Us and Them — A Study of Privacy Requirements Across North America, Asia, and Europe

Swapneel Sheth, Gail Kaiser
Department of Computer Science
Columbia University
New York, NY, USA
{swapneel, kaiser}@cs.columbia.edu

Walid Maalej
Department of Informatics
University of Hamburg
Hamburg, Germany
maalej@informatik.uni-hamburg.de

ABSTRACT

Data privacy when using online systems like Facebook and Amazon has become an increasingly popular topic in the last few years. However, only a little is known about how users and developers perceive privacy and which concrete measures would mitigate privacy concerns. To investigate privacy requirements, we conducted an online survey with closed and open questions and collected 408 valid responses. Our results show that users often reduce privacy to security, with data sharing and data breaches being their biggest concerns. Users are more concerned about the content of their documents and personal data such as location than their interaction data. Unlike users, developers clearly prefer technical measures like data anonymization and think that privacy laws and policies are less effective. We also observed interesting differences between people from different geographies. For example, people from Europe are more concerned about data breaches than people from North America. People from Asia/Pacific and Europe believe that content and metadata are more critical for privacy than people from North America. Our results contribute to developing a user-driven privacy framework that is based on empirical evidence in addition to the legal, technical, and commercial perspectives.

Categories and Subject Descriptors

D.2.1 [**Software Engineering**]: Requirements/Specifications; K.4.1 [**Computers and Society**]: Public Policy Issues—*Privacy*

General Terms

Human Factors

Keywords

Human factors in software engineering, requirements engineering, privacy, user developer collaboration, empirical studies

1. INTRODUCTION

As systems that collect and use personal data, such as Facebook and Amazon, become more pervasive in our daily lives, users are starting to worry about their privacy. There has been a lot of media and press coverage about data privacy. One of the earliest articles in the New York Times reported how it was possible to break the anonymity of AOL's search engine's users [7]. A more recent article mentions privacy concerns about Google Glass [25]. Both technical and, especially, non-technical users are finding it increasingly hard to navigate this privacy minefield [20]. This is further exacerbated by well-known systems periodically making changes that breach privacy and not allowing users to opt out a-priori [18].

There is a large body of research on privacy in various research communities. This ranges from novel techniques to make privacy settings more understandable [17, 27] to data anonymization techniques for privacy in different domains [12, 22, 28, 36]. Recent studies have shown that there is a discrepancy between users' intentions and reality for privacy settings [23, 24]. The assumption behind most of this work is that privacy is well-specified and important. However, there is very little evidence about what exactly are the user concerns, priorities, and trade-offs, and how users think these concerns can be mitigated. In particular, in the software engineering community, there have been no systematic studies to find out what the privacy requirements are for users and how these requirements should be addressed by developers.

This paper aims to understand the privacy expectations and needs for modern software systems. To this end, we conducted a study using an online survey. We received 595 responses and selected 408 of them as valid. The responses included diverse populations including developers and users, and people from North America, Europe, and Asia. The results of our study show that the biggest privacy concerns are data sharing and data breaches. On the other hand, there is disagreement on the best approach to address these concerns. With respect to types of data that are critical for privacy, people are least concerned about metadata and most concerned about their personal data and the content of documents. Most people are not willing to accept less privacy in exchange for fewer advertisements and financial incentives such as discounts on purchases.

The main contribution of this paper is threefold. First, it illustrates and quantifies the general trends on how people understand privacy and on how they assess different privacy concerns and techniques to address them. Second, the paper

identifies differences in privacy expectations between various groups: developers and users, on one hand, and people from different geographic regions, on the other hand. Finally, the paper gives insights into how software developers and managers can address privacy concerns of their users.

Our analysis for geographic regions, for example, shows that there is a significant difference between respondents from North America, Europe, and Asia/Pacific. People from Europe and Asia/Pacific rate different types of data such as metadata, content, and interaction being a lot *more critical* for privacy than respondents from North America. People from Europe are a lot more concerned about data breaches than data sharing whereas people from North America are equally concerned about the two. Similarly, our analysis for developers versus users shows a marked difference between the two groups. For example, developers believe that privacy laws and policies are *less* effective for reducing privacy concerns than data anonymization.

The rest of the paper is organized as follows. Section 2 describes the design of our study. Sections 3–5 highlight its key results. Section 6 discusses the implications of the results and limitations of the study. We describe related work in Section 7 and conclude the paper in Section 8.

2. STUDY DESIGN

We describe the research questions, methods, and respondents of our study.

2.1 Research Questions

The goal of this study is to gather and analyze privacy requirements for modern software systems. In particular, we want to study the perception of different groups of people on privacy. According to the Merriam-Webster dictionary, privacy is the "freedom from unauthorized intrusion." We are interested specifically in data privacy and other notions of privacy such as physical privacy are beyond the scope of our work.

We focused on the following research questions:

- **RQ 1**: What are developers' and users' perceptions of privacy? What aspects of privacy are more important and what are the best techniques to address them? (Section 3)
- **RQ 2**: Does software development experience have any impact on privacy requirements? (Section 4)
- **RQ 3**: Does geography have any impact on privacy requirements? (Section 5)

By perception, we mean the subjective understanding and assessment of privacy aspects. Since privacy is a very broad term, we are interested in specific aspects such as types of concerns, techniques to mitigate these concerns, types of data that are critical to privacy, and whether people would give up privacy. We focus on these aspects because we think they are most related to software engineering topics.

2.2 Research Method

We created an online survey that consisted of 16 questions. Out of these, 14 questions were closed and respondents had to choose an answer from a given list of options. These questions consisted of 3-point and 5-point semantic scale questions. These types of questions allow for the measurement of subjective assessments of people while allowing for some flexibility of the interpretation [30]. For example, one of the questions was: "Would users be willing to use your system if they are worried about privacy issues?" and the answer options were: "Definitely yes - Users don't care about privacy," "Probably yes," "Unsure," "Probably not," and "Definitely not - if there are privacy concerns, users will not use this system." We used the 3-point scale for the majority of the questions as we did not need respondents to have higher discriminative powers as needed by the 5-point scale. Jacoby and Matell [21] have shown that 3-point scales do not result in any significant reduction in reliability or validity.

We chose a survey instead of direct observation or interviews because of the following reasons: First, a survey is very scalable and allowed us to get a large number and broad cross-section of responses. Second, we were interested in the subjective opinion of people and this can be different from real behavior. Third, the closed questions were purely quantitative and allowed us to analyze general trends for our survey respondents. In addition, the survey also had two open-ended questions. This helped us get qualitative insights about privacy and gave an opportunity for respondents to report aspects that were not already included in the closed questions.

