Fakultät für Informatik
Technische Universität München

Assessing the Impact of Data Practices on User Privacy

Ashwini Rao

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:
Prof. Dr. Claudia Eckert

Prüfende der Dissertation:
1. Prof. Dr. Jürgen Pfeffer
2. Prof. Dr. Simon Hegelich

Die Dissertation wurde am 26.04.2018 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 08.10.2018 angenommen.

Abstract

Interactions with online, mobile and Internet-of-Things (IoT) technologies generate intimate data about our lives and how we think, feel and behave. Products and services can collect, share and combine large amount of user data including sensitive data such as personal health, income and religion data. To understand user privacy, we need to identify data practices of products and services and assess their impact on data privacy. In this thesis, we use qualitative and quantitative methods to identify data practices of websites, trackers, aggregators and mobile apps, and assess their impact on data privacy. We identify website data practices by analyzing website privacy policies and assess their impact on user privacy expectations. We propose and validate a conceptual model for privacy expectation with four types of privacy expectations: Desired, Predicted, Deserved and Minimum. We identify mismatched privacy expectations by comparing the privacy expectations elicited from users with the website data practices extracted from privacy policies. We propose a conceptual model for different types of mismatches and discuss how they can impact user privacy differently. We use network analysis to identify tracker data practices and study their impact on linking of user data across multiple contexts and siphoning of user data. We analyze user behavioral profiles to identify aggregator data practices. We study user concerns regarding data in their behavioral profiles, estimate the extent of errors in user behavioral profiles, evaluate the extent of transparency provided by profile access mechanisms, and identify usability issues of profile access mechanisms. Lastly, we identify data practices of mobile apps and study how they enable integration of user data from online, mobile and IoT contexts.

Zusammenfassung

Interaktionen mit Online-Services, mobilen Endgeräten und Internet of Things-Technologien erzeugen intime Informationen über unser tägliches Leben, unsere Gedanken und unser Verhalten. Produkte und Services können große Mengen solcher sensiblen Nutzerinformationen sammeln und weitergeben – darunter auch Informationen über Gesundheit, Einkommen und religiöse Einstellungen. Um die Privatsphäre von Nutzern zu vermessen, müssen Praktiken im Umgang mit Nutzerinformationen solcher Produkte und Services identifiziert und ihre Auswirkungen auf die Privatheit von Daten greifbar gemacht werden. In dieser Arbeit werden qualitative und quantitative Methoden eingesetzt, um übliche Praktiken von Webseiten, "Trackern", Aggregatoren und mobilen Anwendungen zu identifizieren und ihre jeweiligen Auswirkungen auf die Privatheit von Daten abzuschätzen.

Der Umgang von Webseiten mit Nutzerinformationen wird durch die systematische Analyse von Datenschutzbestimmungen analysiert und an den Erwartungen an die Privatheit von Nutzerdaten gemessen. Dabei wird ein theoretisches Modell für die Erwartungen an die Privatheit von Nutzerdaten vorgeschlagen und evaluiert. Dieses Modell besteht im Kern aus vier Arten von Erwartungen: den erwünschten, den prognostizierten, den verdienten und den minimalen Erwartungen. Anhand einer Umfrage unter Nutzern ausgewählter Webseiten wird evaluiert, inwiefern die Erwartungen an die Privatheit der eigenen Daten durch die in den Datenschutzbestimmungen definierten Praktiken erfüllt werden.

Auf der Basis dieser Analysen werden die identifizierten Arten verschiedener verfehlter Erwartungen kategorisiert und in einem Konzept systematisiert. Dieses Konzept erlaubt erstmals eine fundierte Diskussion über die Diskrepanz zwischen Erwartungen an die Privatheit eigener Daten und übliche Praktiken im Umgang mit diesen Daten.

Daran schließt eine Analyse von sogenannten Website-Trackern an, die Nutzer über mehrere Webseiten hinweg identifizieren und nachverfolgen können. Mithilfe von Methoden der Netzwerkanalyse wird untersucht, welche Tracker Nutzer über welche Webseiten hinweg verfolgen und so deren Nutzerverhalten über mehrere Kontexte hinweg verlinken können.

Weiterhin werden Verhaltensprofile analysiert, um zu verstehen, wie Aggregatoren mit Nutzerdaten umgehen. Dazu werden auch Bedenken von Nutzern in Bezug auf die Daten mit einbezogen, die in solchen Profilen enthalten ist. Das erlaubt erstmals systematische Erkenntnisse, inwiefern solche Verhaltensprofile von der Realität abweichen,

in welchem Maße Transparenz über die in den Profilen enthaltenen Daten geschaffen wird und ob Nutzern der Einblick in die eigenen Daten gewährt wird.

Schließlich werden mobile Anwendungen auf ihren Umgang mit Nutzerdaten untersucht. Das erlaubt neue Einblicke, wie diese Anwendungen die Integration von Nutzerdaten von Online-Services, mobilen Endgeräten und Internet of Things-Technologien ermöglichen.

Acknowledgments

A great person said that it takes a village to raise a child. Although not a child, my doctoral journey has been personally a great undertaking, and I could not have completed it without help from many individuals. I thank my advisor Prof. Dr. Jürgen Pfeffer for his guidance. I thank my committee members and colleagues for their help. I am grateful to my coauthors because they taught me a lot about research. I am indebted to my friends who supported me when I needed it the most. I thank my husband Birendra Jha who was patient and encouraging throughout my doctoral studies; I could not have asked for a more loving and caring partner.

Contents

List of Figures

List of Tables

1 Introduction

In an increasingly technological world, user interactions with online, mobile and Internet-of-Things technologies are becoming an inherent and unavoidable part of everyday life. Our interactions with these technologies are gradually increasing in number and frequency. We use them when working, eating, playing, entertaining and even while sleeping. They have revolutionized and improved the way we live – services such as email, social networking and search have enabled us to communicate, socialize and access information in novel ways. However, interactions with these technologies can generate intimate data about our lives and how we think, feel and behave. They can enable entities, private and government, to collect, share and combine large amount of user data including sensitive data such as personal health, income and religion data. Hence they can impact our data privacy.

Informational or data privacy is related to data practices such as collection, use, sharing and retention of users' data by products and services. There exist other conceptualizations of privacy such as privacy as a right to intimate decisions about one's body or sexuality [1]. If products and services collect data regarding a person's sexuality or intimate decisions, then such data practices fall under the scope of informational privacy. In a technological world, user identity and user data are entwined, and understanding user privacy is incomplete without understanding informational or data privacy. Hence, to understand user privacy, it is important to identify data practices of products and services and assess their impact on data privacy.

Identifying data practices of products and services can be challenging. For example, data practices may be described in privacy notices that are long, verbose and time consuming to read and understand [2, 3, 4]. When descriptions of data practices are unclear in privacy notices or altogether missing, we need alternative means of identifying data practices. It is also challenging to assess the impact of data practices on user data privacy. For example, data practices that do not match user privacy expectations may violate user privacy and cause privacy concerns [5]. However, we need to develop practical approaches for understanding whether data practices meet user privacy expectations [6]. The focus of this thesis is on identifying data practices of products and services and studying how they impact user data privacy.

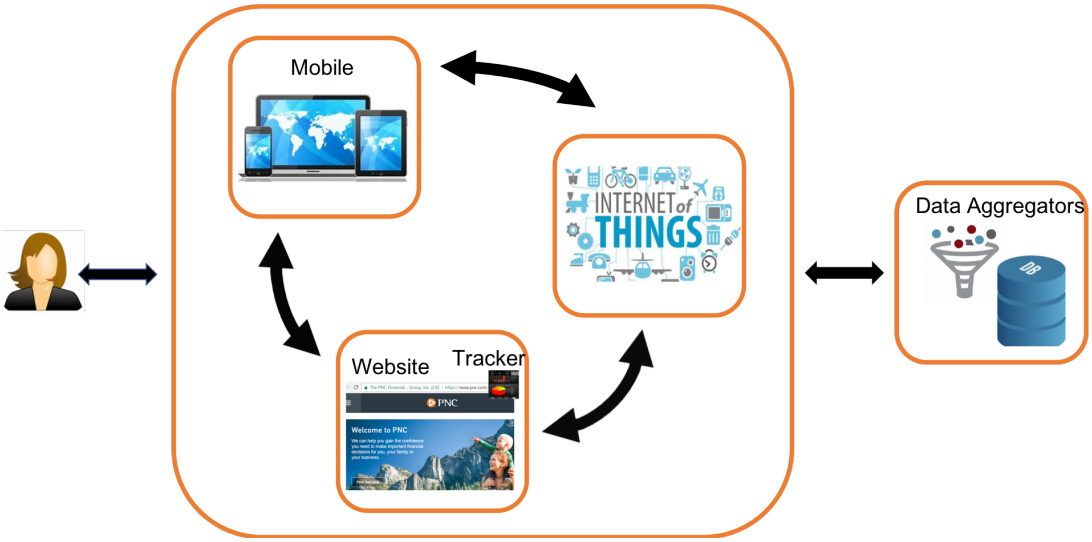


Figure 1.1: User interactions with Internet, Mobile and IoT entities.

1.1 Research Problem

Figure 1.1 shows a simplified data ecosystem in which different entities collect and process user data. The figure shows entities from Internet (e.g. websites and trackers), Mobile (e.g. apps and tablets), Internet-of-Things (e.g. smart bikes and smart watches) and Data Aggregators. Entities can collect data directly via user interactions, or they can collect data indirectly from other entities. For example, generally users directly interact with entities such as websites, mobile apps and smart bikes, but not with entities such as trackers and aggregators. Trackers are small pieces of code that companies can embed within a websites and they can collect information about users' activities on the websites [7, 8]. Currently there are more than 4400 trackers in the wild [9]. Data aggregators are companies that collect and combine user data from public and private sector service providers, and trade them on data marketplaces [5].

Research Questions In this thesis, we focus on data practices of four entities: websites, trackers, aggregators and mobile apps. We use qualitative and quantitative methods (interviews, surveys, content analysis, network analysis, and parametric and non-parametric statistics) to answer two research questions:

R1: How can we identify data practices of products and services?
R2: How can we measure the impact of data practices on user privacy?

We apply research questions **R1** and **R2** to four common scenarios involving websites, trackers, aggregators and mobile apps entities. We order the scenarios based on increasing order of complexity of interactions between users and the entities.

S1: Users interact with websites of products and services, and websites collect data directly from users. This is the simplest scenario.

S2: Trackers are embedded on websites, and they collect data from users indirectly while users interact with websites.

S3: Data aggregators collect user data indirectly from websites and trackers.

S4: Users directly interact with mobile apps that provide access to a smart bikes. Users also interact with smart bikes while using them. Mobile apps and smart bikes collect data directly from users. Both embed trackers that indirectly get data from users. Mobile apps, smart bikes and trackers share user data with data aggregators. This is the most complex scenario.

1.2 Prior Art and Contributions

Below we discuss the main contributions of this thesis. In Table 1.1, we group them into contributions to privacy theory and contributions to privacy methodology. We discuss how our contributions differ from prior research. Table 1.2 shows the organization of the thesis and maps the contributions to chapters in the thesis.

C1: Types of Privacy Expectations and Mismatches We propose a conceptual model that treats privacy expectation as a multi-level construct. The model proposes four "types" of privacy expectations: Desired, Predicted, Deserved and Minimum. Further, it proposes that the types represent distinct levels of user privacy, and, hence, there can be an "order" among the types.

We propose a conceptual model for different types of mismatches. We examine the types of mismatches that can occur due to a single expectation type as well as the types of mismatches that can result from interaction of multiple expectation types. We discuss how different types of mismatches can impact user privacy differently.

To the best of our knowledge, privacy research has not focused on the potential for multiple types of privacy expectations and mismatches. Empirical studies that measure privacy expectations [10, 11, 12, 13, 14, 15, 16] have largely considered "privacy expectation" as a single construct. while empirically eliciting privacy expectations, studies have either focused on privacy expectations in the desired sense [13] or have not clarified the meaning of privacy expectation [10, 14, 15, 16]. Theoretical work on conceptual definition of privacy expectation [17, 10, 18, 19] has considered privacy expectation as a single construct. However, they consider privacy expectation as more than just desires.

Table 1.1: Main contributions of this thesis.

Contribution	Type	Description
C1	Theory	Types of Privacy Expectations and Mismatches
C2	Method	Measure Privacy Expectations and Mismatches
C3	Method	Identify Tracker Data Practices Using Network Analysis
C4	Method	Identify Aggregator Data Practices Using Behavioral Profile Analysis

C2: Measure Privacy Expectations and Mismatches We discuss the design and implementation of an empirical study for measuring different types of privacy expectations. The study tests the validity of the proposed conceptual model for privacy expectation. Empirical evidence from the study supports the conceptual model and the hypothesis that there are several types of privacy expectations and that there can be an ordering among them. Using the study design we can operationalize measuring different types of privacy expectations.

We propose a practical approach for identifying mismatches between user privacy expectations and website data practices [6]. We extract website data practices by analyzing website privacy policies. We elicit privacy expectations from users and compare them to extracted data practices. We study if users predict whether a website will collect, share or delete data. In contrast to prior work, we propose an approach that facilitates direct comparison of individuals' expectations of what a website's data practices are to the website's actual claims of what they do as stated in their privacy policy. While studying mismatched expectations, prior work has either implicitly captured expectations in the sense of desired preferences [13] or not clarified the type of expectation [14, 15, 16]. Earp et al. studied Internet users' privacy values and analyzed privacy policies for respective statements [14]. They find that Internet users' concerns and values are not adequately reflected in privacy policies. Gomez et al. also compared websites' data practices with practices users find concerning [15]. Milne and Bahl examined differences between consumers' and marketers' expectations regarding use of eight information technologies [13]. Liu et al. measured disparity between expected and actual Facebook privacy settings. In contrast to our study on website data practices, Lin et al. studied expectations regarding data practices of mobile apps [10]. Further, their work did not differentiate between different types of expectations; while eliciting expectations, they did not clarify the type of expectation being elicited.

C3: Identify Tracker Data Practices Using Network Analysis We use network analysis to identify tracker data practices. We focus on identifying data practices that enable linking user activities across different website categories. We study the impact of tracker data practices on linking of user data and siphoning of user data. We use the term "siphon" to indicate a one-way channel that once set up will result in a continuous flow of personal information from the source to the destination. We study whether and how current tracking mechanisms can be

Table 1.2: Outline mapping chapters to study scenarios and thesis contributions.

Title	Scenario	Contributions
Chapter 1: Introduction		
Chapter 2: Impact of Website Data Practices	**S1**	**C1**, **C2**
Chapter 3: Impact of Tracker Data Practices	**S2**	**C3**
Chapter 4: Impact of Data Aggregator Practices	**S3**	**C4**
Chapter 5: Impact of Mobile App Data Practices	**S4**	
Chapter 6: Discussion and Conclusion		

used to siphon data from one country to another. Prior research has investigated the prevalence of tracking on the Internet [7, 20], user awareness and concerns regarding tracking [21, 15, 5], technologies used for tracking (cookies, flash cookies, fingerprinting etc.) [22, 23, 24] and defenses against tracking[8, 25].

C4: Identify Aggregator Data Practices Using Behavioral Profile Analysis We propose a novel approach for identifying aggregator data practices [5]. We identify the types of data that aggregators collect about users by examining behavioral profiles of users and user data sold on data marketplaces. The United States Federal Trade Commission has investigated the types of data that companies may potentially use to build behavioral profiles. However, they did not look at contents of actual behavioral profiles [26].

Using behavioral profiles analysis, we study the impact of aggregator data practices [5]. We study user concerns regarding data in their behavioral profiles, estimate the extent of errors in user behavioral profiles, evaluate the extent of transparency provided by profile access mechanisms, and identify usability issues of profile access mechanisms. Prior studies have focused on user concerns and perceptions regarding use of behavioral profiles for advertising [27, 28]. We focus on user privacy concerns regarding actual contents of behavioral profiles. Our approach of using user's own behavioral profile for eliciting concerns and surprises leads to a more contextualized and nuanced understanding of user concerns regarding online behavioral profiles.

