"If You Put All The Pieces Together..." – Attitudes Towards Data Combination and Sharing Across Services and Companies

Igor Bilogrevic
Google
Zurich, Switzerland
ibilogrevic@google.com

Martin Ortlieb
Google
Zurich, Switzerland
mortlieb@google.com

ABSTRACT

Online services often rely on processing users' data, which can be either provided directly by the users or combined from other services. Although users are aware of the latter, it is unclear whether they are comfortable with such data combination, whether they view it as beneficial for them, or the extent to which they believe that their privacy is exposed. Through an online survey (N=918) and follow-up interviews (N=14), we show that (1) comfort is highly dependent on the type of data, type of service and on the existence of a direct relationship with a company, (2) users have a highly different opinion about the presence of benefits for them, irrespectively of the context, and (3) users perceive the combination of online data as more identifying than data related to offline and physical behavior (such as location). Finally, we discuss several strategies for companies to improve upon these issues.

Author Keywords

Privacy; Attitudes; User Study; Data Combination; Data Sharing.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

The online experience of users is more personalized and contextual than ever. From accurate movies recommendations to virtual, voice-operated assistants, online services are able to tailor the content provided to users by extracting relevant information derived from their interactions with the services.

Usually, the extent and limits to the collection and use of users' data are described in the provider's data or privacy policies, which are part of the general "Terms of Use" of the service. By agreeing to such terms, users give consent to the collection and processing of their data in accordance with the data policy. Over time, online services can undergo a series of changes, either in terms of their functionality or as a consequence of mergers, acquisitions or new laws, which can result in updates to their data policy [17, 20, 31]. In spite of the fact that such policies are nominally written for the users, they are often unread or perceived as being unclear [3, 53], mainly due to overly complex language that is beyond the grasp of most of the Internet users [25, 36].

The collection of personal data often raises privacy concerns, as users of all ages find it difficult to understand the benefits they get from such a collection [3, 30, 41, 58]. Furthermore, the use of multiple accounts from different providers to manage personal and professional identities makes it even more challenging to have a correct representation of where different types of data are stored and who can access them [13]. As a result, users often struggle to understand what could be the consequences of online information sharing [22].

Several studies investigated the concerns raised by online users regarding their sensitivity when sharing different types of personal data with third-parties [3, 11, 18, 62]. For instance, [33] showed that data related to health, communications, location and online browsing are considered as highly sensitive. As nowadays more and more data is combined from different services under a more unified privacy policy (e.g., through mergers and acquisitions), different data types can be combined in a single service to provide an enhanced experience for the users. Yet, little is known about their privacy attitudes when different types of data are combined and shared across different services and companies, or about how risky different combinations of data could be in terms of user re-identification.

In this paper, we answer these two questions by means of a brand-blind, online survey with 918 participants, followed by in-depth interviews with 14 participants. Specifically, we investigate users' privacy attitudes when multiple types of personal data are shared and combined across different online services and companies. By quantifying their comfort with sharing data, as well as the perceived benefits and re-identification risks of several combinations of individual data types, we provide a comprehensive assessment of privacy attitudes that is representative for the three most popular online domains (search engines, social networks and shopping services) [39]. Finally, based on our data, we propose mitigation strategies for companies that could increase users' comfort when sharing data and perceived benefits. To the best of our

CHI'16, May 07-12, 2016, San Jose, CA, USA
ACM 978-1-4503-3362-7/16/05.
http://dx.doi.org/10.1145/2858036.2858432.

knowledge, this is the first study to investigate users' privacy attitudes towards data combination and sharing across different services and companies.

Our results show that, in general, comfort with sharing data with third-parties depends on the type of data, the type of service and the existence of a direct relationship with the service provider. One additional aspect that emerges as a positive driver for user comfort is expected liability for first-party data holders in case of a possible information leak by third-party companies who may have received parts of users' data. In terms of perceived benefits, users unsurprisingly believe that there is an imbalance between the benefits that they experience, as compared to the benefits that the company gets when accessing and sharing their data. Furthermore, across the three different online scenarios, a combination of not-so-identifying pieces of information is perceived as being more personally identifying than the single most identifying type of data, suggesting that the availability of different complementary information is a significant aspect of the data collection context that determines sensitivity, and that it should be an important factor in defining communication strategies about user data practices.

Our work focuses on dimensions related to users' perceptions and attitudes towards privacy when combining information. Technical solutions to provide privacy for data aggregation and inference, although important, are only of marginal interest for the presentation of this paper. Hence, we refer the reader to [16, 19, 27, 65] for a comprehensive treatment of some of the technical solutions.

The remainder of the paper is structured as follows. First, we introduce the related work, by treating three different subgroups of studies focusing on personalization, data sensitivity and re-identification. Next, we define the research questions and describe the study methodology. Afterwards, we present the results and discuss their implications, before concluding the paper and outlining the future work.

RELATED WORK

Prior studies that investigated the challenges related to people's privacy attitudes towards data combination can be grouped, from a high-level perspective, into three different categories. First, we start by providing some background about (i) online users' attitudes towards personalization and privacy [5, 6, 14, 56, 58, 60], which is directly related to users' opinions about (ii) data sensitivity and aggregation [4, 9, 28, 33, 34, 62]. Finally, we tackle the topic of (iii) user re-identification [7, 23, 24, 32, 42, 48, 54, 55, 57] , as it follows the previous two (personalization and aggregation) by exploring users' concerns about one potential negative consequence of personalization and data aggregation.

Personalization and Privacy Attitudes

Awad et al. [6] are among the first to discover the existence of an apparent paradox between online privacy and personalization: Users who value transparency (a dimension of privacy) are also less willing to be profiled (thus resisting personalization). The authors posit that, instead of developing features for the users who value privacy, firms should focus on those who value personalization. In addition to transparency, Chellappa et al. [14] identified trust as a dimension of privacy that positively influences users' willingness to use personalization services. In particular, the authors recommend that companies shall build user trust and allow different levels of personalization, in order to acquire and use users' information. By taking a broader perspective towards technology, Toch et al. [60] point to three emerging areas of privacy risks for personalization services, which are social-based personalization, behavioral profiling (through the aggregation of online data from different sources) and location-based personalization. To mitigate such risks, the authors propose strategies based on data anonymization, client-side processing and algorithmic privacy-preserving solutions. Moreover, Acquisti [1] shows how the latter can be used to enhance the aggregate welfare when used for personalization. Regarding client-side personalization, Sutanto et al. [56] studied how users' comfort with using personalized services changes depending on where the personalization takes place, i.e., locally on the mobile device or in a marketer's datacenter. By building on the uses and gratifications theory (UGT, [37, 49]), they show that users can both benefit from personalization and mitigate their privacy concerns if their data is stored and processed locally on their mobile device, instead of being processed by a third-party marketer. Our work explores personalization aspects related to comfort with sharing data, perceived benefits and re-identification risks for the users.

Data Sensitivity and Aggregation

Numerous studies have shown that different types of data (such as age, financial information, health records and location) have different levels of sensitivity for the users [11, 33, 34, 62], and that sensitivity depends on the situation in which the data is collected and used [4, 28, 34]. According to a recent study conducted in the US [33], the most sensitive pieces of data for online users are (in decreasing order): social security number, health record, content of communications and location. In order to capture the nuances in sensitivity across different contexts, in this study we focus on the three most popular online contexts [39]: search engines, social networks and online shopping services. Within each of these contexts, we study users' comfort with sharing different types of data with third parties.