Respondents could choose to fill out our survey in two languages: English or German. For each language, there were two slightly different versions based on whether the respondents had experience in software development or not. The difference in the versions was only in the phrasing of the questions in order to reduce confusion. For example, developers were asked: "Would users be willing to use your system if they are worried about privacy issues?" whereas users were asked: "Would you be willing to use the system if you are worried about privacy issues?" In total, the survey took 5-10 minutes to answer.

To increase the *reliability* of the study [30], we took the following measures:

- Pilot Testing: We conducted pilot testing of our survey in four iterations with a total of ten users that focused on improving the timing and understandability of the questions. We wanted to reduce ambiguity about the questions and answers and ensure that none of the semantics were lost in translation. We used the feedback from pilot testing to improve the phrasing and the order of questions for both the English and German versions.
- Random order of answers: The answer options for the closed questions were *randomly* ordered. This ensures that answer order does not influence the response [38].
- Validation questions: To ensure that respondents did not fill out the answers arbitrarily, we included two validation questions [3]. For example, one of the validation questions was: "What is the sum of 2 and 5?" Respondents who did not answer these correctly were not included in the final set of valid responses.
- Post sampling: We monitored the number of respondents from each category of interest: developers versus users and geographic location. We conducted post-sampling and stratification to ensure that we got sufficient responses for each category and that the ratio

	Developers	Users
North America	85	44
Europe	116	65
Asia	61	30
South America	3	2
Africa	2	0

Table 1: Summary of study respondents based on location and software development experience

of developers to users for each geographic location was roughly similar. For categories that did not have sufficient respondents, we targeted those populations by posting the survey in specific channels. We stopped data collection when we had a broad spectrum of respondents and sufficient representation in all the categories of interest.

Finally, to corroborate our results and analysis of data, we conducted a number of statistical tests. In particular, we used the Z-test for equality of proportions [32] and Welch's Two Sample t-test to check if our results are statistically significant.

2.3 Survey Respondents

We did not have any restrictions on who could fill out the survey. We wanted, in particular, people with and without software development experience and people from different parts of the world. We distributed our survey through a variety of channels including various mailing lists, social networks like Facebook and Twitter, and personal and professional colleagues. We circulated the survey across companies with which we are collaborating. We also asked specific people with many contacts (e.g., with many followers on Twitter) to forward the survey. As an incentive, two iPads would be raffled among the survey respondents.

In total, 595 respondents filled out our survey between 10 November 2012 and 8 September 2013. Filtering out the incomplete and invalid responses resulted in 408 valid responses (68.6% completion rate). Table 1 shows the respondents based on location and software development experience. The four versions of the survey along with raw data and summary information are available on our website[1]. Among the respondents, 267 have software development experience and 141 do not. For respondents with software development experience, 28 have less than one year of experience, 129 have 1-5 years of experience, 57 have 5-10 years of experience, and 53 have more than ten years of experience. 129 respondents live in North America, 181 in Europe, and 91 in Asia/Pacific. 166 are affiliated with industry and the public sector, 182 are in academia and research, and 56 are students.

3. PRIVACY PERCEPTIONS

We asked respondents: "How important is the privacy issue in online systems?" They answered using a 5-point semantic scale ranging from "Very important" to "Least important." 66.7% of the respondents chose "Very Important," 25.3% chose "Important," and the remaining three options ("Average," "Less Important," "Least Important") combined were chosen by a total of 8.1% of the respondents.

The *location* of the data storage was a key concern for the respondents. We asked respondents whether privacy concerns depend on the location of where the data is stored and provided a 5-point semantic scale with options: "Yes," "Maybe yes," "Unsure," "Maybe not," "No." 57.7% of the respondents chose "Yes" and 28.6% chose "Maybe yes" while the remaining three options were chosen by a total of 13.7% of the respondents.

On the other hand, there was disagreement about whether users would be *willing* to use such systems if there were privacy concerns.. The answer options were: "Definitely yes - Users don't care about privacy," "Probably yes," "Unsure," "Probably not," and "Definitely not - if there are privacy concerns, users will not use this system." 20.8% of the respondents choose "Unsure," while 34.8% and 29.4% chose "Probably yes" and "Probably not" respectively.

3.1 What factors would increase and reduce privacy concerns?

We asked respondents if the following factors would *increase* privacy concerns:

- Data Aggregation: The system discovers additional information about the user by aggregating data over a long period of time.
- Data Distortion: The system might misrepresent the data or user intent.
- Data Sharing: The collected data might be given to third parties for purposes like advertising.
- Data Breaches: Malicious users might get access to sensitive data about other users.

For each concern, the respondents could answer using a 3-point semantic scale having options: "Yes," "Uncertain," and "No." We also asked respondents if the following would help to *reduce* concerns about privacy:

- Privacy Policy, License Agreements, etc.: Describing what the system will/won't do with the data.
- Privacy Laws: Describing which national law the system is complaint with (e.g., HIPAA in the US, European privacy laws).
- Anonymizing all data: Ensuring that none of the data has any personal identifiers.
- Technical Details: Describing the algorithms/source code of the system in order to achieve higher trust (e.g., encryption of data).
- Details on usage: Describe, e.g., in a table how different data are used.

The overall answers of the respondents for both questions are shown in Figure 1. In the figure, each answer option is sorted by the number of "Yes" respondents.

Most respondents agreed that the biggest privacy concerns are data breaches and data sharing. There is disagreement about whether data distortion and data aggregation would cause privacy concerns. To check if these results are statistically significant, we ran Z-tests for equality of proportions. This would help us validate, for example, if there is a statistically significant difference in the number of respondents who said "Yes" for two different options. The results for *increasing* concerns about privacy are shown in Table 2. For