1.3 Outline

We organize the rest of the thesis as follows. In Table 1.2, we map how the chapters in the thesis map to the four scenarios that we study and to the contributions listed in Table 1.1. At the end of each chapter, we provide a brief summary of the contents of the chapter. In Chapter 2, we use privacy policy analysis to identify website data practices, and study their impact on user privacy expectations. We propose and validate a conceptual model for privacy expectation types, examine types of mismatches,

and measure mismatched privacy expectations. Parts of this chapter were previously published by USENIX [6]. In Chapter 3, we discuss the impact of tracker data practices. We use network analysis to identify tracker data practices and examine how they impact linking and siphoning of user data. In Chapter 4, we discuss the impact of aggregator data practices. We use behavioral profile analysis to identify aggregator data practices and explain how they impact privacy concerns, data quality, data transparency and profile usability issues. Parts of this chapter were previously published by ASE [5]. In Chapter 5, we discuss the impact of mobile app data practices. We identify mobile app practices and examine how they impact user privacy. In Chapter 6, we discuss implications of our work for public policy and regulations, development of privacy enhancing technologies and privacy research. We conclude in Chapter 6.

2 Privacy Impact of Website Data Practices

Internet users interact with websites to access products and services. Websites generally inform Internet users about websites' data practices, such as collection, sharing and retention of personal information, via a website privacy policy [29]. Website privacy policies, written in natural language, can be long, time consuming to read [2, 3], and difficult to understand for users [30, 4]. Therefore, users often do not read the policies [31, 32]. Since users do not read privacy policies, their expectations regarding data practices of websites may not match websites' data practices. Expectations influence decision making [33] and mismatches between users' expectations and website data practices may lead to incorrect privacy-related decisions. Users may expose themselves to unanticipated privacy risks e.g. using a health website that shares users' health data with insurance companies.

In order to improve consumer data privacy, regulatory agencies have sought to understand user privacy expectations [34, 35]. The European Union General Data Protection Regulation (GDPR), which comes into effect in 2018, emphasizes "taking into consideration the reasonable expectations of data subjects" and requires that companies do "[. . .] careful assessment whether a data subject can reasonably expect at the time and in the context of the collection of the personal data that processing for that purpose may take place." [34] To comply with GDPR regulations, companies have to understand how user privacy expectations vary based on website data practices. They have to elicit and measure privacy expectations and identify whether expectations match actual practices. Enforcement of GDPR will depend on the ability to understand expectations and identify mismatches in expectations.

One approach to help users understand website data practices is to provide more concise privacy notices in addition to privacy policies [36]. Such privacy notices may be based on privacy policies, but are generally shorter and more usable. Although unexpected data practices may be described in a privacy policy, they are likely to be overlooked among descriptions of a large number of data practices that may or may not be relevant for the user's current transactional context. In order to make data practices transparent to users, privacy policies should be complemented with short form notices tailored to the user's transactional context [36] and should warn users about unexpected practices [37]. The challenge, however, lies in identifying unexpected practices.

Parts of this chapter were previously published by USENIX [6].

We analyze the impact of website data practices on user privacy expectations [6]. We examine how expectations vary based on data practices. We propose a novel practical approach for identifying unexpected website data practices; we elicit user privacy expectations and compare them with website data practices extracted from website privacy policies. In order to accurately elicit privacy expectations, we propose and validate a conceptual model for privacy expectation. Our model defines four "types" of privacy expectations: Desired, Predicted, Deserved and Minimum. Further, the types represent distinct levels of user privacy, and, hence, there the model proposes that there can be an "order" among the types. To the best of our knowledge, privacy research has not focused on the potential for multiple types of privacy expectations.

2.1 Identifying Website Data Practices

Website privacy policies are the dominant mechanism through which websites disclose their data practices. Hence, to identify website data practices, we can extract and analyze data practices disclosed in website privacy policies. A website data practice can be discomposed into components such as *source*, *target*, *action*, *data*, *purpose* and *consent*. A source performs an action, involving certain data, on the target for a certain purpose with or without the consent of the target. Table 2.1 shows 17 website data practices. The source can be a user using a website. The target can be the operator of the website. Action component includes collection, sharing, deletion of personal information etc. Actions related to surreptitious collection, unauthorized disclosure and wrongful retention of personal information are more concerning to users than other actions [29]. Examples of the data component include *contact information* (e.g. email or postal address), *financial information* (e.g. bank account information, credit card details, or credit history), *health information* (e.g. medical history or health insurance information) and *current location* (e.g. from where a user is accessing the website). Generally health and financial data are considered to be privacy-sensitive. Purpose of the action could be classified as *core purpose* or *other purpose*. A core purpose is performing the action such as collection for core services that a website provides e.g. a museum website sells entry tickets for the museum. Purposes not directly relevant for providing core services can be classified as other purposes e.g. a museum website may share data with advertisers. A website may ask explicitly as for users' consent before performing an action or assume that the user implicitly consents to an action.

2.1.1 Privacy Policy Analysis

To demonstrate the process of identifying website data practices, we extract and analyze data practices from the privacy policies of a set of 16 websites shown in Table 2.2. We extract 17 data practices shown in Table 2.1. The data set contains websites from three types of website categories: finance, health and dictionary. The financial category includes banking, credit card and online payment websites. The health category

Table 2.1: Types of website data practices.

Action	Scenario	Data
Collection	With account	Contact
		Financial
		Health
		Current location
	Without account	Contact
		Financial
		Health
		Current location
Sharing	For core purpose	Contact
		Financial
		Health
		Current location
	For other purpose	Contact
		Financial
		Health
		Current location
Deletion	–	Personal data

includes pharmacy, health clinic and health reference websites. Website categories were determined using Alexa website categories [38]. The websites vary *popularity*, as determined by their website traffic rankings [38]. Lastly, websites are either owned by the government or by private companies.

Extracting Data Practices

We can extract data practices manually or semi-automatically. In the manual approach, people read and annotate text from privacy policies. To establish the ground truth, manual approaches generally employ experts with knowledge of privacy practices and privacy law. When multiple annotators annotate the same policy, we can compute inter-annotator agreement statistic to understand the level of agreement among the annotators. The manual approach limits the number of policies that we an analyze. Semi-automated approaches combine machine learning and natural language processing techniques with crowdsourcing to extract data practices from privacy policies [39, 40,

Table 2.2: Sample website dataset.

Website	Type	Subtype	Context	Rank
Webmd.com	Health	Reference	Private	107
Medhelp.org	Health	Reference	Private	2,135
Medlineplus.gov	Health	Reference	Government	558,671
Walgreens.com	Health	Pharmacy	Private	315
Bartelldrugs.com	Health	Pharmacy	Private	54,737
Mayoclinic.org	Health	Clinic	Private	297
Clevelandclinic.org	Health	Clinic	Private	2,629
Americanexpress.com	Finance	Credit	Private	76
Discover.com	Finance	Credit	Private	324
Bankofamerica.com	Finance	Bank	Private	33
Woodlandbank.com	Finance	Bank	Private	915,921
Banknd.nd.gov	Finance	Bank	Government	5,267
Paypal.com	Finance	Payment	Private	21
V.me	Finance	Payment	Private	27,289
Merriam-webster.com	Dictionary	–	Private	266
Wordnik.com	Dictionary	–	Private	8,412

Rank as of 3/10/2015

41, 42, 43]. This is an active area of research and recent results [39, 40] show the possibility of achieving acceptable level of agreement with non-expert crowd workers. Semi-automated techniques can enable scaling up to a large number of websites.

Annotating Data Practices

Annotators can code website data practices in many ways. For example practices such as collection and sharing may be coded as *yes, no, unclear* or *not addressed* [30]. A *yes* indicates that the website is clear engages in collection or sharing of data, and a *no* implies that the website clearly does not engage in a in collection or sharing of data. For example, the statement "When you use our Websites, we collect your location using IP address." makes it clear that the website collects location information. A website policy may be ambiguous or not clear about collection or sharing of data. For example, the statement "We collect the IP address from which you access our Website." mentions collecting IP address but is unclear whether the website collects

location information derived from the IP address. Annotators can code such actions as *unclear*. If a website policy does not disclose information about a specific practice, annotators can use the code *not addressed*. A deletion practice may be annotated as *full deletion* (websites allows deletion of all user data), *partial deletion* (deletion of only some data), *no deletion*, *unclear* or *not addressed*.

We employed two annotators, one with legal and another with privacy expertise, to manually extract data practices shown in Table 2.1 from the privacy policies of websites shown in Table 2.2. Each annotator independently read each of the 16 privacy policies and extracted the relevant collection, sharing and deletion data practices. Agreement was generally high, for instance, among the 17 data practices, the highest inter-annotator agreement was κ=1 and lowest agreement was κ=0.718. All disagreements were resolved jointly after initial independent coding. Table 2.3 shows the annotations for the privacy policy of one of the websites (Bank of America). Annotators coded the 17 data practices as follows: 7 "Yes," 4 "No" and 7 "Unclear."

Table 2.3: Annotations for the 17 data practices of BankofAmerica.com's privacy policy.

Data practice	**Annotation**
Collect contact – with account	Yes
Collect contact – without account	Unclear
Collect financial – with account	Yes
Collect financial – without account	No
Collect health – with account	Yes
Collect health – without account	No
Collect location – with account	Unclear
Collect location – without account	Unclear
Share contact – core purpose	Unclear
Share contact – other purpose	Unclear
Share financial – core purpose	Yes
Share financial – other purpose	Yes
Share health – core purpose	Yes
Share health – other purpose	No
Share location – core purpose	Unclear
Share location – other purpose	Unclear
Deletion	No

2.1.2 Website Data Practices

Figure 2.1 shows how collection, sharing and deletion data practices vary in clarity. A website privacy policy may clearly say "Yes" or "No" about what the website does, may be unclear, or not contain any statements that address the data practice topic. Collection data practices appear to be more clear than deletion, which is more clear than sharing (58.6% vs. 50.1% vs. 47.7%).

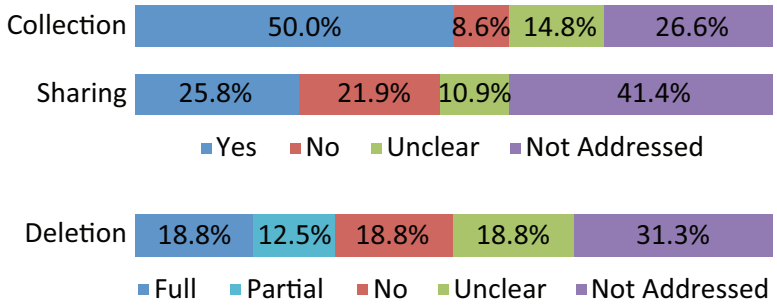


Figure 2.1: Clarity of collection, sharing and deletion data practices.

Figure 2.2 gives an overview of collection and sharing data practices among financial, health and dictionary categories. It shows the percentage of collection and sharing data practices that are clear, unclear or not addressed in the privacy policies. We find that policies in all three website categories are mostly clear about practices concerning the collection or sharing of contact information, i.e. they make explicit statements about whether they collect or not collect contact information and make clear statements about sharing (dominantly yes for core purposes; no for other purposes). The analysis shows that some data practices are common across different website categories, whereas others are category-specific or even vary within a category.

Not surprisingly, finance websites make explicit statements about collection and sharing of financial information. Note that credit card and online payment finance websites collect financial information even from non-registered users, e.g., when users buy products, but banking websites do not. About half of the health websites' privacy policies also make explicit statements concerning financial information, however, the other half is silent on whether they collect or share financial information. Interestingly, the dictionary websites make statements that leave it unclear if they may collect financial information, but are either explicit or silent on sharing of financial information. Dictionary sites mention processing payments or posting transactions but not explicit collection of financial information.

All dictionary websites and all but one of the financial websites do not address collection or sharing of health information. One of the finance websites, BankofAmerica.com is explicit about collecting health information from registered users and sharing it with third parties for core purposes. It does so via its insurance-related affiliates, which may not be obvious to users. However, all but two of the health websites are explicit

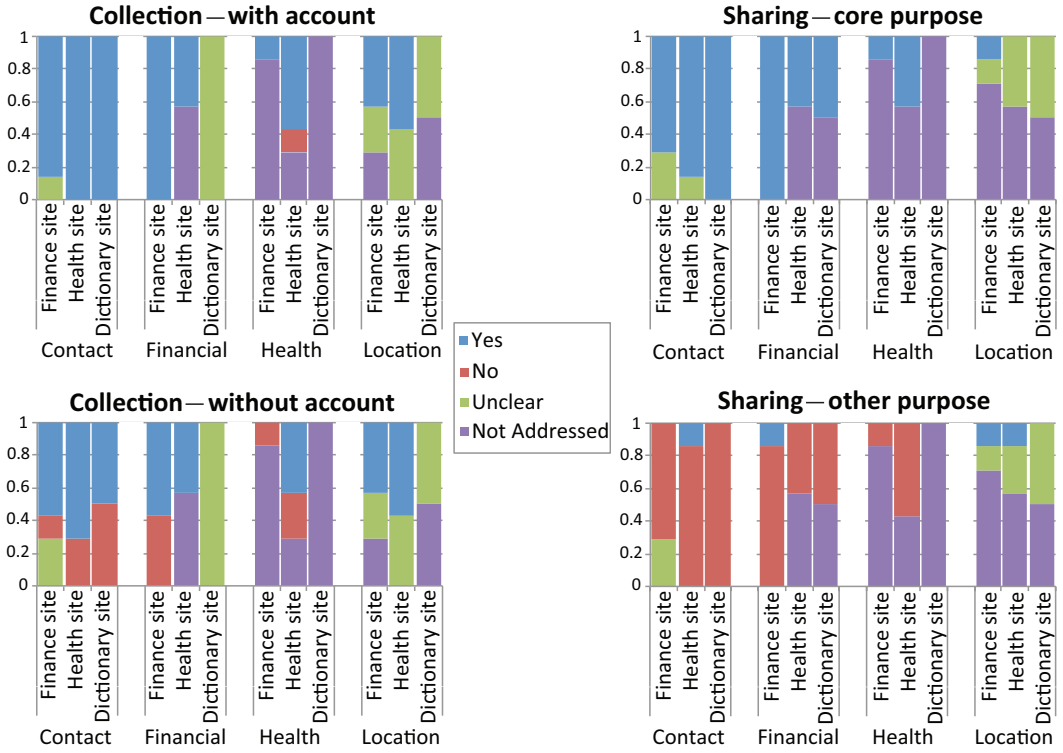


Figure 2.2: Collection and sharing data practices of websites in sample data set.

about whether they collect health information. Both health clinic websites do not address collection of health information in their website privacy policy, but contain links to additional policies, which may disclose their collection practices. Health websites are less explicit about sharing of health information compared to collection of health information.

About half of the financial and health websites are clear about collection of current location information, but none of the dictionary sites are clear on this aspect. Almost all website privacy policies are unclear or silent on whether they share location information with third parties. Only one finance website explicitly states that it shares user location for core and other purposes. Only one health website explicitly states that it shares user location for other purposes, but it is unclear whether it shares it for core purposes.

Financial websites are more explicit about deletion data practices compared to health and dictionary websites. Nearly 71% (5) of the financial websites clearly disclose their practice in contrast to 50% (1) of the dictionary websites and 28% (2) of the health websites. However, nearly half of the financial websites (3) do not allow any deletion of data and two only allow partial deletion. In contrast, when clear about the practice, health websites (2) and dictionary websites (1) allow full deletion.

2.2 Types of Privacy Expectations

We address theoretical and empirical questions related to privacy expectations of users. We focus on expectations related to informational privacy or data privacy and not on other conceptualizations of privacy such as privacy as a right to intimate decisions about one's body or sexuality. Informational or data privacy is related to collection, use, sharing, retention etc. of users' data by products and services. Products and services generally describe their data practices in a privacy policy. If products and services collect data regarding a person's sexuality or intimate decisions, then such data practices fall under the scope of informational privacy.

We do not focus on legal doctrines such as "right to privacy" or "expectations of privacy" as defined in law. Legal doctrines and laws related to privacy vary widely across the world. It is beyond the scope of our work as well as expertise to explain how they may be related to this work. The results from our work, however, could be used to understand whether privacy laws and doctrines are grounded in users' expectations related to informational privacy.

To the best of our knowledge, privacy research has not focused on the potential for multiple types of privacy expectations. In contrast to privacy domain, other domains such as Consumer Satisfaction/Dissatisfaction (CS/D) and service quality treat expectation as a multi-level construct [44, 45, 46]; CS/D literature supports four types of "consumer expectations": ideal, expected, deserved and minimum tolerable [44, 45], and service quality literature supports three types of "service expectations": desired, adequate and predicted [46].

