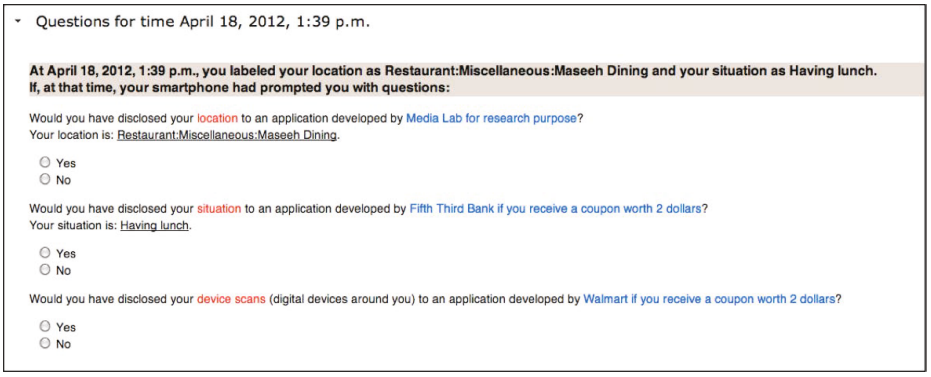


Fig. 2. Example of a personalized questionnaire based on "contextual information"

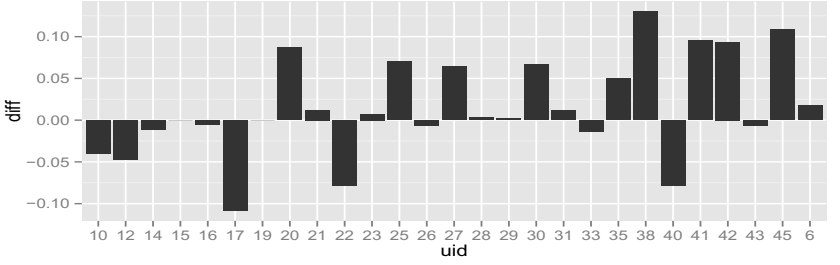


Fig. 3. The percentage of *yes* responses for disclosing locations annotated as *home* vs. the percentage of *yes* responses specifically at time slots after 6pm or before 6am

describing the general outcome from the survey questions. Then we look into the responses of each individual and how the results relate to contextual factors using outputs from the decision tree algorithm. Lastly, we use interview data 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 are most likely to disclose activity data (62% yes), followed closely by location data (59% yes), but are much less likely to disclose Bluetooth data (49% yes). The interview data revealed that the participants are more reluctant to disclose Bluetooth data due to the unsureness of what information can be disclosed by Bluetooth data.

As for the general trend across individuals, we found that the preference for disclosing information are dependent on the participant's location at the time of sharing. For instance, the participants are 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 are 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 contrast, our results also showed the difference in time did not significantly affect the participant's willingness to disclose location. For example, Figure 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.

When considering the data requestor, the participants are 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 show that users are more willing to disclose information to people who they are closer to – in this case local businesses as opposed to larger web services.

Individual Preference Model. We ran C4.5 decision tree algorithm and produced rules from each participant's responses. Our results showed that about

Table 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)				

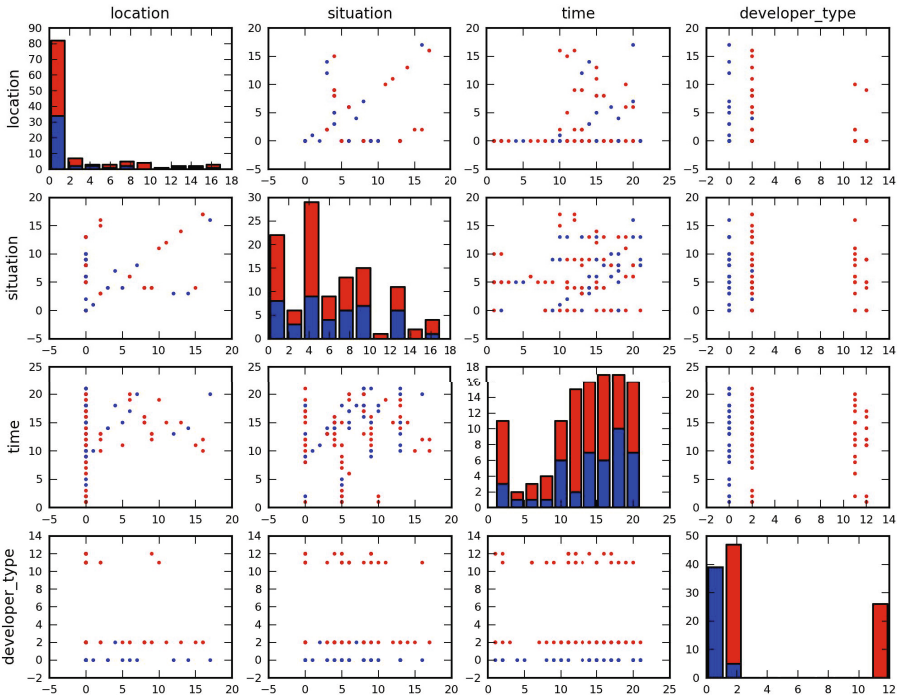


Fig. 4. Responses (··· Deny; ··· Allow) showing privacy preferences are biased towards certain companies (*developer type*)