[1] `http://www.psl.cs.columbia.edu/1476/privacy-requirements/`

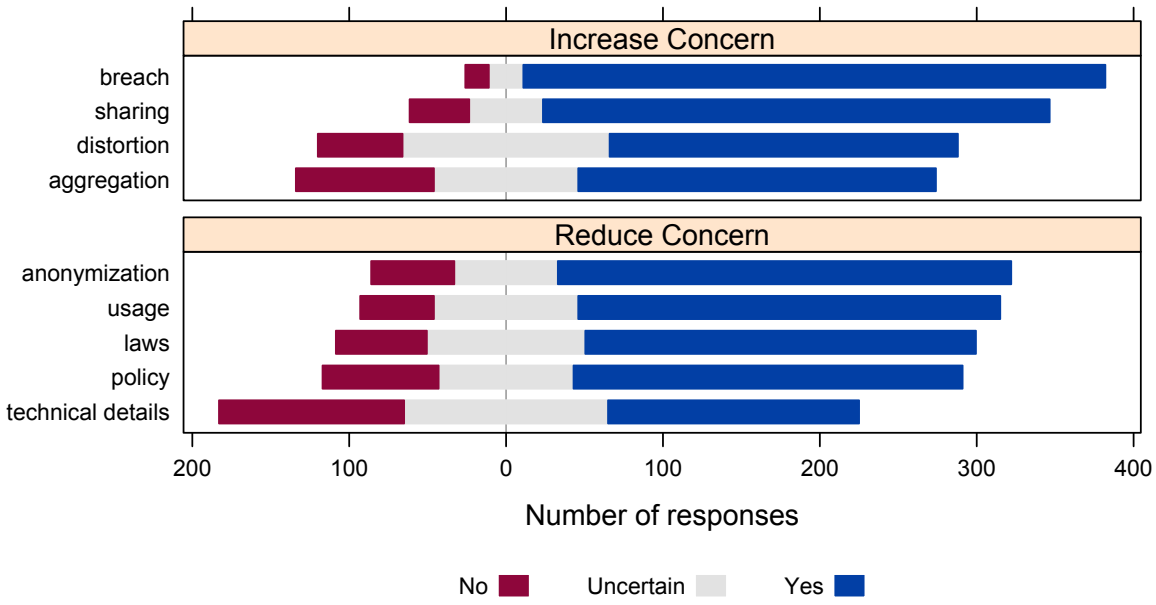

Figure 1: What increases and reduces privacy concerns?

	p-values
Sharing > Aggregation	$p = 1.231e^{-12}$
Sharing > Distortion	$p = 6.036e^{-14}$
Breach > Aggregation	$p < 2.2e^{-16}$
Breach > Distortion	$p < 2.2e^{-16}$

Table 2: What *increases* privacy concerns? For each privacy concern, $X > Y$ indicates that X is a bigger concern than Y for the respondents. (We used the Z-test for equality of proportions and only statistically significant results for $p < 0.01$ are shown.)

all of these tests, the null hypothesis is that a similar fraction of the respondents chose "Yes" for both options. Our results show that the concerns about data breaches and data sharing are statistically significantly higher than data aggregation and data distortion ($p \leq 1.231e^{-12}$).

Hypothesis 1: People are more concerned about the security aspects of privacy and in particular, data sharing and data breaches than data distortion and data aggregation.

As far as *reducing* concerns, respondents consider technical details the least effective option. It is statistically significantly the worst option (with p-values ranging from $7.218e^{-10}$ for comparing to policy to $2.2e^{-16}$ for comparing to anonymization). Respondents think that anonymization is the most effective option for mitigating privacy concerns and statistically significantly better than privacy laws ($p = 0.003$) and privacy policy ($p = 0.002$). There is, however, no statistically significant difference between anonymization and providing usage details ($p > 0.15$). The remaining three options (privacy policy, privacy laws, and usage details) had similar responses and none of their combinations for the Z-test yielded statistically significant results for $p < 0.01$.

Hypothesis 2: Different people assess the importance of various privacy mitigation measures differently.

3.2 Qualitative feedback

Overall, we collected 135 comments from 408 respondents on our open questions. We analyzed these comments manually in three steps. First, we read each comment and annotated it with a few summarizing keywords. Thereby, we tried to reuse the keywords whenever possible. Second, we unified and grouped the keywords into topics, making sure that no important comments are lost. Finally, we read the comments again and assigned each of them to one of the identified concerns:

3.2.1 Additional Privacy Concerns

With respect to additional privacy concerns, we collected 66 comments. 15 comments were useless as they just repeated the standard response options, were not understandable, or without content (e.g., "no comment," or "nothing more"). The remaining 51 comments gave interesting insights, which can be grouped into the following topics:

Authorities and intelligent services: 13 respondents mentioned authorities and intelligent services as an additional privacy concern. One wrote: "Government access is not exactly a data breach, but still a privacy concern." Another commented: "anyway there is prism." It is important to mention that about half of the responses were collected after the NSA PRISM scandal [13, 15].

APIs, program correctness, and viruses: Nine respondents mentioned concerns related to the program behavior, including malicious programs and viruses. Respondents also mentioned that privacy concerns are "transmitted" through the application programming interfaces of the tools collect-

ing data. One respondent wrote: "Sharing data over API" while others mentioned specific systems such as Google Analytics or Facebook API. Three respondents specifically pointed the correctness of privacy implementation as a specific privacy concern.

Unusable and nontransparent policies: Seven users complained about unusable privacy implementations with unclear, nontransparent policies. These respondents were concerned because most users simply do not know which data is being collected about them and for what purposes. One respondent wrote: "Companies and software developers shield themselves [...] by making consumers agree on a long, convoluted, and often a hard to understand hard to read [...] policy. Companies know that people do not read them a tactic on which they are banking." Another gave a more concrete example: "Sometimes sharing sensitive data is activated by default in applications (unaware users would leave it so)." One respondent wrote: "Transparency and letting the user choose make a huge difference. Maybe not in the beginning and maybe not for all users but definitely for a specific user group."

Intentional or unintentional misuse: At least seven respondents mentioned different forms of misusing the data as main concerns. This includes commercial misuse such as making products of interest more expensive, but it could also be misused for social and political purposes. Apart from abusing the data to put pressure on users, respondents mentioned using fake data to manipulate public opinions or inferencing sensitive information about groups of people and minorities. One respondent wrote: "Whenever something happen the media uses their data accessible online to 'sell' this person as good or evil."

Lack of control: Seven respondents mentioned the lack of control and in particular, options to delete data collected about them as their main concern. One wrote: "if we agree to give the data, we are not able anymore to revise this decision and delete the data. Even if the service confirms the deletion, we don't have any mean of control." Another respondent explicitly mentioned the case where companies owning their data are bankrupt or sold and in this case, the privacy of their data is also lost. "Company A has a decent privacy policy, Company B acquires the company, and in doing so, now has access to Company A's data."

Combined data sources: Five respondents explicitly mentioned combining data about users from different sources as a main privacy concern. In most cases, this cannot be anticipated when developing or using a single system or a services. One respondent wrote: "it's difficult to anticipate or assess the privacy risk in this case." Another claimed: "continuous monitoring, combined with aggregation over multiple channels or sources, leads to complex user profiling and it's disturbing to know that your life is monitored on so many levels."