User Re-identification

Re-identification of users from publicly available information is a topic that has received significant attention from the research community. From the initial work by Sweeney [57], based on US Census data from 1990, to more recent de-anonymization attacks on search queries [7], movie ratings [42], music listening history [32] and genomic data [23, 24, 29], researchers have shown that it is possible to identify an individual or some of her traits with high precision, by accessing data that is easily (and most of the time publicly) available on the Internet. For instance, a combination of [5-digit US ZIP code, gender, date of birth] is sufficient to uniquely identify 87% of the US citizens [57], individuals could be identified from a set of 650,000 anonymized search queries [7], and genomic data released by one member of a

family can be used to infer health predispositions of the other members via side-channel information (such as kin relationships, available on online social networks [23]). As a complement to prior studies, our work provides a user-centric assessment of their perceptions about the re-identification risks, by analyzing both individual data types (such as age, financial information and location), as well as their combinations.

STUDY GOAL AND METHODOLOGY

The goal of our study was to identify and quantify users' privacy concerns when their data is used and shared across different online services and companies. As users' perceptions about privacy vary according to the context, we focused on a comprehensive but limited set of data types and online contexts, as described in the following subsection.

Privacy as Data Sensitivity, Benefits and Risks

As privacy is a multi-faceted concept for which there is no single definition [40, 63], we needed to frame it in clear and understandable terms for the purpose of this study. To this end, we chose to focus on a subset of data types in a limited set of popular online contexts [39]. In particular, we solicited our participants' opinions about the following 5 categories of personal data, which include most of the sensitive data types listed in [33]: Information about you (such as age, gender and interests), contact details (such as home address and phone number), online browsing history (the list of websites you visited), online search history (the terms you searched for) and payment details. For each of the aforementioned data categories, we elicited responses with respect to (i) data sensitivity (by quantifying users' comfort with sharing their data with different types of online services), (ii) incentives (by quantifying the perceived benefits for the users and for the company when sharing users' data) and (iii) users' perceived risk of re-identification (by quantifying and ranking different data categories and combinations thereof in terms of their potential to identify a single user).

Experimental Scenarios: Search Engine, Online Social Network and Online Shopping Service

In order to reduce inter-subject variability when answering our questions, and to retain as much contextual integrity as possible [8, 43] we assigned each participant to one of the following three popular online scenarios: (1) Search engine, (2) social network, (3) shopping service. In each scenario, the participant was told that a fictitious company runs three different services: The main service (i.e., the company SearchCo/SocialCo/RetailCo that runs the search engine/social network/shopping service, respectively), and two other secondary services (e.g., SearchCo runs its main service, the search engine, as well as a social network and a shopping service). Moreover, in each scenario, participants were told to assume that they have an account on the main service, but not on the secondary services of the same company. For instance, Figure 1 shows the diagram for the search engine scenario. We chose to assign a participant to only one of the three scenarios to reduce respondents' fatigue and cognitive load, as well as to increase focus during the study.

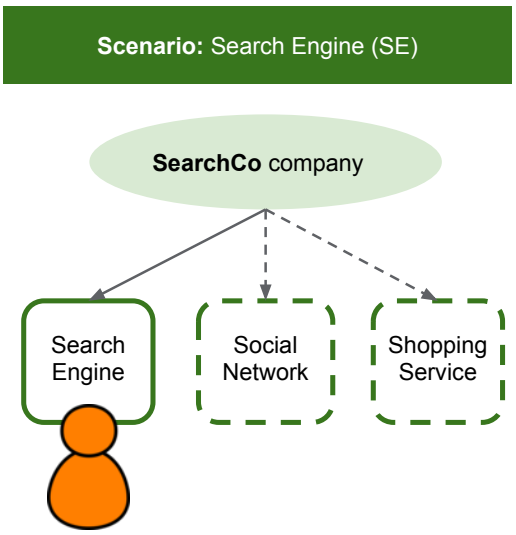

Figure 1. Search Engine scenario diagram, where the participant is assumed to have an account on the main service, (Search Engine) of SearchCo, but not on the other two secondary services (Social Network and Shopping service).

User Study

We structured the study in two phases. In Phase I, we explored the breadth of the research question by eliciting responses in an online survey; in Phase II, we performed an in-depth analysis of the salient findings and trends that we observed in the previous phase, by means of semi-structured interviews, which aimed to provide greater details and interpretation cues as to the previous findings.

Phase I: Online Survey

We created an online survey to elicit a broad range of responses about users' comfort with sharing personal data, the perceived benefits for them and for companies when sharing users' data, and the risks of user re-identification from individual data types and for some of their combinations. We administered our survey through a third-party online vendor, where we removed any logo or text that would link our survey with our company, in order to preserve brand-blindness and avoid any response bias due to factors such as brand reputation and past experience. All subjects were recruited on the Amazon Mechanical Turk (MTurk) platform, a widely-used service in the research community to recruit subjects for human-intelligence tasks [26, 35]. In order not to have any association with our company account or prior reputation, we created and used a new MTurk account for this study.

Survey structure: The survey had a total of 26 questions, including a screener, data access expectations, scenario-specific questions related to comfort and benefits, followed by questions about risks of re-identification when sharing data, demographics and Internet usage (the full questionnaire can be found as a Supplementary Information (SI) document). The format of the possible answers included free-text, multiple-choice selections and 5-point ordinal scales. [1] The struc-

[1]More details about each specific scale will be provided where appropriate.

ture of the survey was the following. First, a participant was randomly assigned to one of the three scenarios (i.e., either search engine, social network or shopping service), where she answered a question related to how she normally uses the main service. For example, in the search engine scenario, a participant answered the multiple-choice question "How do you normally use Online Search Engines (e.g., Google, Bing, Yahoo, Baidu, AOL)?" with either "I do not use Online Search Engines", or "I use at least one of them but I do not have an account on it", or "I use at least one of them but I do not sign in with my account", or "I use at least one of them and I sign in with my account". Participants that selected either one of the two last possible choices were allowed to continue with our survey, whereas those who selected either one of the first two choices were screened out. This approach would ensure that all participants are familiar with providing some personal information (such as email, name, address, interests) to the online service, as usually required during the account set-up process.

Participants who passed the screener question were first presented with 3 questions in which we asked them to rate, on a 6-point scale (where 1 means "multiple times per day" and 6 means "never") how often they expected their data to be accessed by the main service and shared with other companies. Next, they provided answers on a 5-point ordered scale (from "Not at all comfortable" to "Extremely comfortable") to a set of scenario-specific questions about their comfort when different types of data are accessed and shared by different services of the same company. Following that, participants were presented with questions related to the perceived benefits that there for them and for the company, when it accesses and shares different types of data with other companies. The answers to these questions were coded on a 5-point Likert scale as to the agreement with, for example, the statement "There is a benefit for me/SearchCo if it has access to my contact details (e.g., name, home address, email, photo)".

After the scenario-specific questions about comfort and benefits, the participants answered a set of common questions about the risks of re-identification for 10 individual types of personal information and for 4 combinations of these 10 individual types of information. In other words, they were asked to rank, from the most to the least personally identifying, 10 types of personal information, such as email address, ZIP code of their home, age, gender, interests, trace of GPS position over the last week and browsing history. Finally, they finished the survey with a set of demographics and Internet usage questions.