Empirical studies that measure privacy expectations [10, 11, 12, 13, 14, 15, 16] have largely considered "privacy expectation" as a single construct. while empirically eliciting privacy expectations, studies have either focused on privacy expectations in the desired sense [13] or have not clarified the meaning of privacy expectation [10, 14, 15, 16]. Studies that elicit users' privacy preferences have been conducted in many contexts [47, 48, 49, 50]. They may be implicitly studying privacy expectations in a desired sense. Hence, in the privacy context, empirical work has focused on privacy expectations in the desired sense or preferences, or has not clarified the meaning of privacy expectation.

Theoretical work on conceptual definition of privacy expectation [17, 10, 18, 19] has considered privacy expectation as a single construct. However, they consider privacy expectation as more than just desires. Altman considers desired privacy and achieved privacy as two important aspects of privacy [17]. He describes the desired level as a subjective ideal internal state at any given moment. If the achieved level of privacy, as perceived by an individual or group, matches the desired level, then satisfaction results otherwise individual or groups are unsatisfied. Altman's work primarily focuses on physical interaction, and Palen and Dourish extend Altman's theory to a world with information technology [51]. They discuss how technology can disrupt privacy

management by violating personal desires and social expectations of the social settings in which the technology is present. In Nissenbaum's privacy as contextual integrity theory, privacy expectations are obligatory norms [19, pp. 138-139], which vary by context and govern the flow of information in terms of who, what and how [19]. Martin's privacy as social contract theory extends privacy as contextual integrity theory [18] and views privacy expectations as social contracts that are mutually beneficial, sustainable and unstated agreements within a context. In both the theories, meeting privacy expectations requires respecting obligatory norms. Lin et al. propose the concept of privacy as expectations as "people's mental models of what they think an app does and does not do" in a mobile context [10].

2.2.1 Conceptual Model

We first discuss the observations from which we induce our conceptual model. We then present our conceptual model and discuss its relationship to existing privacy theories. We summarize our conceptual model in Table 2.4.

Table 2.4: Summary of conceptual model for privacy expectation types.

Type	Keywords	Description	Critical[a]	Privacy Level
Desired	ideal, want	"what people ideally want to happen"		Highest
Predicted	think, will	"what people think will happen"	Knowledge	
Deserved	feel, should	"what people feel should or ought to happen"	Investment	
Minimum	tolerate, must	"what people would tolerate if something must happen"	Essentiality	Lowest

[a] *Critical* determinant of Type, *Knowledge* of privacy practices and *Investment* in effort, time, money etc.

Observations

Our first observation comes from an in-person interview we conducted to study user privacy expectations. The participant in the interview was asked, "Do you expect the website to ask for your consent for sharing your information?" In response, the participant replied "I think the expect question is a little hard to answer because I am thinking whether you are asking me what I think should be done or what I perceive how they are doing it now." While thinking about privacy expectations, the participant differentiated between expectation of how the world should be and expectation of how the world actually is. Further, the participant suggested that "I think it is helpful if you make it clear otherwise you will get different answers and you don't know what they are answering to because some people might answer the question what they think it should be done, some people might answer the question as what it is, some people might not distinguish those [...]" The observation from the interview suggests that people have different types of privacy expectations and simply asking what they expect can lead to different interpretations.

Our second observation comes from review of Consumer Satisfaction and Dissatisfaction (CS/D) literature. In the CS/D domain, consumer expectations are considered "an influence on, if not determinant of, levels of satisfaction or dissatisfaction." [44] Models of satisfaction consider two determinants of satisfaction: expected performance of a product and evaluation of its perceived actual performance [45]. If the perceived actual performance is greater or equal to the expected performance consumers are satisfied otherwise dissatisfied. Miller extended this basic model of satisfaction to include "types" of expectations consumers might use as comparison standards for performance evaluation [45]. As per Gilly et al., "Miller contends that simply asking the consumer what he or she 'expects' can result in different interpretations by different consumers." [44]

Miller conceptually recognized four types of consumer expectation types: Ideal, Expected, Deserved and Minimum Tolerable [45]. The Ideal represents "wished for" level that reflects what consumers feel performance "can be." The Expected reflects what consumers feel performance "will be." It represents an objective calculation of probabilities and does not have an affective dimension. The Deserved has an affective dimension and represents what consumers feel performance "should be." Lastly, the Minimum Tolerable represents what consumers feel the lowest performance "must be." It is a "better than nothing" option. Miller suggested an ordering among the types with Ideal at the highest level and Minimum Tolerable at the lowest. He contended that the Deserved would be higher than the Expected if consumer investment in terms of time, effort, money etc. is high. Empirical work by Gilly et al. found partial support for the types and ordering among them [44].

We note the similarity between Altman's privacy theory and Miller's conceptual model of satisfaction. Both involve a comparison process where the perceived level is compared to the "expected" level. In Altman's privacy theory, people evaluate perceived actual privacy against their desired level of privacy, and they are satisfied when there is a match. Altman associates only one level, the desired level, with the expected level. Miller, however, considers four different levels, including the desired level, for the expected level.

Miller contends that consumer expectations probably vary among consumers based on experiences, demographics, psychographics etc. and within a consumer temporally based on recent experience, situation etc. [45] Similarly, privacy expectations may vary among people based on demographic characteristics, privacy concern, privacy knowledge, geographic location etc. and within a person based on context, recent experience etc. Expectations in general influence decision making [33]. We contend that, similar to consumer expectation types used as comparison standards to evaluate performance, people use privacy expectation types as comparison standards to evaluate privacy. Given the conceptual similarities between consumer expectations and privacy expectations, we build our conceptual model for privacy expectation types on the work in consumer expectation types in the CS/D domain.

Privacy Expectation Types

Based on our qualitative observations, we propose a conceptual model for privacy expectation with multiple types. Our model, inspired from Miller's model of consumer expectation types, consists of four privacy expectation types: Desired, Predicted, Deserved and Minimum. These are the types of expectations that people have about privacy, and we distinguish them from actual data privacy practices. We discuss below the four types and their relationship to existing privacy theories.

Our conceptual model builds on an important element common to the privacy theories that we discussed: **standard of evaluation**. In each of the theories, there is a single standard against which evaluation is carried out and action is taken when the evaluation fails to meet the standard. We propose that, instead of a single standard, there are multiple standards of evaluation both among and within individuals. When asked "What do you expect[...]," in a scenario, they can use any of these standards to evaluate what they expect in the scenario. Use of multiple standards of evaluation results in multiple "types" of privacy expectations.

1. *Desired Type*: It is what people ideally want to happen. It is similar to the desired level of privacy used as the standard of evaluation in Altman's privacy theory [17]. The desired level of privacy as per Altman is an ideal internal state at any moment.

2. *Predicted Type*: It is what people think will happen. Here "will" indicates a definite future action or likely prediction. The Predicted type is similar to privacy as expectations concept proposed by Lin et al. because their standard of evaluation is what people think a mobile app does or does not do [10]. Accurately predicting website data practices, may require knowledge of privacy practices. For example, a user who understands how IP address works may have different expectation about collection of location information than a user who does not. Hence, it is possible that privacy knowledge will impact the Predicted type more than the other three types.

3. *Deserved Type*: Compared to the other types, the Deserved type has an affective dimension that focuses on feelings. We consider that it is critically determined by evaluation of "investment and rewards" in a scenario. Therefore, it is what people feel should or ought to happen given their investment. People can feel that they deserve a reward if their investment is high in terms of effort, time, loyalty etc. They may feel that they do not deserve a reward if their investment is low. They may even feel that they deserve a penalty if their investment is low. For example, if people paid for a website service, for a considerably long time, they may feel that they deserve a bonus or promotion i.e. reward. On the contrary, they may feel that they deserve to view unwanted advertisements, a penalty, if they did not pay for a website service. The penalty may be a decrease in privacy [52] or may not be a decrease in privacy [18]. We purposely use "investment and

reward/penalty" and not "cost and benefit" to emphasize the affective dimension of the Deserved type.

The Deserved type is somewhat related to the standards of evaluation in Nissenbaum's privacy as contextual integrity theory and Martin's privacy as social contract theory. In Nissenbaum's theory, the standard of evaluation is based on context-relative informational norms [19]. Nissenbaum considers norms that are obligatory [19, pp. 138-139]. She attributes four key elements to norms the first two being "[...] (a) a prescriptive "ought element"; (b) a norm subject upon whom the obligation expressed in the norm falls [...]." Martin considers that individuals make decisions about sharing and use with obligations in mind [18]. Because the Deserved type focuses on what people feel ought to happen, it is obligatory in the sense considered by Nissenbaum and Martin. Martin considers that people use a rule-utilitarian approach that analyzes costs and benefits to develop norms [53], but the cost need not be a decrease in privacy. This is similar to analysis of investment and reward/penalty for evaluating the Deserved type.

4. *Minimum Type*: It is what people would tolerate if something must happen; something is essential to fulfill a need and there is not much choice. Here "must" indicates a stronger obligation than "should" or "ought." The Minimum type is critically determined by a lack of options from which people can choose based on desires or investment-reward analysis. For example, people may not generally tolerate collection of health information on a job website, but they may do so if it is required to apply for a specific job. The Minimum type is not strongly related to any standard of evaluation in existing privacy theories. Hence, it is a contribution to privacy theory.

Ordering of Privacy Expectation Types

We hypothesize that there would be an ordering among the types; different types of privacy expectations represent different levels of user privacy. If people were to assign a score to each type, there would be an ordering among the scores. Given that the Desired type is the most ideal type, it would have the highest score. In contrast, the Minimum type is something that is just tolerated, and, hence, would have the lowest score. We hypothesize that the scores for Predicted and Deserved would be between the scores for Desired and Minimum. The Deserved score could be higher than the Predicted score if "investment" is high. Otherwise its score would be lower than the Predicted score.

2.2.2 Validation

We validate our conceptual model using an empirical study. Evidence from the study supports the conceptual model and the hypothesis that there are several types of privacy expectations and that there can be an ordering among them.

2.2.2.1 Method

We designed the empirical study to test the following hypotheses:

- Are there statistically significant differences among Desired, Predicted, Deserved and Minimum privacy expectation types?
- Is there a statistically significant difference between the orderings $Desired > Predicted > Deserved > Minimum$ and $Desired > Deserved > Predicted > Minimum$?
- Is the impact of knowledge significantly more on the Predicted type? Is the impact of investment significantly more on the Deserved type?

Sample and Procedure We conducted our study in August 2017 with an initial sample consisting of 1437 adults (18+ years) selected from a United States online survey panel [54]. The sample consists of US adults with access to the Internet and is age and gender balanced as per US census. It is a stratified random sample, which reduces self-selection bias. A total of 1249 participants completed the survey (completion rate 86.91%). The final sample consisting of participants who completed the survey (N=1249) is representative of the US online population with a margin of error of $\pm$2.8%.

Participants selected from the online survey panel were invited to take a self-administered questionnaire that elicited their privacy expectations regarding a realistic privacy-sensitive scenario involving collection of health-related browsing activity by a bank. We received informed consent at the beginning of the survey. Participants completed the survey to donate $.50 to their preferred charity and enter a sweepstakes to win a $100 gift card (odds of winning 1/60,000). On average, panelists can take two surveys per week [54]. There were no repeat participants in our survey.

Variables The independent variable in our study is the privacy expectation type with four levels: Desired, Predicted, Deserved and Minimum. The dependent variable is the participant's privacy expectation rating for a given expectation type. Participants expressed their privacy expectations by rating their level of agreement or disagreement for four statements corresponding to four expectation types. We used a repeated-measures design where each participant rated all expectation types. We manipulated

the independent variable by varying the description of four statements. To reduce order effects, we reversed the expectation questions for half the participants.

Study Scenario The study elicited users' data privacy expectations regarding a privacy-sensitive scenario. The scenario described a data practice found in the banking context. Similar to prior empirical studies on privacy expectations [10, 12, 11], we decomposed the data practice into five components – action, data, source, target and purpose – necessary for eliciting privacy expectations.

One of the main goals was to test whether type of privacy expectation significantly impacted user privacy expectations. Hence, we chose an extreme case scenario where significant impact was unlikely. An extreme case scenario is more interesting than an average case scenario because an impact in the extreme case suggests a larger impact in the average case. The extreme case scenario involved health and banking contexts. People consider these contexts as privacy-sensitive at least in the US. Given the sensitive nature of health information, people probably would not desire banks to collect health information. Further, people may not predict that banks collect health information. Hence, impact based on the type of privacy expectation (Desired, Predicted and Minimum) was unlikely.

While expressing their privacy expectations, participants had to make privacy decisions regarding a realistic scenario that they could encounter in real life. By using a realistic scenario, we tried to elicit privacy expectations that are meaningful. The scenario elicited participants' expectations regarding the data practice of banks collecting health-related browsing activities of its users. The scenario described five components of the data practice. The action was collection of data. The data was health-related browsing activity, which was defined as browsing activities on websites such as WebMD, MedlinePlus or MedicineNet that people might use/visit to find information on health conditions, symptoms or treatments. The source of the data was the participant, and the target was a bank. Privacy policies of top banks in the US disclose collection of users' browsing activities from third-party websites [55, 56]. Banks collect user information through mechanisms such as website trackers. Data collected includes IP addresses, which can be mapped to users' full name, postal address and mobile number [5]. For example, the three health-information websites (webmd.com, medlineplus.gov and medicinenet.com) mentioned in the scenario description contain third-party trackers which enable banks to collect health-related browsing activities of users. Ghostery (ghostery.com), a browser tool for identifying trackers, shows trackers on WebMD (71), MedlinePlus (4) and MedicineNet (16) websites. Lastly, the purpose of data collection was specified as "to identify financial needs and provide relevant service." To ensure realism, it was based on a data practice disclosed in the privacy policy of a prominent bank in the United States [56].

Questionnaire Design We discuss our questionnaire design decisions. The full survey questionnaire is in Appendix A. As per survey best practices, the survey wording was iteratively improved based on feedback from cognitive interviews [57] and pilot surveys. During the cognitive interviews (N=2), we asked the participants to the express in their own words what they understood from each question. We wanted to ensure that the wording conveyed what we wanted to measure. At the end of the first pilot survey (N=130) and the second pilot survey (N=60), we asked the participants if they had difficulty answering any question, and if yes, what about the question made it difficult to answer. Participants were also given the option of suggesting improvements to the questions. Overall feedback suggested that the target group was able to understand and differentiate among the four privacy expectation types, and answer the survey questions.

We did not use attention check questions as per advice from recent research on survey methodology [58, 59, 60]. Such questions may increase Social Desirability Bias [60], which is an important issue for surveys related to privacy. Discarding responses based on attention check questions can introduce demographic bias related to gender, age and education [58, 59, 60], which can impact nationally representative surveys. Lastly, our pilot results indicated that the median completion time was ∼3min. Due to the short duration of our survey, decline in attention due to satisficing behavior is reduced.

Survey Introduction
To reduce the impact of demand characteristics, we informed the participants that the purpose of the survey was to understand their opinions regarding websites. Asking for opinions also reduces the threat of knowledge questions and decreases guessing [61]. We did not mention "privacy" to avoid priming effects. To reduce social desirability bias, participants were assured that their answers were anonymous, and that we did not collect any personally identifiable information including IP address.

In the survey instructions, we told the participants that their answers were important to us, and they could take their time reading and answering the questions. To reduce guessing, we told them that they should answer the questions as accurately as possible, but it was OK to say "Don't know."

Pre-Questionnaire
To provide context regarding our study scenario, we asked participants about their usage of health-information and banking websites. To reduce order effects, the two blocks related to health (2 questions) and banking (3 questions) were shown in random order. Questions were worded to reduce social desirability bias e.g. "Some people use [...] Other people do not[...]" Answer options included "Don't know/ Not sure" and "Decline to answer."

The health block had two closed-ended questions. First, we asked whether participants had used websites such as WebMD, MedlinePlus or MedicineNet to find information on health conditions, symptoms or treatments. Second, we asked them to think about

their last visit and tell us whether they recalled the information they were trying to find.

The banking block had three closed-ended questions. First, we asked if participants had used websites to check their Checking/Savings account balance. Second, we asked if they currently had a Checking/Savings account. Lastly, we asked the approximate year in which they opened their account. In our analysis, we used the number of years since they opened their account as an indication of their "investment" that could impact the Deserved type.

Main Questionnaire
We instructed the participants to imagine a scenario where they were a customer of a bank, and they had a Checking/ Savings account with the bank. Each participant rated four Likert-type items one each for Desired, Predicted, Deserved and Minimum privacy expectation types. Ratings were used as the dependent variable to analyze the impact of the independent variable, privacy expectation type.