Collecting and storing data: Five respondents wrote that collecting and storing data is, on its own, a privacy concern. In particular, respondents complained about too much data being collected about them and stored for too long time. One respondent mentioned: "the sheer amount of cookies that are placed on my computer just by landing on their website." Another claimed: "collecting the data and storing for a long period of time is seen more critical than just collecting."

Other issues: Three respondents mentioned problems with the legal framework and in particular, the compatibility of laws in the developer and user countries. Three respondents said that in some cases there is no real option to not use a system or service, e.g., due to a "social pressure as all use Facebook" or since "we depend on technology."

3.2.2 Suggestions for Reducing Privacy Concerns

In total, 69 respondents answered the open question on additional measures to reduce user concerns about privacy. Ten of these answers either repeated the standard options or were useless. The remaining 59 comments showed more convergence in the opinion than the comments on the additional concerns, possibly because this question was more concrete. The suggestions can be grouped into the following measures:

Easy and fine-grained control over the data, including access and deletion: 17 respondents recommended allowing the users to easily access and control the collected and processed data about them. In particular, respondents mentioned the options of deactivating the collection and deleting the data. One respondent wrote: "to alleviate privacy concerns, it should be possible to opt out of, or 'not agree' to certain terms." Another wrote: "allow users to access to a summary of all the data stored on their behalf, and allow them to delete all or part of it if they desire." The respondents also highlighted that this should be simple and easy to do and embedded on the user interface at the data level.

Certification from independent trusted organizations: 14 respondents suggested introducing a privacy certification mechanism by independent trusted authorities. A few also suggested the continuous conduction of privacy audits similar to other fields such as safety and banking. Respondents also suggested that the results of the checks and audits should be made public to increase the pressure on software vendors. One respondent even suggested "having a privacy police to check on how data is handled."

Transparency and risk communication, open source: 13 respondents mentioned increased transparency about the collection, aggregation, and sharing of the data. In particular, respondents mentioned that the risks of misusing the data should be also communicated clearly and continuously. Three respondents suggested that making the code open source would be the best approach for transparency. One wrote: "tell users (maybe in the side-bar) how they are being tracked. This would educate the general public and ideally enable them to take control of their own data." The spectrum of transparency was from the data being collected to physical safety measures of servers and qualifications of people handling data to how long the data is stored.

Period and amount of data: 11 respondents recommended always limiting and minimizing the amount of data and the period of storage, referring to the principle of minimality. The period of time for storing the data seems to be crucial for users. One wrote: "Not allowing users data being stored

in servers. Just maintaining them in the source."

Security and Encryption: We noted that respondents strongly relate privacy issues to information security. At least seven suggested security measures, mainly complete encryption of data and communication channels.

Trust and Education: Seven respondents mentioned building trust in the system and vendor as well as education of users on privacy as effective means to reduce privacy concerns.

Short, usable, precise and understandable description, in the UI: At least six respondents mentioned increasing the usability to access data and policy as an important measure to reduce privacy concerns. One wrote: "the disclaimer should be directly accessible from the user interface when conducting a function which needs my data." Another respondent wrote: "short understandable description and no long complicated legal text."

3.3 What types of data are critical?

To get a deeper insight into the privacy criticality of different types of data, we asked respondents to rate the following types of data on a 5-point semantic scale ranging from "Very critical" to "Uncritical."

- Content of documents (such as email body)
- Metadata (such as date)
- Interaction (such as a mouse click to open or send an email)
- User location (such as the city from where the email was sent)
- Name or personal data (such as email address)
- User preferences (such as inbox or email settings)

The results are shown in Figure 2. Respondents chose content as most critical, followed by personal data, location, preferences, and interaction and metadata are the least critical as far as privacy is concerned.

We used Welch's Two Sample t-test to compare if the difference among the different types of data is statistically significant. The null hypothesis was that the difference in means was equal to zero. The results are summarized in Table 3. The results show, for example, that there is no statistically significant difference between content and personal data. On the other hand, there is a statistically significant difference between content and location for $p < 0.01$.

Hypothesis 3: People are more concerned about content and personal data than interaction and metadata.

3.4 Would one give up privacy?

We asked respondents if they would accept *less* privacy for the following:

- Monetary discounts (e.g., 10% discount on the next purchase)
- "Intelligent" or added functionality of the system (such as the Amazon recommendations)
- Fewer advertisements

For each option, the respondents could answer using a 3-point Semantic scale having options: "Yes," "Uncertain," and "No." The results are shown in Figure 3.

36.7% of the respondents said they would accept less privacy for added functionality of the system while only 20.7% and 13.7% would accept less privacy for monetary discounts and fewer advertisements respectively. Added functionality seems to be the most important reason to accept less privacy. These results are statistically significant using the Z-test for equality of proportions ($p < 3.882e^{-5}$ for monetary discounts and $p < 1.854e^{-9}$ for fewer advertisements).

Even though these results are statistically significant, it is important to note that *less than half* of the respondents would accept less privacy for added functionality of the system.

Previous studies, such as the one conducted by Acquisti et al. [1], have showed, however, that people's economic valuations of privacy vary significantly and that people *do* accept less privacy for monetary discounts. This contrast in results might be due to a difference between people's opinion and their actual behavior.

Hypothesis 4: People say that they are not inclined to give up privacy for additional benefits.

4. PERCEPTIONS OF DEVELOPERS

The results from the previous section describe the broad concerns for the respondents of our study overall. Here, we show the important results of doing a differential analysis for respondents who are developers (267 out of 408 respondents) versus users of software systems.

4.1 Privacy Concerns

Data distortion: 49.1% of developers believe that data distortion is an important privacy concern. The percentage of users, on the other hand, is 64.5%. The difference between these two groups is statistically significant ($p = 0.003$). (Note: We used the Z-test for equality of proportions for the rest of this section, unless otherwise noted.)

Data aggregation: 52.1% of developers believe that data aggregation is an important privacy concern. The percentage of users, on the other hand, is 63.1%. The difference between them is statistically significant ($p = 0.04185$). It seems that developers trust their systems more than users when it comes to wrong interpretation of sensitive data.

Data criticality: Developers believe that "name and personal data" ($p = 0.038$) and "interaction" ($p = 0.082$) are more critical for privacy compared to users. On the other hand, for the remaining four categories (content, location, preferences, metadata), there is no statistically significant difference between the perceptions of developers and users ($p > 0.2$ for all). We used Welch's Two Sample t-test here.

Less privacy for added functionality: A larger fraction of developers (43.3%) would accept less privacy for added or intelligent functionality of the system compared to 31.2% of users ($p = 0.002$).

Hypothesis 5: Developers are more concerned about interaction, name, and personal data whereas users are more concerned about data distortion and data aggregation.