Survey validation: We assessed and refined the questionnaire by means of 3 cognitive walk-throughs, 2 expert checks and a pilot with 30 survey participants on MTurk. To detect possible misbehavior, we relied on the time to complete of the survey (as compared to the average) and we provided a survey completion code after participants submitted their last answer. We did not include any verification (or "dummy") questions, as prior work has shown this has does not significantly increase the quality of the responses [11, 12]. To qualify for our survey, MTurk participants were required to have at least 1 previous HIT approved and not to have previously participated in our pilot survey. Finally, several researchers with privacy and ethics training reviewed the survey questions before the experiment.

Participants recruitment: We recruited our participants on the MTurk platform, as privacy-related concerns from MTurk participants are found to be in between those from participants recruited through other providers [50]. The target population was comprised of US adults who reported having an account on at least one of the three online services (search engine, social network or shopping service). Each participant was remunerated with $2 for completing the survey (avg. completion time of 17 minutes, which corresponds to $7.1/hour).

Phase II: Semi-structured interviews

We structured the interviews that we conducted in Phase II with the goal of focusing on specific patterns and concepts that emerged from the analysis of the survey results from Phase I (more details are provided in the next section). We followed a similar structure as we did in Phase I, i.e., the same three online scenarios, but we allowed participants to articulate their opinions to a greater extent, and we invested more time for in-depth discussions of specific statements.

Interview structure: We created an interview script for each of the three online scenarios (search engines, social networks and shopping services). According to the answer to a pre-interview screening question (which is the same as the online interview screening question), we assigned each interviewee to one of the three online scenarios. We then followed the script and took short notes about the participants' answers in a separate document.

Logistics and participants recruitment: We informed the participants that we would be recording the audio and video of the session, and that they were free to stop the interview at any time, or to skip any question they wanted. Each interview lasted approximately 60 minutes, and participants were given an appreciation token of $75 in the form of a coupon. We recruited participants through an internal panel of human subjects volunteering to be part of user studies conducted by our company. Each participant was informed about the logistics of the study, signed a Non-Disclosure Agreement (NDA) and gave written consent to participate to this study.

Dataset Description

During Phase I of our study (September 15th – 23rd, 2014), 954 participants completed the survey, out of the 987 participants that started it. The average survey completion time was 17 minutes, with a standard deviation of 15 minutes. Out of the 954 who completed the survey, we registered 31 screen-outs (3%) and 4 exclusions due to potential cheating (2 with a wrong completion code, and 2 with a completion time smaller than 3 minutes, i.e., one standard deviation from the average) and 1 invalid entry in our database. In the end, we collected 918 valid survey response sets (96% of the total). In terms of demographics, 53% of the respondents were male, 68% were between 24 – 40 years of age, 13% reported working in science, IT or engineering, 9% were students and 98% stated

that they use the Internet more than 1 hour per day. Following the random scenario assignment, we obtained 32% of all the valid responses for the Search Engine scenario, 24% for the Online Social Network scenario, and 44% for the Online Shopping scenario.

In Phase II (November 18th – 21st, 2014) we conducted 14 semi-structured remote interview sessions with US adults, lasting 60 minutes each, via a proprietary video-conferencing application. The participants were selected from a large internal pool and had similar demographics as the ones of Phase I. To keep a balanced split across the three scenarios, we assigned 4 participants to the Search Engine scenario, and 5 to both the Online Social Networks and the Online Shopping scenario. The first author conducted and coded all the interviews, and the video recordings were used to complete the related notes. After the interviews, the authors presented the summary of the notes to our research group, and together we open-coded the notes (16 codes in total). Next, the authors clustered the codes into concepts, and we iteratively analysed the concepts to identify and refine the topics that we report hereafter.

RESULTS

In the following, we present our results by themes rather than by the study phases, as it enables us to present the findings from both phases within their context.

Comfort with Sharing Data

Overall, participants reported feeling the least comfortable with sharing information about their payment details (such as bank account and credit card numbers); on average, 78% of the respondents answered "Not at all comfortable" (on the 5-point scale [Not at all / Slightly / Moderately / Very / Extremely comfortable]) to a question related to their comfort when different services and companies access such information (question #8 of the questionnaire). Next, we find online search history (61%), browsing history (60%), contact details (59%) and information about you (45%). Interestingly, the level of comfort for the online search and browsing histories, two aspects related to the online identity and persona, was lower as compared to two aspects related to the physical world (contact details and generic demographic and behavioral information). We discuss this aspect further in the subsection "Risks of re-identification".

A within-scenario analysis shows that the differences in the comfort level across information types are significant, in all three scenarios (Figure 2). [2] In general, we notice a sharp difference between the first row (i.e., other company) and the last 3 rows (secondary and main service of the first-party company) across all scenarios, indicating that respondents' opinions are markedly influenced by the presence of a direct interaction with the company. The follow-up interviews revealed that for 50% (7 out of 14) of the interviewees, the absence of a first-party interaction with a third-party company makes them significantly less comfortable in knowing that it could access some of their data. For instance, participant #10 (P10) stated:

P10: *"Well... if I'm not even interacting with that company online, I don't really... I don't want them to have my info."*

In addition to a first-hand relationship with a company, interviewees mentioned four factors that influence their comfort level: (i) The control over the access to information (57% - in particular P3-P5,P7,P10-P13), (ii) whether their data is anonymized before being shared by a third-party company (43% - P1,P5,P7,P12-P14), (iii) the transparency about what information is accessed and for what purpose (29% - P3,P7,P9,P13), and (iv) the level of trust in the third-party company (29% - P2,P7,P12,P13). Moreover, participants often mentioned a combination of the aforementioned factors when explaining their lack of comfort.

Regarding control and transparency, interviewees noted that

P7: *"... if it [access to information] is, like, without my consent or something, or if it's automated or something, I probably would be 2 or 1 [on the 5-point scale where 1 means "Not at all comfortable" and 5 means "Extremely comfortable"]. But if it's like an option... that might be ok, only if it's like something that is brought to my attention."*

P4: *"It's about what you've allowed... I think it's when you... when you don't know that you're allowing all these people to access it [information about you], is the problem."*

Control and transparency are usually implemented through forms of notice and consent [2, 64], by which users are informed about the modalities of data access and have a choice on whether to agree to it. Often, however, such a choice is limited, as companies are under legal obligation to store data about user interactions for a fixed amount of time, hence reducing the level of perceived control for users over their online data.

With respect to anonymization, P13 stated that:

P13: *"That [level of comfort] depends on whether or not it [the information] is linked to me. If it's... if the person is anonymous... uhm... if... they could figure out who I am, it's something that I'm uneasy about."*

Similarly, P7 and P12 raised concerns about the possibility of linking their online information with their identity, in which case the comfort level would drop from 5 to 1 and from 5 to 2, respectively.

Finally, four interviewees mentioned trust as a driving factor for their comfort level. In particular, P12 argued that:

P12: *"... I actually care for what company is running it [the main service], and when I sign up for that, I actually place a bit of trust in the company, and when a third-party is using it, I am not exactly sure who these third-parties are or maybe I just don't trust them as much as I would..*

According to Unsworth [61], the challenges in managing trust in cloud environments are mainly due to the uncertainty about

[2] Cochran's Q test results across information types are as follows: Search Engine ($89.6 < \text{Chi-sq} < 258.32$, df = 4, adjusted $p < .001$ with Bonferroni correction), Online Social Network ($54 < \text{Chi-sq} < 182.3$, df = 4, adjusted $p < .001$), Online Shopping ($123.6 < \text{Chi-sq} < 290.2$, df = 4, adjusted $p < .001$).