For the rating task, participants were instructed "In this scenario, tell us how much you **agree or disagree** with the statements below. Use a scale from 0 to 10, with 0 indicating strongly disagree and 10 indicating strongly agree." To distinguish neutral from undecided, we included a "Don't know/ Not sure" option at the end of the rating scale. Providing a "Don't know" option reduces guessing [61], which is important for accurately measuring expectations that may depend on respondents' knowledge e.g. the Predicted type. Because of the privacy-sensitive scenario, we provided a "Decline to answer" option after the "Don't know" option.

Our rating scale is similar to the scales used in empirical studies that measure and compare user privacy expectations for multiple items [11, 12]. Measuring the level of agreement allows us to use a single scale, which is required to compare ratings across multiple expectation-related items. With an 11-point scale we can measure finer differences among four expectation types; it allows participants to distinguish among four items with a probability ($\sim$54%) greater than chance (50%). Scales with fewer points, 9-point ($\sim$46%), 7-point ($\sim$35%) and 5-point ($\sim$19%), have a probability of less than chance. Likert-type item data measured on a 11-point scale is closer to interval level of scaling [62], which can be used with more powerful statistical tests.

We empirically elicited four types of privacy expectations similar to the way Gilly et al. empirically elicited four types of consumer expectations [44]. Participants rated how much they **agree or disagree** with four statements. The statements used keywords identified in the conceptual model (Table 2.4) to capture the impact of four privacy expectation types: "**want**" for Desired, "**think...will**" for Predicted, "**deserve**" for Deserved and "**tolerate...must**" for Minimum. The four statements were as follows:

1. "I **want** my bank to collect my health-related browsing activity to identify my financial needs and provide service relevant to me."

2. "I **think** my bank **will** collect my health-related browsing activity to identify my financial needs and provide service relevant to me."

3. "I **deserve** that my bank collect my health-related browsing activity to identify my financial needs and provide service relevant to me."

4. "I would **tolerate** if my bank **must** collect my health-related browsing activity to identify my financial needs and provide service relevant to me."

Stating both sides of the attitude scale (**agree or disagree**) and use of a repeated-measures design reduced the impact of acquiescence bias. Each statement was a positive affirmative statement without double-negatives. This ensured that higher scores corresponded to higher level of agreement to collection of data. We stated the purpose of collection as "to identify my financial needs and provide service relevant to me." As discussed earlier, to make the rating task more realistic, the purpose-related text was based on data practices disclosed in the privacy policy of a prominent US bank [56]. Instructions for the rating task included a definition for "health-related browsing activity."

Post Questionnaire
We asked demographic questions at the end of the survey. We received gender (male, female), age range (18-29, 30-44, 45-59, 60+), household income and US location information for each panelist from the survey panel. Hence, we asked only one question regarding the highest education level completed. In addition to demographic questions, we asked two open-ended questions soliciting feedback regarding difficulty in answering questions and anything else participants considered important.

2.2.2.2 Analysis

We analyzed all completed survey responses ($N = 1249$). The median time to complete the survey was 3min & 8sec. Table 2.5 presents participant demographics and shows the distribution of gender, age range, highest education level, household income and US geographical region in our nationally representative sample. For analysis, we set the level of significance $\alpha = 0.05$. We used a Bonferroni correction for pairwise comparisons e.g. $\alpha = 0.008$ for comparisons between four expectation types.

Privacy Expectation Types We analyzed the frequencies of "Know (0-10)", "Don't know/Not sure" and "Decline to answer" responses for the four expectation types (Table 2.6). The number of responses that contain "Decline to answer" is small and similar among the four expectation types (20, 1.60%; 21, 1.68%; 26, 2.08%; 20, 1.60%). However, the number of "Don't know" responses is higher for Predicted (151, 12.09%) compared to Desired (52, 4.16%), Deserved (67, 5.36%) and Minimum (49, 3.92%) expectation types. We removed "Decline to answer" responses from analysis. We compared frequencies of "Don't know" with "Know" responses for the four expectation

Table 2.5: Participant demographics ($N = 1249$).

Gender	**N**	**%**
Female	672	53.80%
Male	577	46.20%
Age Range (years)		
18 – 29	231	18.49%
30 – 44	338	27.06%
45 – 59	229	18.33%
60+	451	36.11%
Education		
Grade 1-8/ no formal school	5	0.40%
Grade 9-11/ 12 no diploma	29	2.32%
Grade 12 with diploma	128	10.25%
Some college, no degree	256	20.50%
Two-year college degree	113	9.05%
Four-year college degree	308	24.66%
Some postgraduate school	102	8.17%
Postgraduate degree	291	23.30%
Decline to answer	17	1.36%
Household Income		
$0 to $9,999	92	7.37%
$10,000 to $24,999	133	10.65%
$25,000 to $49,999	231	18.49%
$50,000 to $74,999	164	13.13%
$75,000 to $99,999	137	10.97%
$100,000 to $124,999	120	9.61%
$125,000 to $149,999	59	4.72%
$150,000 to $174,999	32	2.56%
$175,000 to $199,999	26	2.08%
$200,000 and up	61	4.88%
Decline to answer	194	15.53%
US Region		
East North Central	186	14.89%
East South Central	63	5.04%
Middle Atlantic	151	12.09%
Mountain	111	8.89%
New England	71	5.68%
Pacific	222	17.77%
South Atlantic	226	18.09%
West North Central	98	7.85%
West South Central	109	8.73%
Decline to answer	12	0.96%

types by considering the responses as dichotomous nominal values. Cochran's Q test for related observations [63] indicated a significant overall difference in the frequency of "Don't know" responses among the expectation types $Q(3) = 177.27$; $p < 0.001$. Pairwise comparisons between expectation types indicated that the frequency of "Don't know" responses was significantly different for the Predicted expectation type compared to Desired $Q(1) = 80.03$; $p < 0.001$, Deserved $Q(1) = 59.11$; $p < 0.001$ and Minimum $Q(1) = 87.19$; $p < 0.001$ types. This supports the hypothesis that knowledge impacts the Predicted type more than other types.

Table 2.6: Responses for privacy expectation types ($N = 1249$).

Type	Know (0-10)	Don't know	Decline to answer
Desired	1177 (94.24%)	52 (4.16%)	20 (1.60%)
Predicted	**1077 (86.23%)**	**151 (12.09%)**	21 (1.68%)
Deserved	1156 (92.55%)	67 (5.36%)	26 (2.08%)
Minimum	1180 (94.48%)	49 (3.92%)	20 (1.60%)

To examine participants' ratings (0-strongly disagree to 10-strongly agree) for the four expectation types, we considered responses that contain numerical values, but no "Don't know" or "Decline to answer" values ($n = 1038$). We treat the ratings, measured on a fine grained 11-point scale, as interval data. Table 2.7 lists mean, SD, minimum, quantiles and maximum of the ratings for four expectation types and their six pairwise comparisons. From Table 2.7, we see that median values for all expectation types are 0. At least 50% of the participants strongly disagreed to collection of health-related browsing activity by a bank; these participants did not desire it (Desired), predict it will happen (Predicted), feel they deserved it (Deserved), or tolerate it under any circumstances (Minimum). Ratings for Desired and Deserved are 0 even at the 75^{th} quantile. However, at the 75^{th} quantile, ratings for Predicted and Minimum are ≥ 1 indicating that participants disagree to a lesser extent. At the 90^{th} quantile, about 10% of the participants somewhat agree (≥ 5) that banks will collect health-related browsing activity data, and they tolerate such collection under some circumstances. However, even at the 90^{th} quantile, participants disagree (≥ 3) that they desire or deserve such collection.

Participant ratings of expectation types are different, but we want to confirm that they are significant. In Table 2.7, the means for the four expectation types are different: Predicted (1.48), Minimum (1.13), Deserved (0.91) and Desired (0.84). However, the distribution of ratings is not normal, and means may not accurately estimate significance. Hence, we compare rank ordering of the ratings by using nonparametric tests that treat rankings as ordinal data.

Friedman test for measuring differences between related observations [63], indicated a significant overall difference among the expectation types $F(3) = 53.4264$; $p < 0.00001$.

Pairwise comparisons of expectation types showed significant differences $p < 0.00003$ within all pairs except one: (Desired, Deserved). The pairs (Predicted, Desired), (Predicted, Deserved), (Predicted, Minimum), (Minimum, Desired) and (Minimum, Deserved) were significantly different. This result supports our hypothesis that people have different types of privacy expectations.

Table 2.7: Participant ratings of privacy expectation types (n=1038)

Type	Mean	SD	Min	Q 5%	Q 10%	Q 15%	Q 20%	Q 25%	Med	Q 75%	Q 80%	Q 85%	Q 90%	Q 95%	Max
Predicted	1.48	2.58	0	0	0	0	0	0	0	2	3	5	6	8	10
Minimum	1.13	2.36	0	0	0	0	0	0	0	1	2	3	5	7	10
Deserved	0.91	2.15	0	0	0	0	0	0	0	0	1	2	3.1	6	10
Desired	0.84	2.12	0	0	0	0	0	0	0	0	1	2	3	6	10
P-Di	0.64	2.14	-10	-1	0	0	0	0	0	0	1	2	3	5	10
P-De	0.58	2.17	-10	-1	0	0	0	0	0	0	0	1	3	5	10
P-M	0.35	2.45	-10	-3	-1	0	0	0	0	0	0.2	1	3	5	10
M-Di	0.29	1.69	-10	-1	0	0	0	0	0	0	0	1	2	3	10
M-De	0.22	1.54	-10	-1	0	0	0	0	0	0	0	0	1	2	10
De-Di	0.07	1.29	-10	-1	0	0	0	0	0	0	0	0	0	1	10

We had two hypotheses regarding ordering of expectation types: either $Desired > Deserved > Predicted > Minimum$ or $Desired > Predicted > Deserved > Minimum$. In our study, higher ratings indicate higher agreement to collection of health-related browsing activities. Hence, higher ratings indicate lower privacy expectations. Therefore, to test the significance of ordering, we analyze the reversed versions $Desired < Deserved < Predicted < Minimum$ and $Desired < Predicted < Deserved < Minimum$. Page test for ordered alternatives [63] is significant for $Desired < Deserved < Predicted < Minimum$ ($T = 5.37$; $p < 0.001$), but not for $Desired < Predicted < Deserved < Minimum$ ($T = 1.77$; $p < 0.039$) at Bonferroni adjusted $\alpha = 0.025$. This supports our hypothesis that privacy expectation types can be ordered.

Using Binomial interval estimation, we compute interval estimates for p, the proportion of the population that would give different scores for two different privacy expectation types. We assume that our sample ($n = 1038$) is representative of the population. Table 2.8 lists the 99.2% confidence intervals, computed using the Wilson score, for pairs of privacy expectation types. The largest proportion of the population, 29% to 36%, rates Predicted and Minimum differently. The smallest proportion, 13% to 19%, rates Deserved and Desired differently. Between 19% to 26% of the population rates Desired and Minimum differently indicating that, even in a privacy-sensitive scenario, the desired level of privacy can differ from the minimum tolerable level of privacy.

Table 2.9 lists Spearman pairwise rank correlations for pairs of privacy expectation types. Correlations indicate whether higher (lower) ratings for one type also corresponds to higher (lower) ratings for another type. Correlations between all pairs are

Table 2.8: Interval estimates for the proportion of the population rating privacy expectation types differently (n=1038).

Pair	**[99.2% Confidence Interval]**
Predicted, Minimum	[0.29, 0.36]
Predicted, Desired	[0.23, 0.30]
Predicted, Deserved	[0.22, 0.30]
Minimum, Desired	[0.19, 0.26]
Minimum, Deserved	[0.17, 0.23]
Deserved, Desired	[0.13, 0.19]

significant ($p < 0.0001$) with strong ($\rho = \pm 0.8$) to moderate ($\rho = \pm 0.5$) correlation. The highest correlation is between Deserved and Desired ($\rho = 0.81$) indicating that people who have higher desire for collection feel more that they deserve the collection. The lowest correlation is between Predicted and Minimum ($\rho = 0.57$).

Table 2.9: Spearman rank correlation (ρ) or pairs of privacy expectation types (n=1038).

Pair	ρ	**Prob**$> \lvert\rho\rvert$
Deserved, Desired	0.81	<0.0001
Minimum, Deserved	0.77	<0.0001
Minimum, Desired	0.73	<0.0001
Predicted, Deserved	0.63	<0.0001
Predicted, Desired	0.61	<0.0001
Predicted, Minimum	0.57	<0.0001

Demographic Effects We analyzed whether demographics (gender, age range, household income and education level), experience (prior use of websites and duration participants have had a bank account) impact ratings for privacy expectation types ($n = 1038$). Accounting for the number of levels of demographic and experience attributes, we set Bonferroni adjusted $\alpha = 0.002$. For pairwise comparisons α value was further divided by the number of comparisons.

- *Gender*: A Mann-Whitney test found no significant differences in scores by gender (male vs. female) for the four privacy expectation types. A person's gender does not seem to influence their expectations about data privacy practices.

- *Age*: Spearman pairwise rank correlations showed significant ($p < 0.0001$), but weak negative correlations between age range and scores of all expectation types: Minimum ($\rho = -0.25$), Deserved ($\rho = -0.21$), Desired ($\rho = -0.19$) and Predicted ($\rho = -0.19$). A person's age influences their expectations about data privacy practices; as age increases, they agree less to collection of health-related browsing activities.

  Kruskal-Wallis Rank Sum test indicated significant overall difference in scores by age range: Minimum $\chi 2(3) = 83.40$, $p < 0.0001$; Deserved $\chi 2(3) = 49.64$, $p < 0.0001$; Desired $\chi 2(3) = 45.68$, $p < 0.0001$; and Predicted $\chi 2(3) = 36.72$, $p < 0.0001$.

  Pairwise comparisons using Wilcoxon test showed the following significant differences by expectation type and age range. For Minimum, 18-29 vs. 30-44, 18-29 vs. 45-59 and 18-29 vs. 60+ ($p < 0.0001$). For Deserved, 18-29 vs. 45-59, 18-29 vs. 60+ and 30-44 vs. 60+ ($p < 0.0001$). For Desired, 18-29 vs. 45-59 and 18-25 vs. 60+ ($p < 0.0001$), and 30-44 vs. 60+ ($p < 0.0003$). Lastly, for Predicted, 18-29 vs. 60+ and 30-44 vs. 60+ ($p < 0.0001$).

- *Education Level*: Spearman pairwise rank correlations showed significant ($p < 0.0003$), but weak negative correlations between education level and Minimum ($\rho = -0.13$), Deserved ($\rho = -0.12$), Desired ($\rho = -0.11$) types, but not the Predicted type. Education does not seem to impact what people predict about a data privacy practice. To a small extent, it impacts what they desire, tolerate, and feel that they deserve.

- *Household Income*: Spearman pairwise rank correlations showed significant ($p < 0.00014$), but weak negative correlations between household income and Minimum ($\rho = -0.23$), Deserved ($\rho = -0.21$), Desired ($\rho = -0.19$) types, but not the Predicted type. Household income does not seem to influence what people predict about a data privacy practice, but it influences what they desire, tolerate, and feel that they deserve.

Experience Effects There was no significant difference in scores based on prior use of health and banking websites. We use the duration participants have had a bank account as an indication of their "investment" in the scenario. In our conceptual model, evaluation of investment and reward may critically determine the Deserved type. Spearman pairwise rank correlations showed significant ($p < 0.0001$), but weak negative correlations between investment and all expectation types: Minimum ($\rho = -0.20$), Deserved ($\rho = -0.19$), Desired ($\rho = -0.19$) and Predicted ($\rho = -0.18$). However, investment has significant moderate correlation with age ($\rho = 0.54$). Older participants tend to give lower scores. Hence, to control for the impact of age, we analyzed the correlation between investment and expectation types for each age group. There is significant weak to moderate correlation ($\rho = -0.28$, $p = 0.0016$) between 18-29 and the Deserved type. Other correlations are not significant. This supports the hypothesis

that investment impacts the Deserved type more than other types; investment in time, effort, money etc. impacts whether people feel that they deserve a data privacy practice or not.