81% (22/27) of the study participants have obvious patterns in their responses. Table 2 summarizes the results of participants' reponses, and it shows that some participants are what Westin called "privacy fundamentalist" and "privacy unconcerned" [13] such that they either rejected or accepted most of the data requests in the survey questions. About 54% (12/22) of these participants 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. For instance, Figure 4 shows that 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.[1]), 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

[1] Each box 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.

disclose her 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 her shopping behaviors. P29 would not disclose all locations labeled as *home* to requestors besides those from the category *academic entities*.

4.1 Post Interview

We invited the participants who completed the study for 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 sharing their information. We identified three characteristics of how some participants evaluate privacy risks based on their privacy expectations that are shaped by context: 1) private or public of their context, 2) sensitivity of the disclosed information, and 3) relations between the purpose of data collection and the context.

One of the deciding factors for disclosing personal information is to consider whether its context is private or public [11]. However, 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. Second, 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, *"I think device scans give information about the devices around me, and it is not personal."* Deciding the sensitivity of information then 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 *"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 the questions by just their instincts. But as the study continued, rules were introduced accumulatively through

relevant contexts. For example, P22 *"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."*

5 Conclusion

Our study of people's preference for information disclosure on smartphones has addressed three challenges in mobile privacy research. Firstly, to record information that approximates an individual's daily contexts, we used an enhanced experience sampling method. The ESM program prompts the user automatically for annotations of locations and activities whenever it detects a new place or that a previous labeled place is re-visited.

Secondly, in order to investigate people's *privacy in context*, we created the personalized survey in which each participant would answer questions with the help of the contextual triggers. The participant would to give her privacy preferences while recalling the experience *in situ*. We then applied the decision tree algorithm C4.5 to generate a preference model for each participant. We found that although people have some default policies, not much can be gleaned about just how much contextual factors can affect people's decisions about data disclosure. Furthermore, both the quantative and the qualitative data showed that other external factors such as types of the data requestors predominate over the contextual factors.

Lastly, the participants had several issues when providing their responses in the study. These issues include the lack of understanding about the privacy impacts of disclosed data and lack of connection between their decision and the purpose of the data collection. Together, these problems lead to indifferent responses during different contexts of data disclosure. Future study should inform people the capability of the technology that is used in data collection and create a sense of real use of the disclosed information for a specfic purpose instead of presenting just hypothetical questions.

References

1. Anthony, D., Henderson, T., Kotz, D.: Privacy in location-aware computing environments. IEEE Pervasive Computing 6(4), 64–72 (2007)
2. Barkhuus, L.: The mismeasurement of privacy: using contextual integrity to reconsider privacy in hci. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2012, pp. 367–376. ACM (2012)
3. Blum, M., Pentland, A., Troster, G.: Insense: Interest-based life logging. IEEE MultiMedia 13(4), 40–48 (2006)
4. Castañeda, J., Montoro, F.: The effect of Internet general privacy concern on customer behavior. Electronic Commerce Research (2007)
5. Consolvo, S., Smith, I.E., Matthews, T., LaMarca, A., Tabert, J., Powledge, P.: Location disclosure to social relations: why, when, & what people want to share. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2005, pp. 81–90. ACM (2005)

6. Madden, M., Boyles, J.L., Smith, A.: Privacy and data management on mobile devices. Technical Report CS-2011-02, Pew Research Center (September 2012)
7. Jedrzejczyk, L., Mancini, C., Corapi, D., Price, B., Bandara, A., Nuseibeh, B.: Learning from context: A field study of privacy awareness system for mobile devices (2011)
8. Kahneman, D., Krueger, A., Schkade, D., Schwarz, N., Stone, A.: A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method (2004)
9. Khalil, A., Connelly, K.: Context-aware telephony: privacy preferences and sharing patterns. In: Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work, CSCW 2006, pp. 469–478. ACM (2006)
10. Larson, R., Csikszentmihalyi, M.: The experience sampling method. New Directions for Methodology of Social and Behavioral Science 15, 41–56 (1983)
11. Mancini, C., Thomas, K., Rogers, Y., Price, B., Jedrzejczyk, L., Bandara, A., Joinson, A., Nuseibeh, B.: From spaces to places: emerging contexts in mobile privacy. In: Proceedings of the 11th International Conference on Ubiquitous Computing, pp. 1–10 (2009)
12. Patil, S., Lai, J.: Who gets to know what when: configuring privacy permissions in an awareness application. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2005, pp. 101–110. ACM (2005)
13. Westin, A., Harris, L. Associates: Equifax-Harris Consumer Privacy Survey. Equifax (1996)