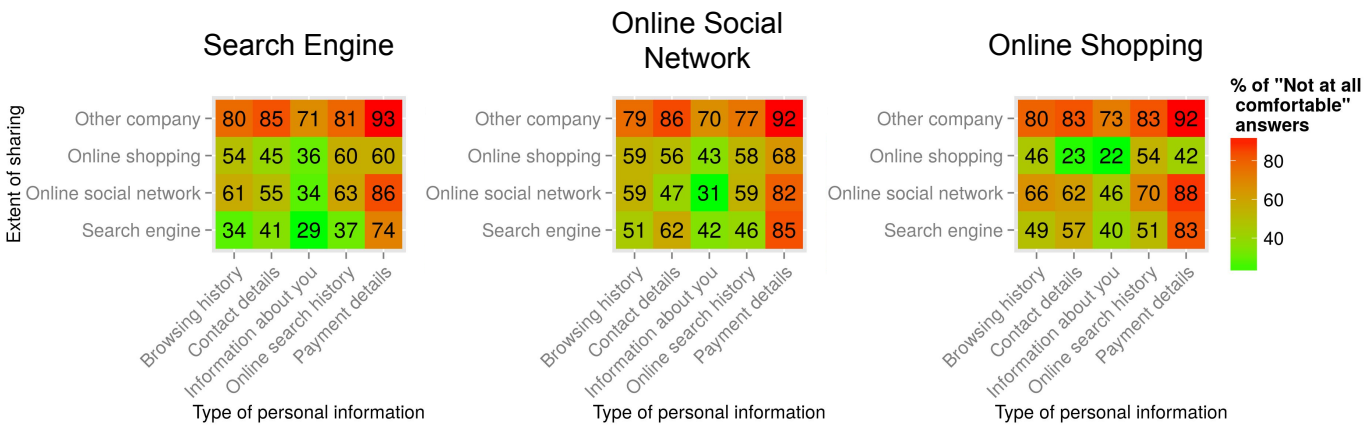

Figure 2. Discomfort with sharing different types of information, for each scenario (left - Search Engine, middle - Online Social Network, right - Online Shopping). The values are rounded for enhanced readability.

and lack of knowledge of third-party data policies, two dimensions of transparency. As users often have limited control over cloud resources and about the data that can be accessed by a third-party, they need to rely on contracts stipulating what use can be performed over users' data, and on the possibility of compensations for the affected users in the event of inappropriate and unauthorized data leakage [47].

In the following, we present findings related to comfort with sharing different types of data within each of the three considered scenarios.

Search Engine Scenario

In the search engine scenario, the participants are overall the least comfortable when their information is accessed by a third-party company: 82% chose the "Not at all comfortable" answers (on average across all 5 information types), followed by the secondary service "Online social network" belonging to the first-party company (60%), the shopping service (51%) and finally by the search engine (43%, the main service). Figure 2 (left) shows the breakdown per information type for this scenario. [3] With respect to the different information types, 78% of the respondents were (on average) not at all comfortable with sharing payment details, 60% with sharing their online search history, 57% with their online browsing history, 56% with their contact details and 43% with information about them (such as age, gender, interests), which is also the type of information that exhibits the least amount of variability across all secondary services (max − min = 7.8%). A post-hoc analysis [4] indicates that the difference between the comfort level of the information type "payment details" and each other type is significant, across all sharing extents (adjusted $p < .001$), except from "Browsing history" and "Online search history" when accessed by the shopping service.

Unsurprisingly for this scenario, we see that the survey respondents are quite comfortable if the search engine accessed their search history, as it is required to provide them with relevant results. However, they seem to be quite comfortable also for other types of information, except for payment details. In the follow-up interviews, the participants in this scenario stated that they could see the value in receiving personalized ads and services based on their browsing history. However, they were quite sensitive when it came to having such information publicly accessible and associated with them, especially by close social circles such as family and friends, as stated by P9:

P9: *"So... if I'm gonna like... look at some porn, I wouldn't want it to... like... immediately coming up on my shopping site for things that I wanna buy... I'll be like... oh no, I'm shopping with my mom and that's really awkward... but... if it's being used in a more subtle way that that I don't mind."*

Online Social Networks Scenario

As shown in Figure 2 (middle), respondents to the survey in the online social network scenario are in general less comfortable with sharing any kind of personal information with the main service (social network), as compared to the previous scenario (search engine). In fact, the difference in the number of responses with the lowest comfort score between these two scenarios is 12.6% (on average, for the same type of information), with the largest single difference occurring for the online search history, where there are 59% more users that are not at all comfortable with the primary service accessing such information in the Online Social Network scenario as compared to the Search Engine scenario, even if they have an account on it. One concern that emerged from the interviews is related to the scope of the search history that could potentially reach the social circles of the users, if used for advertising and personalization. For instance, P2 stated that:

P2: *"... I'd be a little less comfortable with the social network... providing information to... ah... large groups of people... you know... or my grandma."*

[3] Note that the number in each cell represents the percentage of users who replied "Not at all comfortable" for a given information type (column) being accessed by a given entity (row). Hence, each of the cells corresponds to a question/answer item in the questionnaire.

[4] A post-hoc series of McNemar's tests conducted using the Bonferroni correction for all pairwise comparisons can be found in in the Supplementary Information document, together with the full questionnaire.

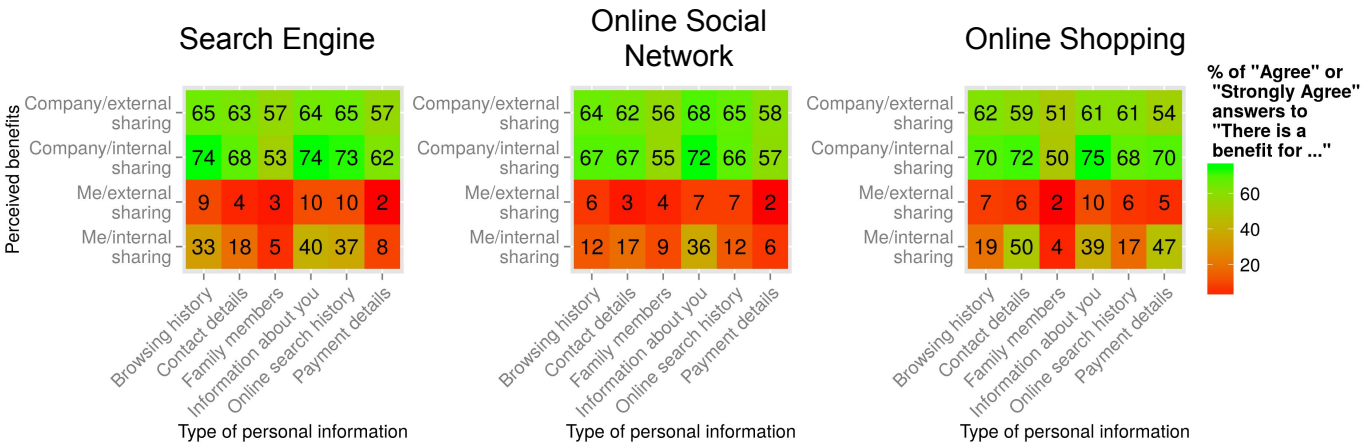

Figure 3. Perceived benefits for either users (i.e., Me/) or the first-party company (i.e., Company/) when personal information is used only internally (i.e., /no sharing) or shared with a third-party company (i.e., /sharing), for each scenario (left - search engine, middle - online social network, right - online shopping). The values are rounded for enhanced readability

For the other types of information, participants reported having the same ranking in terms of comfort as they have in the search engine scenario.