2.2.2.3 Results

Multiple Types of Privacy Expectations Exist Empirical results support the hypothesis that privacy expectation is a multi-level construct with Desired, Deserved, Predicted and Minimum types. Hence, people have different types of privacy expectations. A person can have multiple types of privacy expectations for a given scenario. Different groups may have different types of privacy expectations. For example, none of the types varies between males and females, but all the types vary between 18-29 years old and 60+ years old. Younger people desire, predict, feel that they deserve and tolerate data privacy practices differently than older people. Predicted type does not vary by education level and household income, but the Deserved, Predicted and Minimum types do.

We elicited participants' privacy expectations regarding banks collecting participants' health-related browsing activities. Even in this privacy-sensitive scenario, we observe small, but significant impact of types. It is likely that for less privacy-sensitive scenarios the effect of types will be larger. For example, in our banking context, about 15% of participants "disagree" to Predicted and "strongly disagree" to Desired collection of health-related browsing activity. The same participants, in a search context such as Google search, are likely to "strongly disagree" to Desired but agree more to Predicted collection of health-related activity. Our study estimates that 23% to 30% of the population has different Desired and Predicted privacy expectations. For less privacy-sensitive scenarios, the proportion of the population as well as the magnitude of the difference may be larger.

The overall difference between Desired and Deserved is not significant in a privacy-sensitive scenario. However, that may change in a less privacy-sensitive scenario. For example, in an entertainment context, people may not desire watching advertisements in return for a free service, but they may feel that they deserve watching advertisements since they did not pay. The interaction between context sensitivity and the Deserved type requires investigation.

Privacy Expectation Types can be Ordered As proposed in the conceptual model, results show significant ordering among different types of privacy expectations. The Desired level of privacy is higher than the Minimum level of privacy; a person can ideally wish for a higher level of privacy and yet tolerate a lower level of privacy when essential. Understanding what constitutes "essential" could help businesses balance user privacy and product functionality. Businesses could also benefit from understanding the tradeoffs between addressing the Desired and Minimum privacy expectations of

customers. In our scenario, at least 25% of 18-29 years old "disagree" to the Desired level, but are "neutral" to the Minimum level. However, at least 25% of 60+ years old "strongly disagree" to both the Desired and Minimum levels. A business could achieve different tradeoffs between user privacy and cost by addressing either the Minimal or the Desired levels; by addressing the Minimal level the business can meet the privacy expectations of younger users, and by addressing the Desired level the business can meet the privacy expectations of both younger and older users.

Knowledge Impacts the Predicted Type We hypothesized that knowledge of privacy practices may impact the Predicted type more than the other types. The proportion of "Don't know/ Not sure" responses is significantly higher for the Predicted type. People seem to believe that they need knowledge to express their Predicted privacy expectation. Hence, knowledge does impact the Predicted type more than the other types.

Investment Impacts the Desired Type We hypothesized that evaluation of "investment and reward" in a scenario would critically impact the Deserved type. Results partially support this claim. We used the duration participants have had their banking account as an indication of their investment. In the 18-29 age range, low investment significantly increases the feeling that one deserves collection of data and high investment significantly decreases the feeling that one deserves collection of data. Investment, however, does not impact other types in the 18-29 age range, and it does not impact any type for the other age ranges. To assess the impact of investment better, we have to capture all the investment in terms of time, effort, money etc.

2.2.2.4 Limitations

Our sample consists of United States adults with access to the Internet. The sample is representative of an online population and not necessarily the general population. Our results may not be generalizable to other countries. It would be interesting to conduct further studies on samples from other countries and compare them with our results.

We conducted an online study to elicit user expectations. It would be beneficial to supplement our results with results from in-lab studies conducted under more controlled conditions. In our study, we employed partial counterbalancing where we reversed the expectation-related questions for half the participants. Previous studies [44, 11, 12] have used partial counterbalancing to achieve a balance between controlling for order effects and increasing experimental power. Although a design with complete counterbalancing could further control for order effects, it would require a much larger sample size.

Results from our full study ($N = 1249$) and our pilot study ($N = 130$) are statistically similar. This suggests that our results are repeatable. However, further studies can provide a higher degree of confidence in reliability.

We used a specific health and banking scenario to demonstrate significant differences among four privacy expectation types. Further, certain orderings among the types were significant. However, since expectations can vary by context, results may vary for other scenarios. Nevertheless, our results show that different types of privacy expectations exist and they can be ordered. Further, the design of our empirical study showed how we could measure different privacy expectation types in practice.

Our conceptual model includes four privacy expectation types. These cover the expectation types discussed in prominent privacy theories. Nevertheless, other privacy expectation types may exist. Further qualitative studies are required to identify existence of such types.

2.3 Types of Mismatched Privacy Expectations

A mismatch in privacy expectation occurs when users' privacy expectations do not match websites' actual data practices. To identify mismatched privacy expectations, we can compare privacy expectations elicited from users with website data practices identified from analyzing privacy policies. Mismatches can exist in one or more of Desired, Predicted, Deserved or Minimum expectation types. For example, consider that a banking website collects users' health information. If users do not desire such collection, then there is a mismatch in the Desired expectation type. If users do not predict it, then there is a mismatch in the Predicted type. If users feel that they do not deserve such collection, then there is a mismatch in the Deserved type. Lastly, if users do not tolerate it, then there is a mismatch in the Minimum type.

2.3.1 Mismatches from a Single Privacy Expectation Type

For a given type of privacy expectation, there can be different types of mismatches. Consider for example the Predicted type. Say we interpret expectations elicited from users as Yes or No. When a user predicts that the website will engage in data practice, we interpret that as a Yes; when a user predicts that the website will not engage in data practice, we interpret that as a No. As discussed in Section 2.1, collection and sharing data practices may be annotated as Yes, No, Unclear or Not addressed. In this scenario, comparing user expectations with actual data practices results in eight potential combinations shown in Table 2.10.

It is worth taking a closer look at the implications of the different types of mismatches. Although, both Yes–No and No–Yes are mismatches, they may impact users' perception of privacy violations differently. In the case of Yes–No, the website will collect or share information, but users optimistically expect it not to. Due to lack of awareness that the website shares information, users may decide to use the website. By doing so, they give up data that they do not want to share, resulting in a violation of their privacy. Although the website discloses its data practice in its policy, from a user viewpoint, the

Table 2.10: Potential mismatches from a single privacy expectation type.

	User:	Yes	No
Website:	Yes	Match	Mismatch
	No	Mismatch	Match
	Unclear	?	?
	Not addressed	?	?

(?) indicates that a data practice is unclear or not addressed in the privacy policy, and it cannot be determined if user expectation matches the data practice.

practice could be considered surreptitious unless users are appropriately and explicitly made aware of it. When found out, such data practices may damage a company's reputation.

In contrast, in the case of No–Yes, a website will not engage in a collection or sharing practice, but users pessimistically expect it to. As a result, users may have reservations to use the website or some features, which may affect their utility but not their privacy. In such cases, websites should aim to make users aware of the privacy-protective practices to assuage pessimistic expectations.

The number of unclear website data practices can be high. For the 16 websites analyzed in Section 2.1, ∼40% of collection data practices are unclear. Hence, it is important to analyze the impact of unclear data practices. Consider the Unclear–Yes case. If the website is really collecting information, then it would be a Yes–Yes match. If the website is not collecting information, then it would be a No–Yes mismatch. The same applies to Unclear–No. As discussed, a Yes–No mismatch, could potentially violate user privacy. Hence, for analysis purposes, we could treat Unclear as a likely Yes. We could use a similar approach for Not addressed–Yes and Not addressed–No.

Say deletion data practices are annotated as No, Yes–Full and Yes–Partial. We can analyze mismatches in case of the data deletion practice by considering two types of Yes values, Yes–Full and Yes–Partial, separately. We could also simplify the analysis by combining the two Yes values. In case of deletion, users may use a website if they think that the website allows deletion, whereas for collection and sharing they may not use the website. Hence, in case of deletion, the implications of No–Yes and Yes–No mismatches are reversed compared to implications for collection and sharing data practices.

2.3.2 Mismatches from Multiple Privacy Expectation Types

We discussed mismatches resulting from a single privacy expectation type. We can extend our analysis to more than one expectation type. For example, let us consider the Predicted and the Desired expectation types. For the Predicted type, users may answer Yes or No to whether they predict a website to engage in a given data practice. For the Desired type, users may answer Yes or No to whether they want a website to engage in the same data practice. To simplify our analysis, let us annotate the data practice as Yes or No only. This scenario results in eight potential combinations shown in table 2.11. When we consider only the Predicted type, there are four combinations (Yes–Yes, Yes–No, No–Yes and No–No) out of which two are matched expectations and two are mismatched expectations. When we additionally consider the Desired type, we have only two matches (Yes–Yes–Yes and No–No–No) and six mismatched expectations. Additional information could reveal that a case that looked like matched expectation e.g. Yes–Yes could be in reality a mismatched expectation e.g. Yes–Yes–No.

Table 2.11: Potential mismatches from multiple privacy expectation types.

Website (Yes or No)	**Predicted** (Yes or No)	**Desired** (Yes or No)	**Match or Mismatch?**
Yes	Yes	Yes	Match
Yes	Yes	No	Mismatch
Yes	No	Yes	Mismatch
Yes	No	No	Mismatch
No	Yes	Yes	Mismatch
No	Yes	No	Mismatch
No	No	Yes	Mismatch
No	No	No	Match

Simultaneously analyzing two privacy expectation types provides additional insights into implications of mismatched privacy expectations. Let us consider two cases where considering the Desired type in addition to the Predicted type changes the meaning of match (Yes–Yes) or mismatch (Yes–No) in the Predicted expectation type.

- *Matched Expectation (Yes–Yes)*: Let us analyze the case where users' Predicted expectation matches a website data practice (Yes–Yes). Considering the Desired type in addition to the Predicted type results in either Yes–Yes–Yes or Yes–Yes–No. When we have Yes–Yes–Yes, user expectation truly matches the website data practice. However, Yes–Yes–No is a mismatch i.e. users may predict that the website engages in the data practice, but they desire it to be different. For a collection data practice, Yes–Yes–No indicates that users are aware that website

will collect information, but they do not want it to happen. User may continue to use the website due to lack of awareness of other websites that do not collect information. If a monopoly exists, users may have no choice but to continue using the website. For example, users may know that Google search website collects privacy-sensitive user data, but they may not want Google to collect the data. Further, they may continue using Google because they are not aware of alternative search websites (e.g. DuckDuckGo.com) that do not collect privacy-sensitive user data.

- *Mismatched Expectation (Yes–No)*: Similarly, in case of a mismatch due to a website engaging in unexpected practices, the Desired expectation type may change the meaning of the mismatch. For example, when a Yes–No mismatch for the Predicted type is combined with a Desired expectation the meaning of the mismatch changes. In a Yes–No–No mismatch, users both incorrectly think that a website will not engage in a data practice and desired that it should not. They may decide to use the website and lose data privacy. For Yes–No–Yes, users want the website to engage in a practice, but do not predict it to do so at the moment. For instance, users may want a website to provide personalized services based on their data. In this scenario, users may decide not to use the website and lose utility, but not data privacy.

We can similarly analyze other cases for the Predicted and the Desired types listed in Table 2.11. We could consider Unclear and Not Addressed in our analysis. We could analyze more than two expectation types simultaneously.

2.4 Impact of Website Data Practices

To understand the impact of website data practices on user privacy expectations, we have to elicit and measure privacy expectations for different website data practices. Because people have different types of privacy expectations, simply asking them what they "expect" can lead to different interpretations. Hence, we have to determine the type of privacy expectation that we need to elicit. Otherwise studies may inadvertently use questions such as "would you expect it to access your precise location?" to measure what users think (Predicted type) [10] and "how much did you expect this app to be accessing this resource?" to measure what users want (Desired type) [64]. They may also or ask users to rate the statement "This application meets my privacy expectations" to measure expectations [11, 12] that are related to the Deserved or Minimum types.

We should explicitly measure the privacy expectation type that is relevant. Consider the challenge of designing usable short form notices that could complement long and difficult to comprehend website privacy policies. Short form notices could inform users about website data practices that do not match users' privacy expectations. To design such notices, we have to decide the type of privacy expectation – Desired, Predicted,

Deserved or Minimum – we should measure to identify mismatches; some privacy expectation types may be more effective for designing notices than others. For example, Acquisti et al. [49] note that privacy preferences and privacy decision making are prone to uncertainty, context-dependent, shaped by heuristics and cognitive biases, malleable and easily influenced by framing. Elicited privacy preferences can therefore be difficult to generalize, and actual behavior often deviates from stated preferences [65]. If preferences implicitly measure the Desired type, then the Desired type may not be reliable for designing notices since mismatches in the Desired type may fluctuate often. The Predicted type significantly depends on user privacy knowledge and focuses on expectations of what is likely to happen. Users expectations based on knowledge is less likely to vary than those based on desires. Hence, the Predicted type may be a better than the Desired type for designing short form notices. The Minimum type may also be a good candidate because it measures whether data practices meet the minimum user required standard.

We discuss how we can study the impact of website data practices on user privacy expectations. To identify website data practices, we can analyze website privacy policies as described in Section 2.1. To identify mismatches, we can elicit privacy expectations of users and compare them with actual data practices identified from privacy policies. We examine the impact of 17 data practices described in Table 2.1 on the Predicted privacy expectation type. We elicit user privacy expectations in the sense of "expected occurrence likelihood." By using semi-automated techniques, we could scale the task of identifying website data practices to a large number of websites [43]. Our results show that we could build models that can predict user expectations. Hence, our approach for identifying mismatches could scale up.

2.4.1 Study Details

To assess the impact of different website scenarios on privacy expectations, we conducted an online study involving 16 websites (see Table 2.2 and 240 participants (N=240). We opted for a between-subjects design to prevent fatigue and learning effects and which we asked the participants to answer questions about one website randomly assigned to them. Website type (health, finance, dictionary) and popularity (low, high) were the main independent variables in our study, resulting in a 3x2 design with six conditions. We based website type and popularity on website categories and traffic rankings respectively obtained from Alexa.com [38]. We studied 16 websites across three website types (7 Health, 7 Finance, 2 Dictionary). Fifteen participants were assigned to each website, resulting in the following number of participants per condition: 60 in Health-Low, 45 in Health-High, 60 in Finance-Low, 45 in Finance-High, 15 in Dictionary-Low, and 15 in Dictionary-High.

Survey Questionnaire We designed a questionnaire to measure user expectations for eight collection data practices (4 information types collected with or without account),

eight sharing data practices (4 information types shared for core or other purposes), and one deletion data practice. These website practices, listed in Table 2.1, were treated as 17 dependent variables.

The survey questionnaire consisted of three sections: introduction, main questionnaire and post-questionnaire. Privacy-related questions, which could bias participant responses, were asked in the post-questionnaire. While designing the questionnaire, we used think-aloud and verbal-probing cognitive interviewing techniques [57] in pilot tests with six participants. We tested whether participants understood the questions. We iteratively refined the questionnaire based on participant feedback. We summarize the questionnaire below. The full questionnaire is in Appendix A.

At the beginning of the questionnaire, we explained the purpose of the study. We framed the purpose of the study as understanding user opinions about websites rather than their knowledge of data practices, to avoid self-presentation issues associated with knowledge questions [61]. We also did not mention privacy or data practices to avoid biasing participants. After explaining the purpose, we asked whether participants had visited or used the assigned website before.

We instructed the participants to familiarize themselves with the website assigned to them. Since participants may explore websites in different ways, we wanted them to look at what they considered important and did not want to bias their thinking by providing too specific instructions. Based on participant feedback from our in-lab pilot tests, we asked participants to look at the website for 2–3 minutes. Initially, we had instructed the participants to take their time familiarizing themselves with the website. However, after about three minutes of interaction, our in-lab participants were either ready to provide their opinions or were not sure what else to look at. Two participants specifically told us that it would be helpful if we told them how much time they should spend looking at a website. Because the website was opened in a separate browser window, participants could go back to the website at any point during the study.

After participants interacted with the website, we provided definitions of contact, financial, health and current location information.

- *Contact Information*: Examples include (but are not limited to) email address, postal address, phone number, home phone number, etc.
- *Current location*: Current, real-time location of a user accessing the website (city-level or more precise)
- *Health information*: Examples include (but are not limited to) user's medical history, family medical history, user's health insurance information, etc.
- *Financial information*: Examples include (but are not limited to) bank account details, credit/debit card numbers, credit ratings/history etc.