4.2 Measures to Reduce Concerns

Developers and reducing concerns: A larger fraction of developers feel that data anonymization (71.2%) is a better option to reduce privacy concerns as compared to pri-

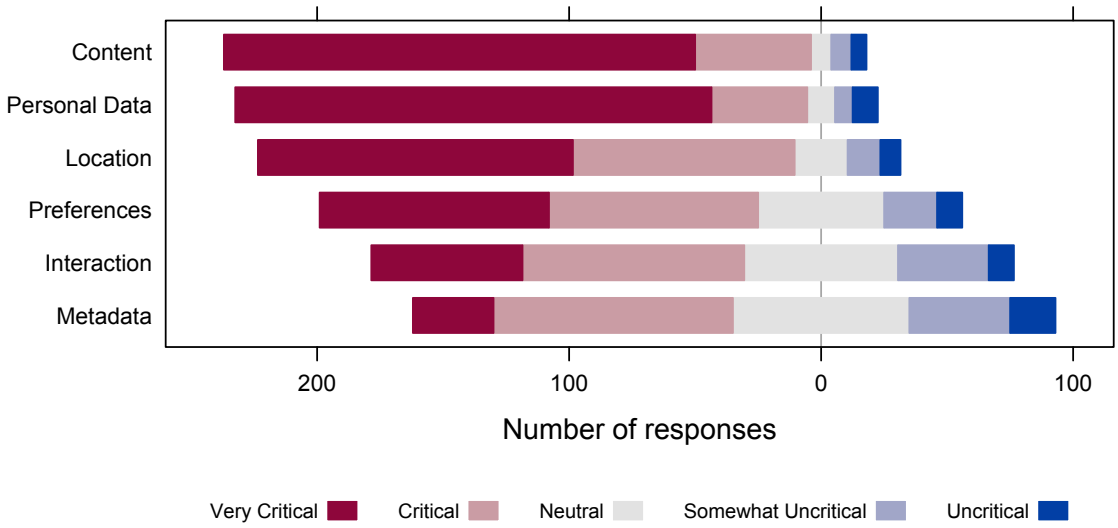

Figure 2: How *critical* would you rate the collection of the following data?

	Content	Personal Data	Location	Preferences	Interaction	Metadata
Content	–					
Personal Data		–				
Location	+	+	–			
Preferences	+++	++	+	–		
Interaction	+++	+++	++	+	–	
Metadata	+++	+++	+++	++	+	–

Table 3: How significant is the difference in *data criticality*? (p-values: 0 '+++' e^{-11} '++' e^{-6} '+' 0.01 ' ' 1) The rows and columns are ordered from most to least critical. For each cell, t-tests compare if the difference in criticality is statistically significant. For example, the difference between interaction and content is statistically significant for $p < e^{-11}$.

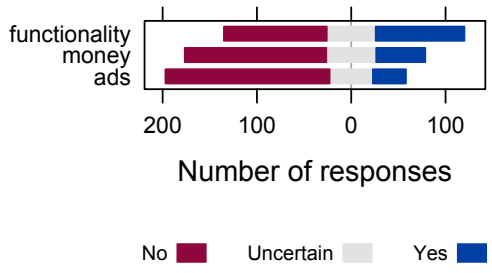

Figure 3: Would users accept *less* privacy for the following?

vacy policies or privacy laws (both, 56.9%) ($p = 0.0006$). 66.3% of developers prefer providing details on data usage for mitigating privacy concerns compared to privacy policies (56.9%) ($p = 0.03$).

Similarly, 20.2% of developers feel that privacy policies will *not* reduce privacy concerns whereas only 11.2% feel that providing details on data usage will *not* be beneficial ($p = 0.004$).

Users and reducing concerns: In contrast, for users, there is no statistically significant difference between their perception on privacy policies, laws, anonymization, and providing usage details. ($0.6 < p < 1$ for all combinations).

Hypothesis 6: Developers prefer anonymization and providing usage details as measures to reduce privacy concerns. Users, on the other hand, do not have a strong preference.

5. THE ROLE OF GEOGRAPHY

In this section, we present the results of the differential analysis based on the geography of respondents. We asked respondents to specify with which region they identified themselves with. The options were North America, South America, Europe, Asia/Pacific, Africa, and other. Since we have only seven responses from South America

and Africa combined, we focus on the differences between the others.

Data Criticality: For the different types of data that are critical for privacy — content of documents, metadata, interaction, user location, name or personal data, user preferences — respondents were asked to rate on a semantic scale from 1-5 on how critical each data item was, with 1 being "Very Critical" and 5 being "Uncritical."

There is a statistically significant difference between respondents from North America, Europe, and Asia/Pacific. We used Welch's Two Sample t-test to compare the ratings given by respondents. Respondents in North America think that all items are *less critical* (overall mean across the six items, i.e., the mean of the means of the six types of data, is 2.31) than respondents in Europe (overall mean: 1.87) for $p = 3.144e^{-8}$.

Similarly, respondents from North America think all items are less critical that those in Asia/Pacific (overall mean: 2.01) for $p = 0.037$. On the other hand, there is no statistically significant difference between respondents in Europe and Asia/Pacific ($p > 0.28$).

Less privacy for added functionality: A larger fraction of respondents in Europe (50.6%) would *not* give up privacy for added functionality. In North America, on the other hand, this fraction is 24.1%. The difference between the two regions is statistically significant ($p = 0.0001$). (Note: We used the Z-test for equality of proportions for the rest of this section, unless otherwise noted.)

Hypothesis 7: People from North America are more willing to give up privacy and feel that different types of data are less critical for privacy compared to people from Europe.

Concerns about data sharing versus data distortion: A larger fraction of respondents in North America (88.9%) feel that data sharing is a concern compared to 46.3% for data distortion ($p = 6.093e^{-6}$). On the other hand, there is no statistically significant difference among respondents in Asia/Pacific ($p > 0.67$).

Concerns about data sharing versus data breach: In Europe, a larger fraction of the respondents (94.3%) are concerned about data breaches as compared to 76.4% for data sharing. The difference is statistically significant ($p = 5.435e^{-6}$). On the other hand, there is no statistically significant difference among respondents in North America ($p > 0.12$).

Laws versus usage details: In Europe, a larger fraction of respondents (75.9%) feel that providing details on how the data is being used will reduce privacy concerns as opposed to 58.1% who feel that privacy laws will be effective ($p = 0.00063$). On the other hand, there is no statistically significant difference among respondents in North America, where the percentage of respondents are 67.9% and 64.2% respectively ($p > 0.43$).