Online Shopping Scenario

In this scenario, Figure 2 (right) shows that, overall, only a minority (42%) of the users are not at all comfortable if the main service (online shopping) accesses their payment details, and such a ratio drops even further (22%) for the contact details, as most of the time these are required for processing and shipping online purchases. For the remaining information type, the respondents follow the same trend as in the other two scenarios.

It is interesting to notice that even though payment details are usually processed by the shopping service in order to provide the service or good to the user, there is still a non-negligible fraction of the respondents (42%) who are not at all comfortable with that. During our interviews, participants pointed to the risks of identity and financial theft (P4, P14), as experienced by one of our interviewees. In particular, P14 stated:

P14: *"I do not pay unless I can use [popular online payment system A or B]. I had... personal situations where my account information was used... without me knowing. I just don't want that to happen again... it's just too easy."*

In other scenarios, participants raised the issue of explicit consent for payment details being accessed and shared even internally within a company. As payment details can be linked to financial transactions, a notoriously sensitive topic for online users, it is unsurprising to see that users would prefer to exert more control over the flow of such information, in order to increase their comfort.

Benefits from Sharing Data

After asking participants about their comfort level with knowing that different services and companies would access and share their personal data, we now focus on the second dimension of users' attitudes towards online data combination and sharing, i.e., the benefits that there could be for users if their data gets shared. To capture their opinions, we ask them about the extent to which they agree that there are benefits for them if (i) different types of data (the same ones as before) are shared by the first-party company with its secondary services, and if (ii) different types of data are shared with a third-party company (questions #12 – #17 of the questionnaire).

Figure 3 shows the ratio of responses "Agree" or "Strongly agree", on a 5-point Likert scale from "Strongly disagree" to "Strongly agree", to 4 statements about the 4 possible combinations of beneficiaries and sharing; for example, one such statement in the search engine scenario is "There is a benefit for me if the Search Engine has access to... [6 different types of information]". We included "family members", in addition to the previous 5 information types, as online services may rely on social features to provide value for the users.

In general, the figure shows that users clearly distinguish between the presence of benefits for them as opposed to the first-party company (i.e., top vs. bottom two rows of the figure). The majority of respondents agree that there is a benefit for the company when it either uses or shares users' data with its own secondary services or with another company; yet, only a minority of users feels the same about the benefits for themselves, if the first-party company uses and shares their data with another company. A Cochran's Q test shows that, within each scenario, the difference across beneficiaries is statistically significant. [5] During our interviews, however, participants struggled to find a real-world example in which a company would clearly benefit from such an activity and a user would not. For instance, P12 stated:

P12: *"I know that a lot of company already do this, so it must be beneficial... otherwise they wouldn't be doing so."*

On average, 66% of the survey respondents agree that there is a benefit for the first-party company if it accesses users' personal information, and 61% agree that there is a benefit for

[5] Cochran's Q test results across information types are as follows: Search Engine ($55.5 < \text{Chi-sq} < 252.4$, df = 5, adjusted $p < .001$ with Bonferroni correction), Online Social Network ($20.9 < \text{Chi-sq} < 129.5$, df = 5, adjusted $p < .01$), Online Shopping ($42 < \text{Chi-sq} < 416.1$, df = 5, adjusted $p < .001$).

the same company if it shares users' data with a third-party company However, only 23% of the respondents agree that there is a benefit for them if the company uses their information, and only 6% agree to the same if the company shares their information with another company. For example, in the interviews candidates suggested unprompted that a better purchasing experience, such as 1-click purchases, and the convenience of not having to manually enter the contact details are two notable benefits that there are for them, if the first-party company made use of their personal information. If sharing such information with a third-party company, however, participants could perceive some benefits for them in terms of a larger and complementary choice of services/products that the first-party company does not offer. The convenience factor seems to arise from the survey responses of the online shopping scenario (Figure 3 - right), where we notice a highly diverse percentage of respondents who agree that there is a benefit for them if such data does not get shared with a third-party company, for different types of personal information. For example, about half of the respondents agree or strongly agree with that statement for their contact and payment details. Our results provide further evidence of the existence of a relationship between perceived benefits and comfort with allowing access to data, which has been observed in a recent study [58]. Although important, the quantification of such a relationship is outside of the scope of this paper, and further research is needed to evaluate its strength.

Risks of Re-identification

The goal of the last part of our survey is to understand how users perceive the risk of re-identification from 10 different types of data. [6] First, we asked users to rank the different types of data in terms of the re-identification risk that they pose (question #18 of the questionnaire), from the most to the least personally identifying type of data. Second, we asked users to indicate, on a scale from 1 to 5 where 1 means "Not at all identifying" and 5 means "Extremely identifying", the extent to which they feel that 4 different combinations of the 10 individual data types can be used to uniquely identify them (question #19).

Individual Data Types

As shown in Figure 4, the "GPS trace over the last week" is perceived as being the most personally identifying type of data (ranked 1st or 2nd in 74% of the responses), followed by their email address (60%) and ZIP code (21%). At the end of the ranking, we find "marital status", "gender" and "interests", aspects related to demographics and personality. Dimensions pertaining to the online behavior ("browsing history" and "online purchases") are present in the middle. A Friedman rank sum test indicates that the differences in the rankings across the different data types are statistically significant. [7]

[6]The data types we considered are: Trace of the GPS position over the last week, email address, ZIP code, browsing history, list of all items purchased online, education (schools you went to), age, interests, gender, and marital status.

[7]Friedman Chi-sq = 2451, df = 9, $p < .001$.

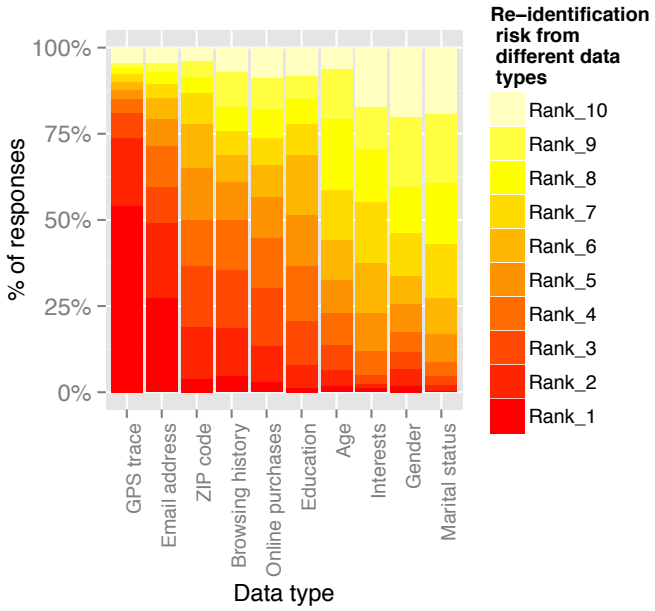

Figure 4. Re-identification risk from different data types, ranked from Rank_1 (most personally identifying) to Rank_10 (least personally identifying). The types are ordered from left to right according to their average ranking, from the highest to the lowest, respectively.