In the main part of the questionnaire, we asked participants about their expectations regarding different website data practices, listed in Table 2.1. First, we asked them questions about data collection practices in two scenarios: collection without account and collection with account. Before asking questions related to a scenario, we showed scenario descriptions. For instance, for the collection without account scenario, we showed the description "*Imagine that you are browsing [website name] website. You do not have a user account on [website name], that is, you have not registered or created an account on [website name].*" We then asked them about their expectations concerning whether and how the website collects different types of data. These questions were framed as likelihood questions: "*What is the likelihood that [website name] would collect your information in this scenario?*" Note that we framed the questions as "would collect" in order to capture participants' objective expectations, and not what they would prefer. We provided a 4-point scale {Likely, Somewhat likely, Somewhat unlikely, Unlikely} as the response option. We wanted respondents' "best guess" and thus did not provide a neutral or not sure option. We did so because users often do not read privacy policies and decide about data practices of a website based on incomplete information, that is, their best guess. We asked an open-ended question to understand how they thought the website collected their information without having an account on the website. After answering questions about the without account scenario, participants read the scenario description for collection with an account and answered the same questions regarding this scenario.

After collection-related questions, we asked participants questions regarding data sharing practices. We first asked them questions about a scenario where data is shared for core purposes, which we defined as sharing only for the purpose of providing a service that the user requested. We them asked them questions regarding a scenario where data is shared for other purposes, which we defined as a purpose unrelated to providing a service that the user requested. To answer the questions, participants had to understand three concepts. First, what are core purposes for the given website? Second, what are other purposes for the given website? Lastly, with whom could the website possibly share information? To encourage them to think about these concepts, we asked them three open-ended questions before asking questions related to sharing. Concerning the data deletion practice, we asked participants whether they expected that the website would allow them to delete all, some or none of their data.

In the post-questionnaire, we captured different user characteristics in order to study their impact on the participants' privacy expectations. We list these characteristics in Table 2.12. We ordered the questions based on ease of answering, level of threat, and effect on subsequent answers [61]. First, we asked questions about their *past experiences* with the assigned website including if they had an account on the website, how much they had used the website, familiarity with the website and the website's perceived trustworthiness. Users' past experience may influence their expectations, for example, having an account may expose them to additional parts of a website that may improve their awareness of the website's data practices. Participants then provided demographic information (gender, age, education, occupation) and whether they had a background in

Table 2.12: Studied website and user characteristics.

Website characteristic	
Type	Finance
	Health
	Dictionary
Popularity	More
	Less
Context	Private
	Government
User characteristic	
Demographic: age, gender, education, occupation computer background, state of residence	
Privacy protective behavior	
Familiarity with privacy concepts and tools	
Knowledge of privacy concepts and tools	
Negative online experience	
Online privacy concern	
Experience with website: amount of recent use, has account, familiarity, trust	

computer-related fields, which may indicate an enhanced understanding of online data practices. We also asked for their U.S. state of residence, to assess whether privacy regulation on the state level, e.g., in California, impacts privacy expectations. We further included questions about privacy-protective behavior [66] and their familiarity and knowledge of privacy concepts and privacy-enhancing technologies [67]. We also asked whether participants had negative online experiences [68], as they may expect data practices to be more privacy invasive. Lastly, we included the 10-item IUIPC scale [69] to assess online privacy concerns.

Study Deployment & Demographics Our study adhered to appropriate practices for human subject studies. To recruit participants efficiently and rapidly, we used the Amazon Mechanical Turk crowdsourcing platform [70]. Research has shown that the Mechanical Turk sample pool is more diverse than traditional sample pools [71], and that data quality is typically good [71, 72, 73]. In February 2015, we recruited 240 participants. We restricted participation to individuals located in the United States,

with at least a 95% approval rate and at least 500 completed tasks on Amazon Mechanical Turk. Participants received $3.50 for completing the study. Each participant was randomly assigned to one of the 16 websites. We implemented our survey on SurveyGizmo. Participants were redirected from Amazon Mechanical Turk to SurveyGizmo to complete the survey. We used a combination of SurveyGizmo and Mechanical Turk features to ensure that participants took the survey only once. We implemented timers to measure how long participants interacted with a website and to measure time spent on survey questions. As instructed, participants, spent on average 1.99 min (SD=2.41, median=1.56) interacting with a website. Statistical analysis did not show a significant impact of the amount of time spent on a website or on the survey questions.

To ensure data quality, we screened for participants that completed the study in less than 10 minutes (pilot tests suggested a 30-minute completion time), and checked whether participants answered two questions about prior experience with the assigned website at the beginning and the end of the survey consistently. All participants passed at least two of three quality criteria.

The 240 participants completed our online survey in 22.5 minutes on average (SD=12.8, median=18.6). The sample was 42% female and 58% male. The average age was 34.4 years (SD=10.3, median=32). The majority (85.3%) had at least some college education and 61.6% reported an Associates, Bachelors or Graduate degree. A fifth of the participants (19.5%) had a college degree or work experience in a computer-related field. The top primary occupations were administrative staff (17.5%), service (14.1%), and business/management/financial (12%).

Scenario Parameters We defined multiple scenarios that varied in key parameters, namely data practices and website characteristics. We hypothesized that these parameters may influence privacy expectations and mismatches.

Data Practices of Interest
We decided to focus on data practices concerning *collection, sharing and deletion of personal information* as prior research has shown that users are especially concerned about surreptitious collection, unauthorized disclosure and wrongful retention of personal information [29]. We considered the collection and sharing of four categories of privacy-sensitive information [74, 75, 48] – *contact information*, *financial information*, *health information* and *current location* – described earlier.

We further distinguished between scenarios in which users have or do not have an *account with the website.* Websites typically collect data when users create an account, often explicitly provided by the user. Hence, users may have different expectations depending on whether they have an account or not. In general, users may not be aware of implicit or automated data collection, e.g., of IP addresses and cookies. Websites may use IPs, email addresses and other information to acquire additional data about individuals, such as purchase history or interests, from social media services and data brokers [5].

Similarly, information sharing with third parties, while abundant, is less visible to users. Websites assume to have the users' permission because they are using the website and therefore implicitly consent to its privacy policy. We distinguish between third party sharing for *core purposes*, such as sharing a user's information to provide the requested service (e.g., payment processing or providing contact information to a delivery service), and sharing for unrelated *other purposes*, such as advertising or marketing. In all, we studied 17 data practices summarized in Table 2.1.

Website Characteristics
To understand whether mismatched privacy expectations vary based on context, we considered three website characteristics: website type, popularity and ownership. *Website type* may influence what information users expect a website to collect [76]. We selected three website categories: finance, health and dictionary. Users may expect finance and health websites to collect sensitive information (health or financial data, respectively). In contrast, users may not expect dictionary websites to collect sensitive information. In the financial category, we included banking, credit card and online payment websites. In the health category, we included pharmacy, health clinic and health reference websites. Website categories were determined using Alexa website categories [38].

Users' expectations may be influenced by their offline interactions with entities affiliated with a website, such as visiting a bank branch or a clinic. Hence, we included websites with *offline interactions* as well as online-only websites in the health and financial categories; dictionary websites were online-only.

Interestingly, popular financial websites have been shown to have more privacy-invasive data practices than less popular ones [77]. Therefore, we studied websites of comparable utility but varying in *popularity*, as determined by their traffic rankings [38].

For a given website type, *government or private ownership* may influence user expectations. Our sample population was limited to the United States, and in the post-Snowden era, people may expect government websites to be more privacy invasive than private websites. Hence, we studied whether user expectations varied between government and privately-owned health and financial websites. Table 2.12 summarizes the website characteristics that we considered in our model.

2.4.2 How Privacy Expectations Vary

Impact of Website Characteristics We find that a website's type has a significant impact on user expectations. This implies that what data practices users expect a website to engage in is influenced by the type of website. We did not find significant differences for popularity or ownership, suggesting they play no or a lesser role in shaping privacy expectations. For example, users expect data practices of BankofAmerica.com, a finance website to be different than those of WebMD.com, a health website. However, they have similar expectations for two finance websites even if one of them is more

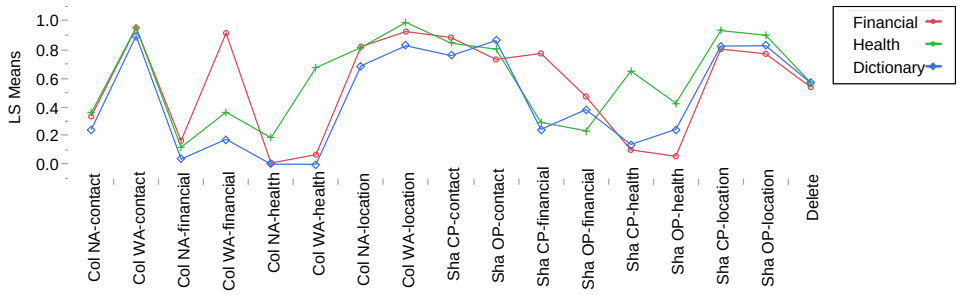
popular than the other (e.g., in our dataset BankofAmerica.com's popularity rank is 33 and WoodlandBank.com's is 915,921). Similarly, expectations do not differ between privately-owned and government-operated websites.

We used a mixed-model ANOVA to analyze the impact of website type and popularity on user expectations. We considered website type (health, finance, dictionary) and popularity (high, low) as nominal between-subjects independent variables. We considered participant expectations concerning the 17 data practices as continuous repeated measures dependent variables (DV), which, as a group, measured users' overall expectation. We verified that the group of DVs has an approximate normal distribution with a normal-quantile plot of a linear combination of the individual DV scores. A Shapiro-Wilk W test showed only moderate departure from normality (W=.988, p=.041).

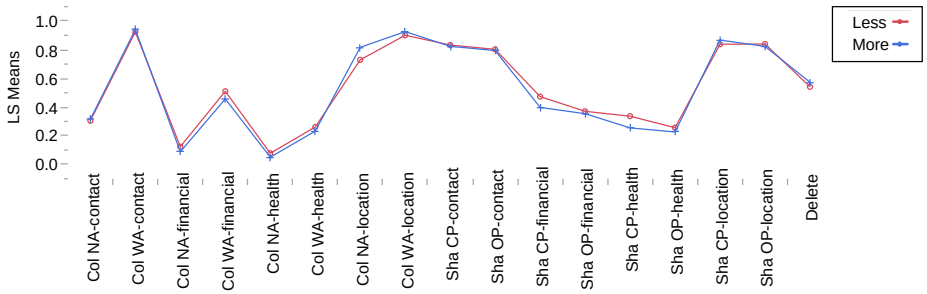
Results showed that interaction of website type and data practices was significant ($F(32.438)$=12.819, p <.0001), see Figure 2.3a for an interaction plot. This interaction effect suggests that website type impacts what data practices users expect. Compare, for instance, the impact of financial website type on users' expectations concerning collection of financial and health information from registered users (*COL WA-financial*), *COL WA-health*). Higher Least Square Means value implies that users are more likely to expect a data practice. Users expect financial websites to collect financial (high $LSMeans$), but not health data (low $LSMeans$). Figures 2.3b–2.3d further show interactions of website popularity and ownership, which were not significant. Note that only the health and finance categories contained government-operated websites, dictionary websites are therefore not shown in Figures 2.3c and 2.3d.

We also studied the impact of website type on individual data practices. The distribution of values of individual data practices was non-normal. We treated them as two-level nominal variables and used a χ^2 statistical test. Figure 2.4a shows what information types participants expect websites to collect from registered users. If *LS Means*>0.5, users are likely to expect the data practice. Type of website has a significant impact for expectations of collection of financial ($\chi^2(2,240)$=87.7, p <.0001, R^2=.302) and health information ($\chi^2(2,240)$=105.826, p <.0001, R^2=.3935), but not for collection of contact and current location information. Users expect all types of websites to collect contact and location information when they have an account. However, they expect only financial websites to collect financial data and health websites to collect health data. A financial website collecting health data would lead to a mismatch in expectations. Most financial websites we studied do not collect health data. However, one financial website in our study, BankofAmerica.com, collects health information when users have an account, which violates user expectations.

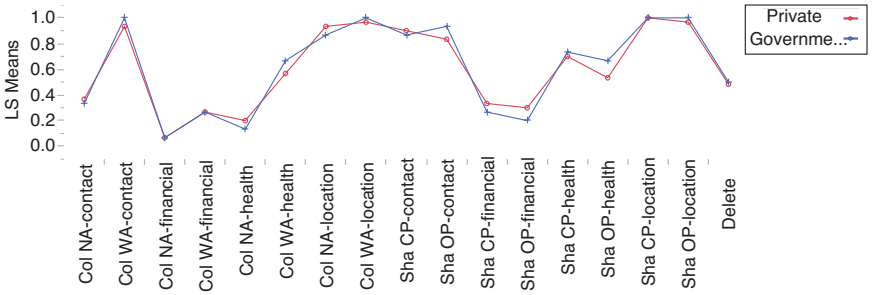
As shown in Figure 2.4c, in the without account scenario, participants expect only collection of location information, but for all types of websites. Participants are unlikely to expect websites to collect contact, financial and health data from users without an account. As we will discuss shortly, websites can collect contact and financial data without an account, leading to a mismatch with expectations.



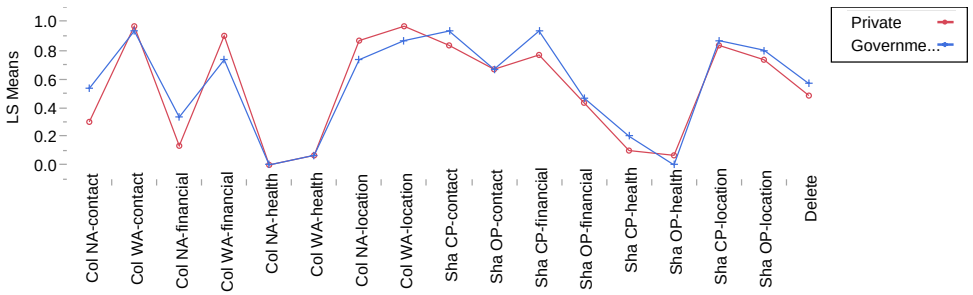
(a) Website type (*sig.*)



(b) Popularity (*n.s.*)

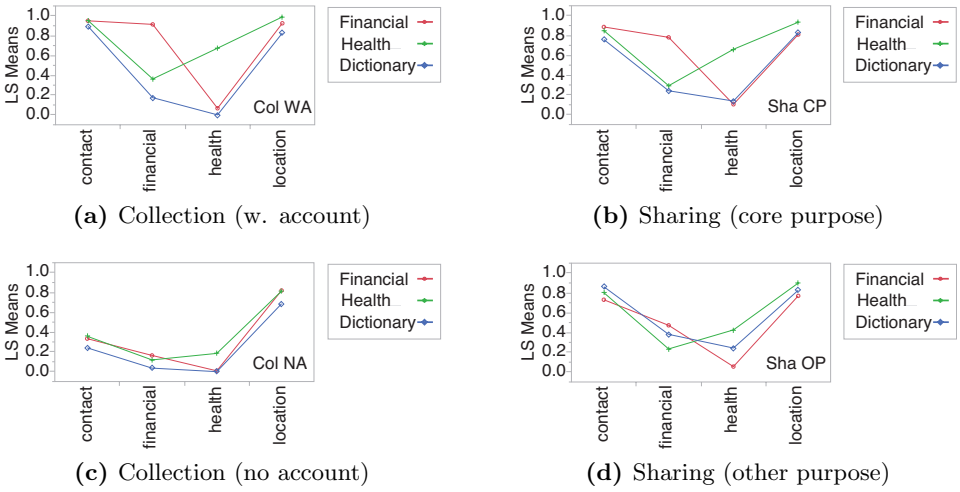


(c) Ownership for health websites (*n.s.*)



(d) Ownership for finance websites (*n.s.*)

Figure 2.3: Interaction of website characteristics and user expectations for the 17 data practices. Higher Least Square Means value implies users expect data practice to be more likely (Col: Collection, Sha: Sharing, WA: With Account, NA: No Account, CP: Core Purpose, OP: Other Purpose).



(a) Collection (w. account) **(b)** Sharing (core purpose)

(c) Collection (no account) **(d)** Sharing (other purpose)

Figure 2.4: Interaction of website type and expectations for specific data practices. Website type significantly interacts with user expectations for financial and health information. Higher Least Square Means value implies users are more likely to expect a data practice.