Usage details versus privacy policies: A larger fraction of respondents in Europe (75.9%) feel that providing usage details can mitigate concerns compared to 63.2% for privacy policy ($p = 0.015$). On the other hand, there is no statistically significant difference among respondents in North America ($p > 0.32$).

Hypothesis 8: People from Europe feel that providing usage details can be more effective for mitigating privacy concerns than privacy laws and privacy policies whereas people from North America feel that these three options are equally effective.

6. DISCUSSION

We discuss our results, potential reasons, and the implications for software developers and analysts. We also reflect on the limitations and threats to validity of our results.

6.1 Privacy Communication Gap

Our results from Sections 4 and 5 show there is a definite gap in privacy expectations and needs between users and developers and between people from different regions of the world. Developers have assumptions about privacy, which do not always correspond to what users need. Developers seem to be less concerned about data distortion and aggregation compared to users. It seems that developers trust their systems more than users when it comes to wrong interpretation of privacy critical data. Unlike users, developers prefer anonymization and providing usage details for mitigating privacy concerns. If the expectations and needs of users do not match those of developers, developers might have wrong assumptions and might end up making wrong decisions when designing and building software systems.

In addition, privacy is not a universal requirement as it appears to have an internationalization aspect to it. Different regions seem to have different concrete requirements and understanding for privacy. Our results confirm that there exist many cultural differences between various regions of the world as far as privacy is concerned. The recent NSA PRISM scandal has also brought these differences into sharp focus. A majority of Americans considered NSA's accessing personal data to prevent terrorist threats more important that privacy concerns [13]. In contrast, there was widespread "outrage" in Europe over these incidents [15]. It also led to an article in the New York Times by Malte Spitz, a member of the German Green Party's executive committee, titled "Germans Loved Obama. Now We Don't Trust Him." [31]. These differences, both in terms of laws and people's perceptions, should be considered carefully when designing and deploying software systems.

Data privacy is often an implicit requirement: everyone talks about it but no one specifies what it means and how it should be implemented. This topic also attracts the interests of different stakeholders including users, lawyers, sales people, and security experts, which makes it even harder to define and implement. One important result from our study is that while almost all respondents agree about the importance of privacy, the understanding of the privacy issues and the measures to reduce privacy concerns are divergent. This calls for an even more careful and distinguished analysis of privacy when designing and building a system. We think that privacy should become an explicit requirement, with measurable and testable criteria. We also think that privacy should also become a main design criteria for developers as software systems are collecting more and more data about their users [14]. To this end, we feel that there is a need to develop a standard survey for privacy that software teams can customize and reuse for their projects and

users. Our survey can be reused to conduct additional user studies on privacy for specific systems. Our results can also serve as a benchmark for comparing the data. This can help build a body of knowledge and provide guidelines such as best practices.

6.2 The Security Dimension of Privacy

We think that people are more concerned about data breaches and data sharing as there have been many recent instances that have gotten a lot of news and media coverage. To list a few recent examples, Sony suffered a massive data breach in its Playstation network that led to the theft of personal information belonging to 77 million users [6]. 160 million credit card numbers were stolen and sold from various companies including Citibank, the Nasdaq stock exchange, and Carrefour [29]. The Federal Trade Commission publicly lent support to the "Do-Not-Track" system for advertising [4]. Compared to these high-profile stories, we feel that there have been relatively few "famous" instances of privacy problems caused by data aggregation or data distortion yet.

There is a large body of research that has advanced the state-of-the-art in security (encryption) and authorization. One short-term implication for engineers and managers is to systematically implement security solutions when designing and deploying systems that collect user data, even if it is not a commercially or politically sensitive system. This would significantly and immediately reduce privacy concerns. For the medium-term, more research should be conducted for deployable data aggregation and data distortion solutions.

As far as mitigating privacy concerns, our results show that there is more disagreement. We believe that the reason for this is because online privacy concerns are a relatively recent phenomenon. Due to this, people are not sure which approach works best and might be beneficial in the long run.

6.3 Privacy Framework

The long-term goal of this study is to develop a universal, empirically grounded framework for implementing privacy requirements. Some of the lessons learned from our study can be translated into concrete qualities and features, which should be part of such a framework. This includes:

- Anonymization: This is perhaps the most well-known privacy mitigating technique and seems to be perceived as an important and effective measure by both users and developers. Developers should therefore use anonymization algorithms and libraries.
- Default encryption: As users are mainly concerned about the loss and abuse of their data, systems collecting user data should implement and activate encryption mechanism for storing and transmitting these data. In Facebook, e.g., the default standard communication protocol should be HTTPS and not HTTP.
- Fine-grained control over the data: Users become less concerned about privacy if the system provides a mechanism to control their data. This includes activating and deactivating the collection at any time, the possibility to access and delete the raw data and processed data, and define who should have access to what data.
- Interaction data first: Users have a clear preference of the criticality of the different types of data collected about them. Therefore, software researchers and designers should first try to implement their systems based on collecting and mining interaction data instead of content of files and documents. Research has advanced a lot in this field in, especially, recommender systems.
- Time and space-limited storage: The storage of data about users should be limited in time and space. The location where the data is stored is an important factor for many respondents. Therefore, systems should provide options to choose the location to store privacy sensitive data.
- Privacy policies, laws, and usage details: Users rated all these options as equally effective for mitigating privacy concerns. Therefore, developers could utilize any of these options, thus giving them better flexibility in the design and deployment of software systems.

6.4 Limitations and Threats to Validity

There are several limitations to our study, which we now discuss. The first limitation is selection bias — respondents who volunteered to fill out our survey were self-selected. Such selection bias implies that our results are only applicable to the volunteering population and may not necessarily generalize to other populations. The summaries have helped us identify certain trends and hypotheses and these should be validated and tested by representative samples, e.g., for certain countries. In contrast, the pseudo-experiment conducted within our set of respondents, enabled us to identify statistically significant relationships and correlations. Hence, many of our results deliberately focus on correlations and cross-tabulations between different populations.

As for internal validity, we are aware that by filling out a brief survey, we can only understand a limited amount of concerns that the respondents have in mind. Similarly, the format and questions of the survey might constrain the expressiveness of some of the respondents. We might have missed certain privacy concerns and techniques to reduce concerns by the design of the survey. We tried to mitigate this by providing a few open-ended questions that respondents could use to tell us additional things they had in mind.

As with any online survey, there is a possibility that respondents did not fully understand the question or chose the response options arbitrarily. We also conducted several pilot tests, gave the option to input comments, and the incompletion rate is relatively small. We included a few validation questions and we only report responses in this paper from respondents who answered these questions correctly. We also provided two versions of the survey, in English and German, to make it easier for non-native speakers.