In the interviews, participants expressed concerns about their physical safety when explaining the risk of re-identification from location information. In particular, P2 stated:

P2: *"GPS position over the last week, that clearly shows my patterns... where I live, where I work, where my kids go to school... it's not that you know where I am, but if you start drawing that pattern... that draws a very good picture of physically where I am. If real-time, if I get close to a store with the GPS, send my location... but tracking me, storing that information is where... you know... I start having problems with that."*

In addition to physical safety and tracking concerns, P3 mentioned concerns about inferring the personality and larger social group of people to whom she belongs to:

P3: *"I live in an area where region strongly defines identity... [with GPS] you can identify the individual but also identify what kind of person they are likely to be, based on the region that they're in."*

Moreover, P8 pointed out that a GPS trace can be used to infer some of the other types of information without having to access them directly:

P8: *"I feel like... for GPS position, that could easily define a lot about what our interests are... who we go to see... it could easily... give you a range... you could easily figure out ZIP code from house you get to visit every night, possibly the gender based on where shopping occurs, interests from where different things happen, where education occurs... "*

Location data, which is often used by online services and mobile applications in order to provide context-related services, has received a significant attention by both media [46, 59]

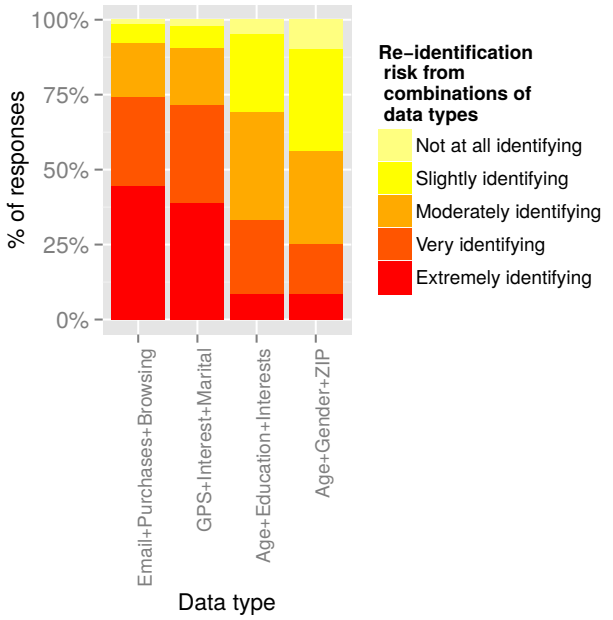

Figure 5. Re-identification risk from different combinations of data types. The combinations are ordered from left to right according to their average ranking, from the highest to the lowest, respectively.

and the research community [10, 27, 45, 51], due to its early adoption and popularity among smartphone users. Research scholars have shown how location data, even if discretized and anonymized, can be used to de-anonymize users successfully 95% of the time [15].

Two aspects related to users' online behavior, i.e., browsing history and list of online purchases, are ranked at the 4th and 5th place, respectively. Interviewees noted that such data can be collected and used passively to infer what the users like, without them knowing about it. For instance, P7, who ranked browsing history as being the most personally identifying type of data, stated that:

P7: *"That [browsing history] is the most ... like ... quantitative and qualitative data, it's a lot of detail as well as breadth... just knowing one day's worth of browsing history is a lot."*

In the literature, Olejnik et al. [44] have analysed the uniqueness of users' browsing histories, showing that 69% of the users in their experiment had a unique browsing history, and that only 4 websites are enough to uniquely identify 97% of the users. In addition to browsing history, Soroush et al. [55] have shown how throughput data related to music streaming on a mobile device can leak one's location trajectory, and similarly Michalevsky et al. [38] have relied on the power consumption of a mobile device to track users' whereabouts.

Combinations of Different Data Types

In addition to the evaluating the perceived risk of re-identification from individual data types, we asked our respondents a similar question related to four different *combinations* of individual data types. Specifically, we wanted to study how such a risk is related to the other types of data it is combined with. We asked respondents to rate on a scale from 1 to 5, where 1 means "Not at all identifying" and 5 means "Extremely identifying", different combinations of data types.

Figure 5 shows the opinions of the respondents for this question. Differences across the combinations are statistically significant. [8] An interesting finding is that, as opposed to the individual case, the combination containing GPS information is not perceived as being the most personally-identifying one. [9] When combined, data related to online activities is perceived by the users as being (on average) the most personally identifying, followed by offline data containing GPS, users' interests, marital status, age, gender and ZIP code. When reasoning about these two combinations during the interviews, the tension was evident:

P5: *"My email address is exactly the same as my legal given name... purchasing and browsing history would tell you quite a bit more. Interest and GPS and marital status are...slightly more anonymous, although because you're throwing in GPS...you probably know where the person lives..."*

43% of the interviewees alluded to the concept of k-anonymity while explaining their thought process for rating the 4 combinations of data types. Although research has shown that such a concept might provide only limited location privacy [52, 66], especially when combined with other publicly-available information (such as the census data), participants often resorted to this concept in order to rationalize their choice.

DISCUSSION

Our study investigated online users' opinions and concerns about data combination that happens across different Internet services and companies. In particular, we (1) analysed users' comfort with knowing that a service or company might access and share some of their data with other services or third-party companies, (2) evaluated the perceived benefits for users and companies from the users' perspective, and (3) studied users' the perceived risk of re-identification for different data types.

The results indicate that the level of comfort with sharing data depends on the data type, the domain in which the first-party company operates and whether users have a direct relationship with the third-party that could be accessing their data. Users are the least comfortable when their financial information is accessed or shared, followed by their online behavior (browsing and search history) and finally by contact details demographic information (age, gender, interests). Although such ranking holds across the three scenarios we considered (search engine, online social network and online shopping service), there are notable differences in relative values within each one of them. In order to increase users' comfort, according to the opinions that emerged during our interviews, first-party companies should adopt more comprehensive commu-

[8] Friedman Chi-sq = 898.2, df = 3, $p < .001$.

[9] Post-hoc Wilcoxon signed rank tests with Bonferroni correction show that the differences across all pairs are statistically significant, with adjusted $p \in [10^{-82}, .049]$.

nication strategies based on a greater transparency (i.e., what and how data is used or shared), provide more control over the data access to users (e.g., through intuitive settings and an opt-in approach) and clarify the extent of data anonymization before it is being shared. According to our participants, the interaction of these three factors would positively affect their comfort when data is accessed and shared. Moreover, the users' trust in the third-party company is especially relevant when such a company accesses users' data, which can be managed through clear and understandable data policies indicating the purpose and scope of the data access, as well as the security mechanisms in place to safeguard data and the liability clauses in case of unintended data leakage.

In terms of perceived benefits, users clearly differentiate between benefits for them versus benefits for the company, where the former are significantly lower (23%) with respect to the latter (66%). Moreover, users feel that there is no significant difference for them in terms of perceived benefits, irrespectively of whether the first-party company shares their data with a third-party company (61%) or not (66%). However, our interview participants often struggle to describe clear examples of benefits the first-party company may get from accessing and sharing users' data. In order to mitigate this perceived imbalance, one possible approach for the companies could be to put forward a better communication strategy, by exemplifying the benefits of accessing and sharing data with both first- and third-party services. As mentioned by our participants, this could include a better user experience through 1-click purchases, more complementary services or goods not offered by the first-party company, and higher relevance of search results by accessing the browsing history.