Concerning expectations of data sharing, Figure 2.4b shows that participants likely expect all types of websites to share contact and current location information for core purposes. Website type has a significant interaction effect for expectations of sharing financial information ($\chi^2(2,240)=59.175$, $p<.0001$, $R^2=.1868$) and expectations of sharing health information ($\chi^2(2,240)=77.935$, $p<.0001$, $R^2=.2642$). Participants expect only financial websites to share financial data and health websites to share health data. One financial website, BankofAmerica.com, shares health information for core purposes, which violates user expectations.

Figure 2.4d shows expectations of websites sharing for other purposes. In this case, users expect all types of websites to share contact and location information for other purposes. They do not expect any type of website to share financial or health information for other purposes. Users expecting websites to share contact information for other purposes is interesting because, as we discuss later, most websites do not do so. Lastly, we did not find significant interactions of website type with participants expectations concerning websites' data deletion practices. Participants expected all website types to permit deletion of data, as shown in Figure 2.5, but this expectation does not match reality.

Further analysis shows that user expectations can vary for individual data types within a larger data type category. For example, for collection of contact information in the with account scenario, participants expected that websites were more likely to collect email address (93.3% participants) than postal address (75%) or phone number (70.8%). Expectations for specific data types can also vary within website sub-categories. For instance, for collection of health information in the with account scenario, participants

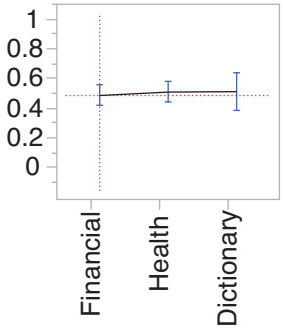


Figure 2.5: Website type does not impact deletion data practice. LS Means (least square mean) higher value implies users expect data practice to be more likely.

expected that pharmacy websites were more likely to collect health insurance information than medical history (66.6% vs. 53.3%), but health clinic websites were more likely to collect medical history than health insurance (67.7% vs. 54.8%). Although we could analyze expectations at a finer granularity, identifying mismatches in expectations at finer granularity is problematic because website privacy policies do not typically disclose data practices at such fine granularity. Privacy policies generally discuss data practices at the level of coarse grained categories such as contact information rather than email address or postal address.

We analyzed the effect of multiple user characteristics on participants' data practice expectations. We find that privacy knowledge, privacy concept familiarity, privacy concern, privacy-protective behavior, negative online experience, age, trust in website, website familiarity, whether participant has an account, and recent use have a significant impact on participants' expectations for certain data practices. Other user characteristics elicited in the survey had no statistically significant impact.

Impact of User Characteristics We analyzed the effect of multiple user characteristics on participants' data practice expectations. We find that privacy knowledge, privacy concept familiarity, privacy concern, privacy-protective behavior, negative online experience, age, trust in website, website familiarity, whether participant has an account, and recent use have a significant impact on participants' expectations for certain data practices. Other user characteristics elicited in the survey had no statistically significant impact.

For analysis, we considered user characteristics as naturally-occurring, continuous IVs. The DVs were the user expectations for the 17 data practices. Distributions of the individual DVs were non-normal. Therefore, we considered them as two-level nominal variables (Yes, No) and built a nominal logistic regression model for each DV. We assessed internal consistency of summated scale responses using Cronbach's α. For responses to online privacy concern, privacy concept familiarity, privacy knowledge, privacy protective behavior and negative online experience scales, reliability estimates

were 0.88, 0.91, 0.63, 0.78, 0.68 respectively. For building regression models, we standardized IV values. To avoid biasing the model due to collinearity of IVs, we computed bivariate non-parametric Spearman rank correlations between IVs and subsequently excluded IVs that had moderate or higher correlation (>0.5). Privacy concept familiarity and privacy-protective behavior were removed from regression models as they correlated with privacy knowledge. Website familiarity and whether the participant has an account were removed because they correlated with the amount of recent use. Our analysis of initial regression models showed that, among demographic variables, only age accounted for a significant amount of variance. Therefore other demographics were removed to improve reliability of the regression models.

As a result, each of the 17 final regression models contained six IVs: privacy knowledge, privacy concern, negative online experience, age, trust in website and recent use. Table 2.13 lists the user characteristics (IV) and regression models in which the IV was statistically significant in predicting user expectation (DV). Below, we explain the user characteristics (IVs) that can significantly predict user expectations (DVs).

- *Privacy Knowledge*: An individual's privacy knowledge impacts user expectations. Specifically, privacy knowledge can impact if a user expects the collection of health information from unregistered users. An individual with a one unit increase on the privacy knowledge scale is two times more likely to expect that a website will not collect health information without an account. Privacy familiarity and privacy protective behavior correlated with privacy knowledge, and are likely to impact users' expectations in a similar way. Recall that users expect websites, especially non-health websites, to collect health information only when they have an account. If a website did collect health information without an account, there would be a mismatch in expectations.

- *Privacy Concern*: Individuals with higher online privacy concern (IUIPC [69]) expect data practices to be more privacy invasive. Specifically, individuals with one unit increase in online privacy concern are twice as likely to expect that a website will collect current location information when users have an account. They are ∼1.6 times more likely to expect that a website will share contact and current location information for core purposes. Although, most users in our study expect such collection and sharing practices, the segment of users with higher privacy concern are even more likely to expect such practices.

- *Age*: Individuals' age impacts expectations regarding deletion; with one unit increase in age, they are ∼1.8 times more likely to expect that a website will not allow deletion of user data. Older users correctly expect websites not to permit deletion of user data. Hence, the likelihood of mismatch is higher in case of younger users.

- *Trust in Website*: User perception of a website's trustworthiness impacts expectations regarding sharing and deletion data practices. With a one unit increase in trust, individuals are ∼1.7 times more likely to expect that a website will not

share health and financial information for other purposes. They are 1.5 times more likely to expect that a website will share location information for core purposes. Lastly, individuals are twice as likely to expect the website to allow deletion of user data. Although, users' expectations based on trust hold for sharing practices, their expectations for deletion does not match reality.

User characteristic (IV)	User expectation (DV)	Model			IV		
		R^2	$\chi^2(6, N{=}240)$	p	Odds(No)	$\chi^2(1, N{=}240)$	p
Privacy knowledge	Collect health info without account	0.10	14.52	0.024	2.09	7.60	0.0058
Privacy concern	Collect location info with account	0.13	13.80	0.0319	0.49	7.22	0.0072
	Share contact info for core purpose	0.09	18.47	0.0052	0.64	5.94	0.0148
	Share location info for core purpose	0.08	15.34	0.0177	0.58	7.67	0.0056
Age	Allow deletion	0.13	30.53	<0.0001	1.77	10.88	0.0010
Trust in website	Share location info for core purpose	0.08	15.34	0.0177	0.65	4.44	0.0352
	Share financial info for other purpose	0.07	21.33	0.0016	1.80	16.82	<0.0001
	Share health info for other purpose	0.05	14.54	0.0241	1.68	1124	0.0008
	Allow deletion	0.13	30.53	<0.0001	0.53	13.64	0.0002
Recent use	Collect location info with account	0.13	13.80	0.0319	1.56	4.01	0.0451
	Share contact info for core purpose	0.09	18.47	0.0052	1.50	6.67	0.0098
	Allow deletion	0.13	30.53	<0.0001	1.56	7.83	0.0051

Table 2.13: Regression models in which specific user characteristics (IV) significantly impact user expectations (DV). *Odds(No)* indicates, for one unit increase in the IV value, the increase in likelihood that a user will not expect a website to engage in that data practice (*Odds(Yes)*=1 / *Odds(No)*).

- *Recent Use*: Participants self-reported use of the website in the last 30 days impacts expectations regarding three data practices. With one unit increase in usage, individuals are 1.6 times more likely to expect that a website will not collect current location information from registered users. Individuals are 1.5 times more likely to expect that the website will not share contact information for core purposes. Lastly, individuals are 1.6 times more likely to expect that website will not allow deletion. User expectations are likely to vary similarly based on website familiarity and whether the participant has an account, because both correlated with the amount of recent use. These results confirm our hypothesis that users who have more access to a website have different expectations. However, it is not always true that their expectations are more accurate. For instance, their expectations regarding deletion are more accurate, but expectations regarding sharing are not.

2.4.3 Mismatched Privacy Expectations

To identify mismatched privacy expectations, we compared participants' privacy expectations concerning a specific data practice with website data practices identified from analyzing privacy policies. The data practices were annotated as Yes, No, Unclear or Not addressed. We elicited the Predicted privacy expectation type. Participants rated their expectation of whether a website will engage in a specific data practice on a 4-point scale (Unlikely–1, Somewhat unlikely–2, Somewhat likely–3, Likely–4). We interpreted the ratings as indications of a positive (Yes) or a negative (No) expectation and compared them with annotations of data practices. The resulting eight combinations are shown in Table 2.10.

As shown in Figure 2.6, overall, expected and unexpected data practices varied for different information types, and collection and sharing scenarios. We analyzed mismatches when websites explicitly disclosed their data practices, as well as when websites were unclear or did not address the data practices. When data practices were explicit, we observed three important mismatches. Collection of contact information without an account was mainly a Yes–No mismatch, that is, participants did not expect websites to collect information, but websites did. Similarly, collection of financial information without an account was a Yes–No mismatch. Sharing of contact information for other purposes was also a mismatch, but a No–Yes mismatch, that is, participants pessimistically and incorrectly thought that websites would share their contact information. For the remaining data practices, participants' expectations either predominately matched website practices or the level of match was equal to the level of mismatch.

For the data deletion practice, 32% of participants expected websites to allow full deletion, but only 19% of the analyzed websites allow it. Similarly, 48% expected partial deletion, but only 12% of websites permit it. However, about 20% of the participants thought that websites would not allow deletion of any data and 19% of the websites do not allow deletion of any data. Participants' expectations were similar across the three

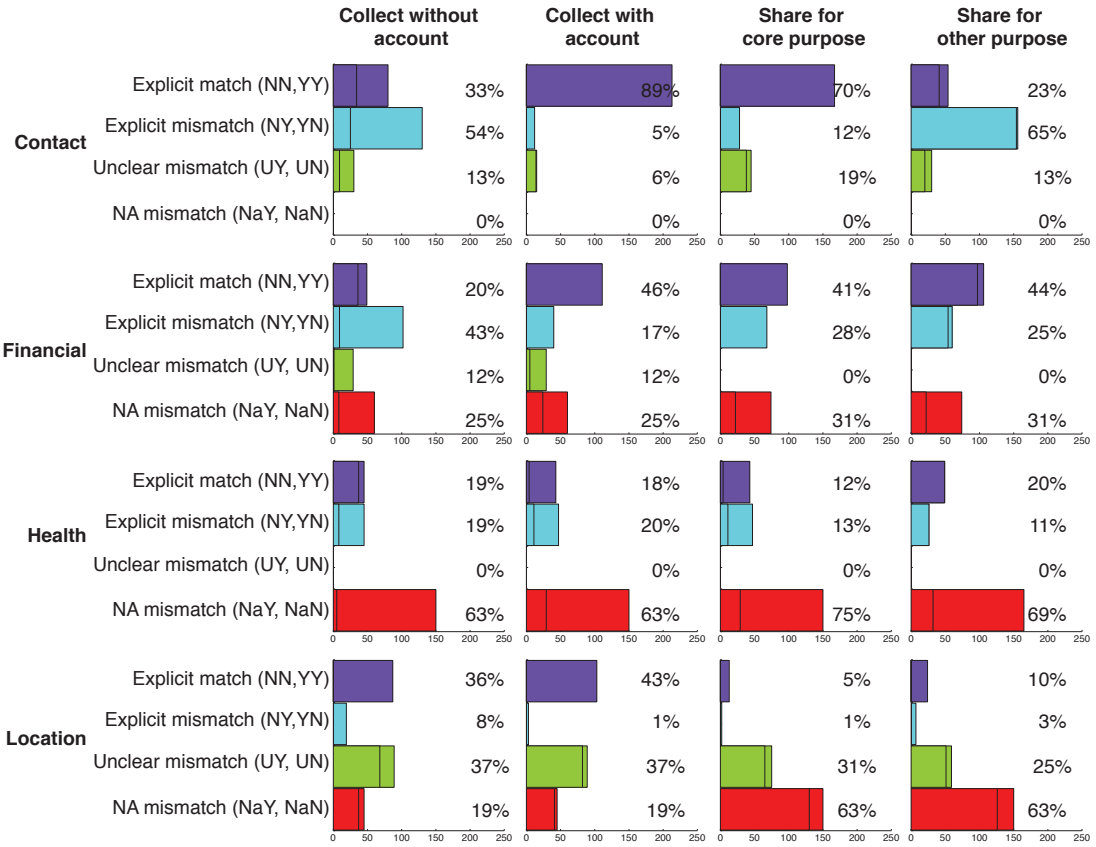


Figure 2.6: Matches and mismatches in user expectations. Explicit match or mismatch occurs when websites are clear about their data practice. When practice is unclear or not addressed, mismatch is not evident.

website types. There is a mismatch in expectations regarding deletion – participants seem to expect websites to allow deletion more than websites actually do.

The number of data practices that are unclear or not addressed in a privacy policy can be high. As shown in Figure 2.6, websites mostly do not address data practices regarding health information. In contrast, they are mostly unclear or do not address data practices regarding location information. Considering Yes–No mismatches to be more privacy invasive, let us assume that a website engages in a data practice when its disclosure is unclear or not addressed. For health information practices, this results in mainly Yes–No mismatches for all scenarios. However, for location information practices, it results in No–Yes mismatches.

2.5 Summary

We proposed a conceptual model for privacy expectation with four types of privacy expectations: Desired, Predicted, Deserved and Minimum. We validated our model using an empirical study. We found that different types of privacy expectations exist, and the types can be ordered based on the distinct levels of user privacy they represent. We use the design of the empirical study to operationalize measuring different types of privacy expectations in practice.

We studied the impact of website data practices on user privacy. We discussed how we can use privacy policy analysis to extract and analyze website data practices. We found that more than 40% of collection, sharing and deletion data practices were unclear or not addressed in privacy policies. We analyzed how the Predicted privacy expectations varied based on website and user characteristics.

We examined the types of mismatches that can occur due to a single expectation type as well as the types of mismatches that can result from interaction of multiple expectation types. For identifying mismatched privacy expectations, we proposed a practical approach that elicits user privacy expectations and compares them with website data practices stated in website privacy policies. Using the proposed approach, we identified mismatches in users' Predicted privacy expectations regarding collection, sharing and deletion data practices of websites.

3 Privacy Impact of Tracker Data Practices

To spread a new idea or influence thinking, we need to know about current norms, beliefs and behavior of people [78]. The more aligned a new idea is with existing norms, the easier it is for people to accept it. Work in cognitive psychology also shows that people are susceptible to cognition bias i.e. people are more likely to accept ideas that are aligned with their current thinking. Tracking user data on the Internet can allow us to gain an intimate understanding of users' opinions, values, behavioral intentions etc. For instance, how people encounter political information on the Internet can reveal their political party affiliation [79], which in turn can reveal whether ideological similarity of personal communication networks can impact perception of media credibility [80]. In a global setting, one country could use such data it obtains from people of another country to its advantage e.g. influencing options or policy in the other country.

Companies can embed small piece of code, called a tracker, within a website and track users' activities on the websites [7, 8]. Currently there are >4400 trackers [9]. Commonly used tracking technologies include beacons and cookies, and may be transparent to a website user [7]. Trackers can track activities such as frequency of visit, search keywords, videos watched and IP addresses. Prior research shows that users are concerned about tracking of user activities on the Internet [21]. Even more concerning is the linking or combining of activities from multiple contexts, e.g. insurance and health-care websites, that can reveal much more than tracking on separate contexts [5].

Prior research has investigated the prevalence of tracking on the Internet [7, 20], user awareness and concerns regarding tracking [21, 15, 5], technologies used for tracking (cookies, flash cookies, fingerprinting etc.) [22, 23, 24] and defenses against tracking[8, 25]. Several browser plugins such as AdBlockPlus, Ghostery, TrackingObserver, DoNotTrackMe, Collusion and PrivacyBadger are available to identify and block trackers present on websites.

We use network analysis to identify tracker data practices that enable linking user activities across different website categories. We analyze how linking of user activities can impact user privacy. We also analyze how trackers enable siphoning of user data. We use the term "siphon" to indicate a one-way channel that once set up will result in a continuous flow of personal information from the source e.g. users to the destination e.g. companies. We study whether and how current tracking mechanisms can be used to siphon data from one country to another.