In spite of these limitations, we managed to get a large and diverse population that filled out our survey. This gives us confidence about the overall trends reported in this paper.

7. RELATED WORK

There has been a lot of research in privacy and security in different research communities. We highlight the important related work in this section.

Many recent studies on online social networks show that there is a (typically, large) discrepancy between users' intentions for what their privacy settings should be versus what they actually are. For example, Madejski et al. report that, in their study on Facebook, 94% of their participants ($n = 65$) were sharing something they intended to hide and 85% were hiding something that they intended to

share [23,24]. Liu et al. [23] found that Facebook's users' privacy settings match their expectations only 37% of the time. A recent longitudinal study by Stutzman et al. [34] shows how privacy settings for users on Facebook have evolved over a period of time. These studies have focused on privacy settings in a specific online system whereas our study was designed to be agnostic to any particular online systems. Further, the main contribution of these studies is to show that there is a discrepancy between what the settings are and what they should be and how settings evolve over time. Our study aims to gain a deeper understanding of what the requirements are and how they change across geography and software development experience.

Fang and LeFevre [17] proposed an automated technique for configuring a user's privacy settings in online social networking sites. Paul et al. [27] present using a color coding scheme for making privacy settings more usable. Squicciarini, Shehab, and Paci [33] propose a game-theoretic approach for collaborative sharing and control of images in a social network. Toubiana et al. [37] present a system that automatically applies users' privacy settings for photo tagging. All these papers propose new techniques that are targeted to making privacy settings "better" (i.e., more usable, more visible) from a user's perspective. Our results help development teams decide when and which of these techniques should be implemented. We focus more on a broader requirements and engineering perspective of privacy than on a specific technical perspective.

There has been a lot of recent work on the economic ramifications of privacy. For example, Acquisti et al. [1] (and the references therein) conducted a number of field and online experiments to investigate the economic valuations of privacy. In Section 3.4, we discussed whether users would give up privacy for additional benefits like discounts or fewer advertisements. Our study complements and contrasts the work of Acquisti et al. as described earlier.

There has also been a lot of work related to data anonymization and building accurate data models for statistical use (e.g., [2, 16, 22, 28, 39]). These techniques aim to preserve certain properties of the data (e.g., statistical properties like average) so they can be useful in data mining while trying to preserve privacy of individual records. Similar to these, there has also been work on anonymizing social networks [8] and anonymizing user profiles for personalized web search [41]. The broad approaches include aggregating data to a higher level of granularity or adding noise and random perturbations. There has been research on breaking the anonymity of data as well. Narayanan and Shmatikov [26] show how it is possible to correlate public IMDb data with private anonymized Netflix movie rating data resulting in the potential identification of the anonymized individuals. Backstrom et al. [5] and Wondracek et al. [40] describe a series of attacks for de-anonymizing social networks.

In the Software Engineering community, there have been recent papers on privacy, which mainly focus on data anonymization techniques. Clause and Orso [12] propose techniques for the automated anonymization of field data for software testing. They extend the work done by Castro et al. [11] using novel concepts of path condition relaxation and breakable input conditions resulting in improving the effectiveness of input anonymization. Taneja et al. [36] and Grechanik et al. [19] propose using k-anonymity [35] for privacy by selectively anonymizing certain attributes of a database for software testing. They propose novel approaches using static analysis for selecting which attributes to anonymize so that test coverage remains high. Our work complements these existing papers as respondents in our study considered anonymization an effective technique for mitigating privacy concerns. Thus, these techniques could be used as part of a privacy framework. There have also been some recent papers on extracting privacy requirements from privacy regulations and laws [9, 10]. These could be part of the privacy framework as well and help in reducing the impact due to cultural differences for privacy.

Finally, many authors of papers in the software engineering and requirements engineering communities mention privacy in their discussion or challenges section. But in many cases, there is little evidence and grounded theory about what, how, and in which context privacy concerns exist and what the best measures for addressing them are. Our study helps in clarifying these concerns and measures and comparing the different perceptions of people.

8. CONCLUSION

In this paper, we conducted a study to explore the privacy requirements for users and developers in online systems, such as Amazon and Facebook, that collect and store data about the user. Our study consisted of 408 valid responses representing a broad spectrum of respondents: people with and without software development experience and people from North America, Europe, and Asia. While the broad majority of respondents (more than 91%) agreed about the importance of privacy as a main issue for modern software systems, there was disagreement concerning the concrete importance of different privacy concerns and the measures to address them. The biggest concerns about privacy were data breaches and data sharing. Users were more concerned about data aggregation and data distortion than developers. As far as mitigating privacy concerns, there was little consensus on the best technique among users. In terms of data criticality, respondents rated content of documents as most critical and metadata as least critical.

We also compared if there was any difference in privacy perceptions based on geographic location of the respondent. We observed many differences in our study. Respondents from North America, for example, consider all types of data as less critical for privacy than respondents from Europe or Asia/Pacific. Respondents from Europe are more concerned about data breaches than data sharing whereas respondents from North America are equally concerned about the two.

Finally, we gave some insight into a framework and a set of guidelines on privacy requirements for developers when designing and building software systems. Our results can help establish such a framework, which can be a catalog of privacy concerns and measures, a questionnaire to assess them, and perhaps a library of reusable privacy functionality.

9. ACKNOWLEDGMENTS

We would like to thank all the respondents for filling out our online survey. We would also like to thank Timo Johann, Mathias Ellman, Zijad Kurtanovic, Rebecca Tiarks, and Zardosht Hodaie for help with the German translations. Sheth and Kaiser are members of the Programming Systems Laboratory is funded in part by NSF CCF-1161079, NSF CNS-0905246, and NIH 2 U54 CA121852-06.

10. REFERENCES

[1] A. Acquisti, L. John, and G. Loewenstein. What is privacy worth? In *Workshop on Information Systems and Economics (WISE)*, 2009.

[2] D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *PODS '01: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 247–255, New York, NY, USA, 2001. ACM.

[3] T. Anderson and H. Kanuka. E-research: Methods, strategies, and issues. 2003.

[4] J. Angwin and J. Valentino-Devries. FTC Backs Do-Not-Track System for Web. `http://online.wsj.com/article/SB10001424052748704594804575648670826747094.html`, December 2010.

[5] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 181–190, New York, NY, USA, 2007. ACM.

[6] L. B. Baker and J. Finkle. Sony PlayStation suffers massive data breach. `http://www.reuters.com/article/2011/04/26/us-sony-stoldendata-idUSTRE73P6WB20110426`, April 2011.