Unintended consequences due to data leakage can be detrimental for both users and companies. The perceived risk of re-identification of different types of data can be analysed from two perspectives. First and foremost, users often mention concerns related to the physical safety when explaining what they perceive as being the most personally-identifying piece of information: The user's GPS trace over the last week. However, concerns are not limited to physical safety. Personality traits, such as interests, preferences and habits, are almost as important, indicating that behavior (including online behavior) is perceived as having a moderate potential to re-identify a person (as highlighted in [21]). Furthermore, our results show that combinations of moderate-risk data types related to online behavior (such as browsing history, email address and list of online purchases) are perceived as bearing more risk than the combination of offline (or physical and demographics) data, such as the GPS trace, interests, age and gender. According to several of our interviewees, one plausible explanation for this significant difference is that online data can potentially reveal personality traits of an individual which provide a greater amount of detail about one's intents their evolution over time. As compared to the physical location, for instance, online behavior could be used to reveal a user's current state and future plans, which could be inferred to a greater extent from her/his browsing history as compared to the regular location and travel patterns.

Some of the limitations of this study are as follows. First, the participants were recruited either directly on MTurk or through an internal participants panel. In both cases, participants needed to opt-in to participate to human-subject studies, which could introduce selection bias in the sample that was recruited. Second, in order to limit participants' fatigue when responding to the questionnaire, we omitted low-sensitivity questions that could have been useful to calibrate the responses to the high-sensitivity questions that were presented. Finally, our participants were asked questions about a hypothetical scenario, which demanded a non-negligible effort to remember; future work should address these issues in a more realistic setting, for instance by focusing more about users' actual experiences rather than imagined situations.

CONCLUSION

The use of online services often requires users to create accounts and to provide some information about them to the service providers. Even though users are aware that such information is stored and processed by the service they register with, they might not be fully aware of the possibility that their data could flow and be aggregated across different services of that company, or even with a third-party company. In order to better understand users' perceptions and feelings about data sharing across online services and companies, we conducted an online survey and follow-up interviews to quantify users' comfort level with sharing different types of data with different types of services, the perceived benefits and the risks of re-identification for different combinations of their data.

Based on 918 survey respondents and 14 interviews, our results indicate that users' comfort level is influenced by three contextual factors, which are the type of service, the type of data, and the existence (or not) of a direct relationship with a third-party company. More polarised, users feel that there is a significant imbalance in the perceived benefits from data access for them, irrespectively of the context, as compared to the company. In terms of re-identification potential of different types of data, our results show that location information is the single piece of information that is perceived as being the most personally identifiable; however, that is not true when data is combined. Inadvertently leaking information pertaining to online behavior (such as browsing history and online purchases) has a similar if not higher risk for the users as leaking other types of data that include location.

Our findings suggest that, in order to address these challenges, companies should actively improve their communication to users (by providing them with more transparency about their data practices), provide a more concrete value proposition for the users, and greater controls over third-party access to data about them.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to Sunny Consolvo, Allison Woodruff, Sebastian Schnorf, Miguel Malheiros, Pauline Anthonysamy, as well as to the anonymous reviewers for the precious feedback that greatly improved the quality of this work.

REFERENCES

1. Alessandro Acquisti. 2006. Price Discrimination, Privacy Technologies, and User Acceptance. *CHI Workshop on Personalization and Privacy* (2006).
2. A. Acquisti, I. Adjerid, and L. Brandimarte. 2013. Gone in 15 Seconds: The Limits of Privacy Transparency and Control. *IEEE Security Privacy* 11, 4 (July 2013), 72–74.
3. Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. 2015. Privacy and human behavior in the age of information. *Science* 347, 6221 (2015), 509–514.
4. Anne Adams and Martina Angela Sasse. 2001. Privacy in multimedia communications: Protecting users, not just data. In *People and Computers XVInteraction without Frontiers*. Springer, 49–64.
5. Lalit Agarwal, Nisheeth Shrivastava, Sharad Jaiswal, and Saurabh Panjwani. 2013. Do not embarrass: re-examining user concerns for online tracking and advertising. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*. ACM, 8.
6. Naveen Farag Awad and M. S. Krishnan. 2006. The personalization privacy paradox: an empirical evaluation of information transparency and the willingness to be profiled online for personalization. *MIS quarterly* (2006), 13–28.
7. Michael Barbaro, Tom Zeller, and Saul Hansell. 2006. A face is exposed for AOL searcher no. 4417749. *New York Times* 9, 2008 (2006).
8. Louise Barkhuus. 2012. The mismeasurement of privacy: using contextual integrity to reconsider privacy in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 367–376.
9. Alastair R Beresford, Dorothea Kübler, and Sören Preibusch. 2012. Unwillingness to pay for privacy: A field experiment. *Economics Letters* 117, 1 (2012), 25–27.
10. Alastair R Beresford and Frank Stajano. 2003. Location privacy in pervasive computing. *IEEE Pervasive computing* 1 (2003), 46–55.
11. Igor Bilogrevic, Kévin Huguenin, Berker Agir, Murtuza Jadliwala, Maria Gazaki, and Jean-Pierre Hubaux. 2015a. A machine-learning based approach to privacy-aware information-sharing in mobile social networks. *Pervasive and Mobile Computing* (2015).
12. Igor Bilogrevic, Kévin Huguenin, Stefan Mihaila, Reza Shokri, and Jean-Pierre Hubaux. 2015b. Predicting Users' Motivations behind Location Check-Ins and Utility Implications of Privacy Protection Mechanisms. In *22nd Network and Distributed System Security Symposium (NDSS" 15)*.
13. Joseph Bonneau and Sren Preibusch. 2010. The privacy jungle: On the market for data protection in social networks. In *Economics of information security and privacy*. Springer, 121–167.
14. Ramnath K. Chellappa and Raymond G. Sin. 2005. Personalization versus privacy: An empirical examination of the online consumers dilemma. *Information Technology and Management* 6, 2-3 (2005), 181–202.
15. Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. 2013. Unique in the Crowd: The privacy bounds of human mobility. *Scientific reports* 3 (2013).
16. Cynthia Dwork. 2011. Differential privacy. In *Encyclopedia of Cryptography and Security*. Springer, 338–340.
17. Facebook. 2015. Updating Our Terms and Policies: Helping You Understand How Facebook Works and How to Control Your Information. (2015).
18. Adrienne Porter Felt, Serge Egelman, and David Wagner. 2012. I've got 99 problems, but vibration ain't one: a survey of smartphone users' concerns. In *Proceedings of the second ACM workshop on Security and privacy in smartphones and mobile devices*. ACM, 33–44.
19. Oded Goldreich. 1998. Secure multi-party computation. *Manuscript. Preliminary version* (1998).
20. Google. 2012. Updating our privacy policies and terms of service. (2012).
21. Serge Gutwirth, Ronald Leenes, Paul de Hert, and Yves Poullet. 2012. *European data protection: coming of age*. Springer Science & Business Media.
22. Mark Halper. 2015. Isabelle Falque-Pierrotin: Privacy Needs to Be the Default, Not an Option. (June 2015).
23. Mathias Humbert, Erman Ayday, Jean-Pierre Hubaux, and Amalio Telenti. 2013. Addressing the concerns of the lacks family: quantification of kin genomic privacy. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security (CCS '13)*. ACM, New York, NY, USA, 1141–1152.
24. Mathias Humbert, Erman Ayday, Jean-Pierre Hubaux, and Amalio Telenti. 2014. Reconciling Utility with Privacy in Genomics. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society (WPES '14)*. ACM, New York, NY, USA, 11–20.
25. Carlos Jensen and Colin Potts. 2004. Privacy Policies As Decision-making Tools: An Evaluation of Online Privacy Notices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 471–478.
26. Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 1301–1318.