[7] M. Barbaro, T. Zeller, and S. Hansell. A face is exposed for AOL searcher no. 4417749. `http://www.nytimes.com/2006/08/09/technology/09aol.html?_r=1`, August 2006.

[8] S. Bhagat, G. Cormode, B. Krishnamurthy, and D. Srivastava. Privacy in dynamic social networks. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 1059–1060, New York, NY, USA, 2010. ACM.

[9] T. D. Breaux and A. I. Anton. Analyzing regulatory rules for privacy and security requirements. *IEEE Transactions on Software Engineering*, 34(1):5–20, 2008.

[10] T. D. Breaux and A. Rao. Formal analysis of privacy requirements specifications for multi-tier applications. In *RE'13: Proceedings of the 21st IEEE International Requirements Engineering Conference (RE'13)*, Washington, DC, USA, July 2013. IEEE Society Press.

[11] M. Castro, M. Costa, and J.-P. Martin. Better bug reporting with better privacy. In *Proceedings of the 13th international conference on Architectural support for programming languages and operating systems*, ASPLOS XIII, pages 319–328, New York, NY, USA, 2008. ACM.

[12] J. Clause and A. Orso. Camouflage: automated anonymization of field data. In *Proceeding of the 33rd international conference on Software engineering*, ICSE '11, pages 21–30, New York, NY, USA, 2011. ACM.

[13] J. Cohen. Most Americans back NSA tracking phone records, prioritize probes over privacy. `http://www.washingtonpost.com/politics/most-americans-support-nsa-tracking-phone-records-prioritize-investigations-over-privacy/2013/06/10/51e721d6-d204-11e2-9f1a-1a7cdee20287_story.html`, June 2013.

[14] L. F. Cranor and N. Sadeh. A shortage of privacy engineers. *Security & Privacy, IEEE*, 11(2):77–79, 2013.

[15] S. Erlanger. Outrage in Europe Grows Over Spying Disclosures. `http://www.nytimes.com/2013/07/02/world/europe/france-and-germany-piqued-over-spying-scandal.html`, July 2013.

[16] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *PODS '03: Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 211–222, New York, NY, USA, 2003. ACM.

[17] L. Fang and K. LeFevre. Privacy wizards for social networking sites. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 351–360, New York, NY, USA, 2010. ACM.

[18] D. Fletcher. How Facebook Is Redefining Privacy. `http://www.time.com/time/business/article/0,8599,1990582.html`, May 2010.

[19] M. Grechanik, C. Csallner, C. Fu, and Q. Xie. Is data privacy always good for software testing? *Software Reliability Engineering, International Symposium on*, 0:368–377, 2010.

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

[21] J. Jacoby and M. S. Matell. Three-point likert scales are good enough. *Journal of Marketing Research*, 8(4):pp. 495–500, 1971.

[22] N. Lathia, S. Hailes, and L. Capra. Private distributed collaborative filtering using estimated concordance measures. In *RecSys '07: Proceedings of the 2007 ACM conference on Recommender systems*, pages 1–8, New York, NY, USA, 2007. ACM.

[23] Y. Liu, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Analyzing facebook privacy settings: user expectations vs. reality. In *Proc. of the 2011 SIGCOMM Conf. on Internet measurement conf.*, pages 61–70, 2011.

[24] M. Madejski, M. Johnson, and S. M. Bellovin. A study of privacy settings errors in an online social network. *Pervasive Computing and Comm. Workshops, IEEE Intl. Conf. on*, 0:340–345, 2012.

[25] C. C. Miller. Privacy Officials Worldwide Press Google About Glass. `http://bits.blogs.nytimes.com/2013/06/19/privacy-officials-worldwide-press-google-about-glass/`, June 2013.

[26] A. Narayanan and V. Shmatikov. How to break anonymity of the netflix prize dataset. *CoRR*, abs/cs/0610105, 2006.

[27] T. Paul, M. Stopczynski, D. Puscher, M. Volkamer, and T. Strufe. C4ps: colors for privacy settings. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, pages 585–586, New York, NY, USA, 2012. ACM.

[28] H. Polat and W. Du. Privacy-preserving collaborative filtering using randomized perturbation techniques. In

Data Mining, 2003. ICDM 2003. Third IEEE International Conference on, pages 625–628, Nov. 2003.

[29] N. Popper and S. Sengupta. U.S. Says Ring Stole 160 Million Credit Card Numbers. `http://dealbook.nytimes.com/2013/07/25/arrests-planned-in-hacking-of-financial-companies/`, July 2013.

[30] R. L. Rosnow and R. Rosenthal. *Beginning behavioral research: A conceptual primer* . Prentice-Hall, Inc, 1996.

[31] M. Spitz. Germans Loved Obama. Now We Don't Trust Him. `http://www.nytimes.com/2013/06/30/opinion/sunday/germans-loved-obama-now-we-dont-trust-him.html`, June 2013.

[32] R. C. Sprinthall and S. T. Fisk. *Basic statistical analysis*. Prentice Hall Englewood Cliffs, NJ, 1990.

[33] A. C. Squicciarini, M. Shehab, and F. Paci. Collective privacy management in social networks. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 521–530, New York, NY, USA, 2009. ACM.

[34] F. Stutzman, R. Gross, and A. Acquisti. Silent listeners: The evolution of privacy and disclosure on facebook. *Journal of Privacy and Confidentiality*, 4(2):2, 2013.

[35] L. Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, 2002.

[36] K. Taneja, M. Grechanik, R. Ghani, and T. Xie. Testing software in age of data privacy: a balancing act. In *Proceedings of the 19th ACM SIGSOFT symposium and the 13th European conference on Foundations of software engineering*, SIGSOFT/FSE '11, pages 201–211, New York, NY, USA, 2011. ACM.

[37] V. Toubiana, V. Verdot, B. Christophe, and M. Boussard. Photo-tape: user privacy preferences in photo tagging. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, pages 617–618, New York, NY, USA, 2012. ACM.

[38] T. L. Tuten, D. J. Urban, and M. Bosnjak. Internet surveys and data quality: A review. *Online social sciences*, page 7, 2000.

[39] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis. State-of-the-art in privacy preserving data mining. *SIGMOD Rec.*, 33(1):50–57, 2004.

[40] G. Wondracek, T. Holz, E. Kirda, and C. Kruegel. A practical attack to de-anonymize social network users. In *Security and Privacy (SP), 2010 IEEE Symposium on*, pages 223–238, 2010.

[41] Y. Zhu, L. Xiong, and C. Verdery. Anonymizing user profiles for personalized web search. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 1225–1226, New York, NY, USA, 2010. ACM.