27. John Krumm. 2009. A survey of computational location privacy. *Personal and Ubiquitous Computing* 13, 6 (2009), 391–399.
28. Ponnurangam Kumaraguru and Lorrie Cranor. 2006. Privacy in India: Attitudes and awareness. In *Privacy Enhancing Technologies*. Springer, 243–258.
29. David Lazer, Alex (Sandy) Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Life in the network: the coming age of computational social science. *Science* 323, 5915 (Feb. 2009), 721–723.
30. Pedro Giovanni Leon, Blase Ur, Yang Wang, Manya Sleeper, Rebecca Balebako, Richard Shay, Lujo Bauer, Mihai Christodorescu, and Lorrie Faith Cranor. 2013. What matters to users?: factors that affect users' willingness to share information with online advertisers. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*. ACM, 7.
31. Linkedin. 2015. Updating LinkedIns Privacy Policy. (2015).
32. Jen-Yu Liu and Yi-Hsuan Yang. 2012. Inferring Personal Traits from Music Listening History. In *Proceedings of the Second International ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies (MIRUM '12)*. ACM, New York, NY, USA, 31–36.
33. Mary Madden. 2014. Americans Consider Certain Kinds of Data to be More Sensitive than Others. (2014).
34. Miguel Malheiros, Sren Preibusch, and M. Angela Sasse. 2013. "Fairly truthful": The impact of perceived effort, fairness, relevance, and sensitivity on personal data disclosure. In *Trust and Trustworthy Computing*. Springer, 250–266.
35. Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazons Mechanical Turk. *Behavior research methods* 44, 1 (2012), 1–23.
36. Aleecia M Mcdonald, Robert W Reeder, Patrick Gage Kelley, and Lorrie Faith Cranor. 2009. A comparative study of online privacy policies and formats. In *Privacy enhancing technologies*. Springer, 37–55.
37. William J. McGuire. 1974. Psychological motives and communication gratification. *The uses of mass communications: Current perspectives on gratifications research* 3 (1974), 167–196.
38. Yan Michalevsky, Gabi Nakibly, Aaron Schulman, and Dan Boneh. 2015. PowerSpy: Location Tracking using Mobile Device Power Analysis. *USENIX Security* (2015).
39. Kate Moody. 2012. Internet usage amongst British consumers. (2012).
40. James Moor. 1989. How to invade and protect privacy with computers. *The information web: Ethical and social implications of computer networking* (1989), 57–70.
41. Anthony Morton. 2015. *Age shall not wither them: But It Will Change Their Priorities About Protecting Their Information Privacy*. Working Paper. Academy of Science and Engineering,USA.
42. Arvind Narayanan and Vitaly Shmatikov. 2009. De-anonymizing social networks. In *Security and Privacy, 2009 30th IEEE Symposium on*. IEEE, 173–187.
43. Helen Nissenbaum. 2004. Privacy as contextual integrity. *Washington law review* 79, 1 (2004).
44. Lukasz Olejnik, Claude Castelluccia, and Artur Janc. 2012. Why johnny can't browse in peace: On the uniqueness of web browsing history patterns. In *5th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs 2012)*.
45. Alexandra-Mihaela Olteanu, Kévin Huguenin, Reza Shokri, and Jean-Pierre Hubaux. 2014. Quantifying the effect of co-location information on location privacy. In *Privacy Enhancing Technologies*. Springer, 184–203.
46. Matthew Panzarino. 2011. Its not just the iPhone, Android stores your location data too. (Aug. 2011).
47. S. Pearson and A. Benameur. 2010. Privacy, Security and Trust Issues Arising from Cloud Computing. In *2010 IEEE Second International Conference on Cloud Computing Technology and Science (CloudCom)*. 693–702.
48. Vasile Claudiu Perta, Marco Valerio Barbera, and Alessandro Mei. 2014. Exploiting Delay Patterns for User IPs Identification in Cellular Networks. In *Privacy Enhancing Technologies*. Springer, 224–243.
49. Alan M. Rubin. 1985. Uses and gratifications: Quasi-functional analysis. *Broadcasting research methods* (1985), 202–220.
50. Sebastian Schnorf, Aaron Sedley, M Ortlieb, and A Woodruff. 2014. A comparison of six sample providers regarding online privacy benchmarks. In *SOUPS Workshop on Privacy Personas and Segmentation*.
51. Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. 2011. Quantifying location privacy. In *Security and Privacy (SP), 2011 IEEE Symposium on*. IEEE, 247–262.
52. Reza Shokri, Carmela Troncoso, Claudia Diaz, Julien Freudiger, and Jean-Pierre Hubaux. 2010. Unraveling an old cloak: k-anonymity for location privacy. In *Proceedings of the 9th annual ACM workshop on Privacy in the electronic society*. ACM, 115–118.
53. Internet Society. 2012. Global Internet User Survey. (2012).

54. Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-Laszlo Barabasi. 2010. Limits of predictability in human mobility. *Science* 327, 5968 (2010), 1018–1021.

55. Hamed Soroush, Keen Sung, Erik Learned-Miller, Brian Neil Levine, and Marc Liberatore. 2013. Turning Off GPS is Not Enough: Cellular location leaks over the Internet. In *Privacy Enhancing Technologies*. Springer, 103–122.

56. Juliana Sutanto, Elia Palme, Chuan-Hoo Tan, and Chee Wei Phang. 2013. Addressing the personalization-privacy paradox: an empirical assessment from a field experiment on smartphone users. *Mis Quarterly* 37, 4 (2013), 1141–1164.

57. Latanya Sweeney. 2000. Simple demographics often identify people uniquely. *Health (San Francisco)* 671 (2000), 1–34.

58. Teradata. 2015. Balancing the Personalisation and Privacy Equation The Consumer View. (2015).

59. New York Times. 2014. How your iPhone is tracking your every move. (Aug. 2014).

60. Eran Toch, Yang Wang, and Lorrie Faith Cranor. 2012. Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems. *User Modeling and User-Adapted Interaction* 22, 1-2 (2012), 203–220.

61. Kristene Unsworth. 2014. Questioning trust in the era of big (and small) data. *Bul. Am. Soc. Info. Sci. Tech.* 41, 1 (Oct. 2014), 12–14.

62. Yang Wang, Gregory Norice, and Lorrie Faith Cranor. 2011. Who is concerned about what? A study of American, Chinese and Indian users privacy concerns on social network sites. In *Trust and trustworthy computing*. Springer, 146–153.

63. Samuel D Warren and Louis D Brandeis. 1890. The right to privacy. *Harvard law review* (1890), 193–220.

64. Edgar A. Whitley. 2009. Informational privacy, consent and the control of personal data. *Information security technical report* 14, 3 (2009), 154–159.

65. Sergey Yekhanin. 2010. Private Information Retrieval. *Commun. ACM* 53, 4 (April 2010), 68–73.

66. Hui Zang and Jean Bolot. 2011. Anonymization of location data does not work: A large-scale measurement study. In *Proceedings of the 17th annual international conference on Mobile computing and networking*. ACM, 145–156.