3.1 Identifying Tracker Data Practices

To track users' activities on the Internet, companies can embed a small piece of code, called a tracker, within a website. As per Evidon, a company that maintains a tracker database, there are more than 4400 trackers in the wild [9]. When users interact with a website, a tracker on the website can collect information such as IP address, click-stream data, websites visited before visiting the current website etc. and send the information to the company that owns the tracker. If the tracker belongs to a company that owns the website, it is called a first-party tracker. Otherwise it is called a third-party tracker. Third-party trackers are more common and are considered more privacy invasive because they generally collect information for purposes such as advertising and marketing that are not directly related to the primary service provided by a website. Trackers occur on different types of websites including news, health, shopping etc. Depending on the functionality of the tracker, trackers can be classified into categories such as social media, analytics, comment, advertising, porn-advertising etc.

It is possible to track user activities on the Internet without using trackers on websites. Direct ownership of a website allows an entity to collect data about the website's users. It is possible to buy data about users from third-party companies [5]. Companies such as BlueKai combine data from offline and online sources, build fully-identifiable behavioral profiles of individual users, and sell the profiles on data marketplaces [5]. Information sources used by the companies include public data such as voter registration databases, occupation data from state license boards and bankruptcy records, and private sources of information such as in-store and online transactions, website interactions and social networking activity. A recent breach of a United States political party database revealed that the political party had bought detailed profiles of 200 million citizens [81], and the profiles contained names, addresses, birth dates, phone numbers and political opinions on a wide range of topics.

Several tools are available to identify third-party trackers on a website. For example, Ghostery (ghostery.com), a commercial tool, and TrackingObserver [8], a research tool, identify trackers on a website. In Figure 3.1, the Ghostery tool is showing 16 trackers on a banking website. Using tools such as the OpenWPM platform [82], we can automate the process of visiting websites and collecting trackers on the websites.

A tracker can not only collect user data from individual websites, but can also link and combine user data from multiple websites when it is present on multiple websites. Further, when the websites belong to different website categories, trackers can combine user data from different contexts. Trackers can build individual profiles of users by combining data from different contexts [5]. For example, if the same tracker is present on a banking website, a health website and a religion website, the tracker can combine user activities from the three website categories. If the tracker shares the user data with the bank, then the bank can infer users' religion and health condition. The privacy policy of a top bank [56] in the United States, for instance, states "Data [...] refers to other information that we collect through your internet or mobile activities or which

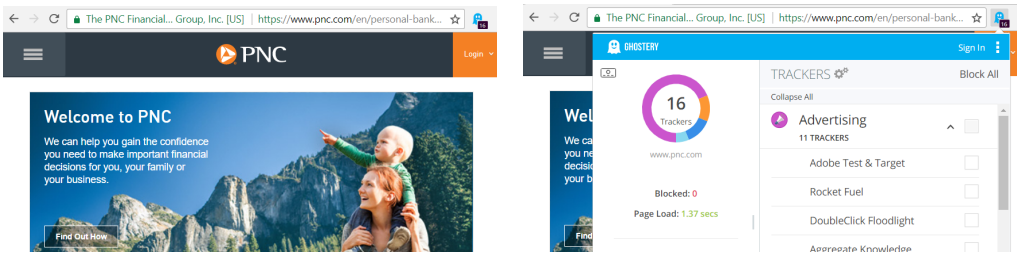


Figure 3.1: Ghostery tool showing 16 trackers on a banking website.

third parties may collect on our behalf. Such data may or may not be personally identifiable to you." The policy further says "For instance, they may keep track of how many of our ads you have seen on other web sites before visiting our Web site. This information is used to understand your browsing behavior and interests so that we can identify your financial needs and provide service and advertising that is tailored to you." The bank can use third-party trackers to collect user data from the bank's website as well as other websites to identify user activities and infer user needs.

3.1.1 Network Analysis

We use network analysis to identify how trackers enable linking of user data from multiple contexts. We consider different website categories as different contexts. We create an undirected two-mode network consisting of two types of nodes – websites and trackers – and analyze the links between them. We identify top websites and trackers that link user activities across website categories. We also identify strongly linked clusters of website categories that indicate linking of user activities among website categories.

Linking of User Activities To illustrate how we can identify linking of user activities, we apply network analysis to a network consisting of 50 most popular websites from each of banking and religion website categories. We identify banking and religion websites using Alexa (Alexa.com) website classification. We use Alexa website ranking to identity most popular websites in each category. Figure 3.2 shows an undirected two-mode network for the websites in banking (left) and religion (right) categories. The red squares indicate the trackers, the blue circles indicate banking websites and the pink circles indicate the religion websites. If a banking website has a tracker, then the blue circular node representing it is connected via an edge to a red square node representing the tracker. Similarly a pink circular religion node is connected to a square red tracker node if the tracker is present on the website. If a website does not have any trackers then it is represented by a circular node with no edges. On an average, the banking category has 3 trackers per website, and the religion category 13 trackers

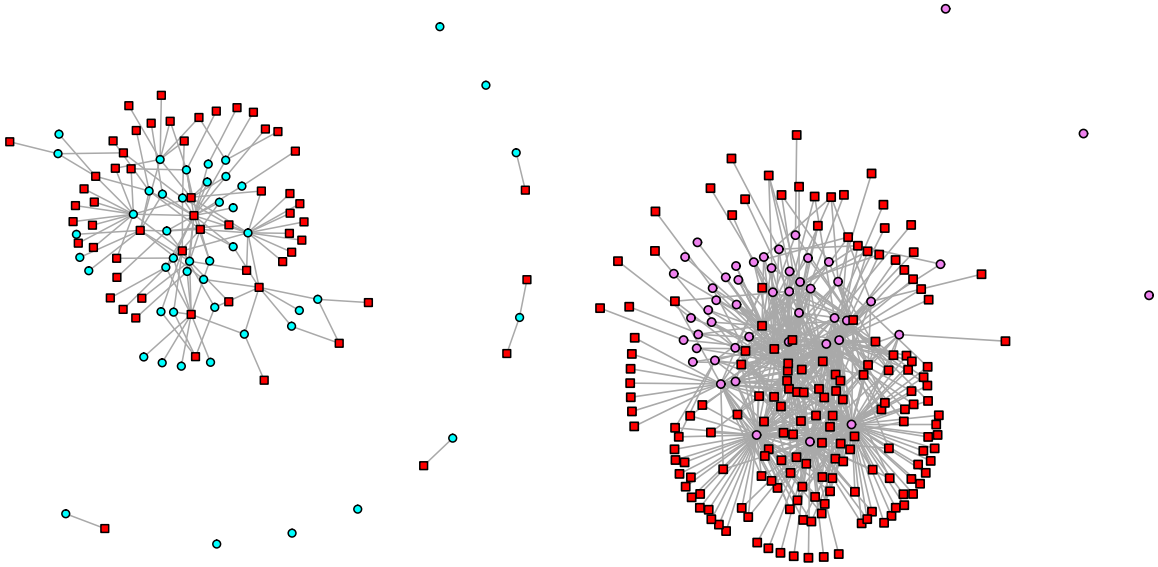


Figure 3.2: Two-mode network for Banking category (left) and Religion category (right).

per website. Figure 3.3 (left) shows a two-mode network that combines websites and trackers from both banking and religion categories. In this figure, we can see trackers that are present on both banking and religion websites. If a tracker is present on both banking and religion websites, it can combine user activities from the two categories.

We can do a network fold operation on a two-mode network to create a one-mode network to visualize how trackers connect websites in the banking category to websites in the religion category. The nodes in the one-mode network are banking and religion websites. If the same tracker is present on a website from a banking category and a website from a religion category, then there is an edge between the nodes representing these websites. Figure 3.3 shows a one-mode network where an edge between two nodes exists if the corresponding websites share at least five common trackers. For example, the edge between Regions bank (regions.com) and Christianity Today (christianitytoday.com) shows that there are at least five trackers that can link the activities of a user who visits both Regions bank website and Christianity Today website. Similarly, there are at least five trackers that link user activities between Regions bank and Jerusalem Post (jpost.com) and Regions bank and Catholic Online (catholic.org). If the trackers share data with the Regions bank, the bank could infer whether users who visit Regions website are Christian, Jewish or Catholic.

To improve visualization of the level of tracking, the size of the nodes correspond to the number of connections to other nodes in the network; the thickness of the edges correspond to the number of shared trackers between two nodes. For example, Bank Rate (bankrate.com) banking website is connected to 11 religion websites, and, hence,

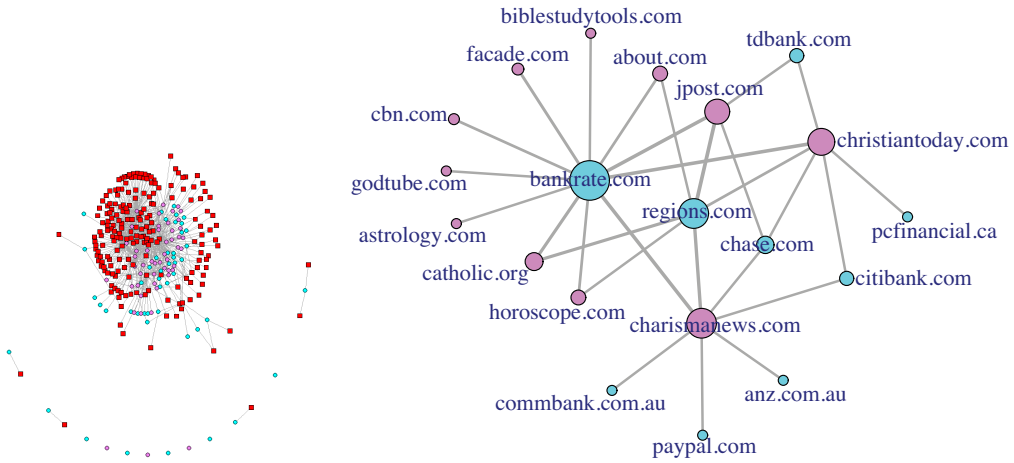


Figure 3.3: Undirected two-mode network combining Banking and Religion categories (left) and corresponding one-mode network with ≥5 trackers between Banking and Religion websites.

the size of the node for Bank Rate is bigger than the size of the node for Chase Bank (chase.com) connected to three religion websites.

Prevalence of Trackers in Website Categories Let us consider another undirected two-mode network consisting of website nodes and tracker nodes. The website nodes correspond to 500 most popular websites in the United States. The tracker nodes correspond to the trackers found on the 500 most popular websites. We used a tracker detection platform tool [8] to identify trackers on each website. As before, we obtained rank and category for each website from Alexa.com. The top 500 websites are split into 15 website categories such as Adult, Arts, Shopping etc. Table 3.1, provides details about categories and the number of websites in each category. In the two-mode network, for each website node, we added a category attribute that indicates the category of the website. For example, the category attribute value "Adult" indicates that the website is an adult website.

To identify prevalence of trackers by website category, we computed the average number of trackers per website for each website category. First, we computed the degree centrality for each website node in the two-mode network. In an undirected network, the degree centrality metric counts the number of edges for a node. In a two mode network of websites and trackers, degree centrality of website node indicates the number of trackers present on the website. We then computed the average degree centrality of all website nodes in a category. Figure 3.4 shows the average number of trackers for each website category. The News category has the highest average number of trackers per website, and the Adult category has the lowest. Trackers are most prevalent on News websites and least prevalent on Adult websites. Each website in the News category has

Table 3.1: Website categories and number of websites in each category.

Website Category	Number of Websites
Adult	8
Arts	63
Business	63
Computers	101
Games	9
Health	8
Home	24
Kids_and_Teens	1
News	35
Recreation	13
Reference	15
Science	2
Shopping	65
Society	11
Sports	14
Total	500

on an average 21.3 trackers, and each website in the Adult category has on an average 4.5 trackers.

Top Trackers in Website Categories To identify trackers that occur frequently in each website category, we computed degree centrality for each tracker in a category. In a two mode network of websites and trackers, degree centrality of tracker node indicates the number of websites on which the tracker is present. Table 3.2 shows the top five trackers in the website categories the News and Adult categories. The tracker Doubleclick.net is the most frequently occurring tracker in both News and Adult categories. It is present on 32 out of 35 news websites and on 5 out of 8 adult websites. It can link user activities from 32 news and 5 adult websites on which it is present. Scorecardresearch.com is another tracker which is present on both news and adult websites. It can combine activities from 31 news and 2 adult websites. Table 3.2 also shows 10 most frequently occurring trackers regardless of the website category. Doubleclick.net is present on 339 out of top 500 websites. The second most frequent tracker is Facebook.com, which occurs on 242 websites. It can combine users' social networking activity with activities on other website categories. However, it does not seem to be present on adult websites.

Extent of Tracking on Websites The number of trackers on a website may not show the extent of tracking on a website. For example, consider a website with only one

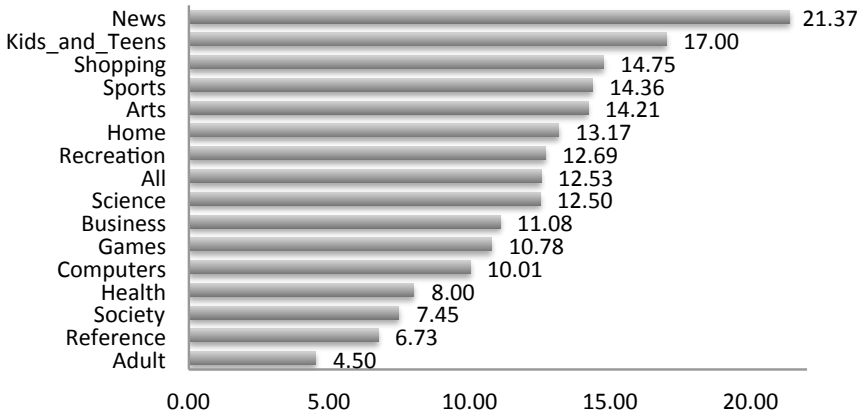


Figure 3.4: Average number of trackers per website in each category.

tracker Doubleclick.net. Because Doubleclick.net is also present on 338 other websites, the extent of tracking the website is high. In contrast, another website may have two trackers that do not occur as frequently as Doubleclick.net. Hence, the extent of tracking on that website with two trackers is lower than the extent of tracking on the website with only one tracker. To identify extent of tracking on websites, we can compute the bipartite projection of website to website in a two-mode network. We can then calculate the degree centrality of each website node in the resulting one-mode network. To understand how websites are linked within each category, we can perform bipartite projection operation on websites in a given category. The degree centrality of a website node in the resulting one-mode network represents the number of connections to other websites in the the same category. Table 3.3 shows five websites with the highest extent of tracking in the News and Adult categories. It also shows the top ten websites with the highest extent of tracking regardless of website category. In the News category, Huffingtonpost.com has 48 trackers and Examiner.com has 51 trackers. However, Huffingtonpost.com has 370 connections to other news websites and Examiner.com has 323 connections to other news websites. Hence, the extent of tracking on Huffingtonpost.com is potentially higher than Examiner.com although the former has fewer trackers than the latter. We can see similar trends in the Adult category. Both Redtube.com and Tubecup.com adult websites are connected to 9 other adult websites. However, Redtube.com has only 4 trackers compared to 12 trackers on Tubecup.com.

Linking of user Activities across Website Categories We can analyze whether trackers are more likely to link user activities from certain website categories. Using network analysis, we can identify patterns of linking among website categories. For example, if we identify a pattern where education websites are more often linked to social network websites than health websites, it may indicate that schools monitor their students'

Table 3.2: Top trackers in News, Adult and across all website categories.

Category	Tracker	Count
News	doubleclick.net	32
	scorecardresearch.com	31
	facebook.com	22
	imrworldwide.com	22
	google.com	17
Adult	doubleclick.net	5
	google.com	3
	google-analytics.com	3
	trafficjunky.net	3
	scorecardresearch.com	2
Across all categories	doubleclick.net	339
	facebook.com	242
	google.com	223
	scorecardresearch.com	211
	google-analytics.com	149
	twitter.com	148
	quantserve.com	137
	adnxs.com	136
	yahoo.com	113
	bluekai.com	99

social networking activities more than heath-related activities. Similarly, if network analysis reveals a pattern involving banking, health and insurance websites, it may indicate that banks and insurance companies closely monitor users' health activities.

To analyze patterns of closely linked website categories, we first perform a bipartite projection from website to website. We then aggregate all website nodes with the same category attribute into a single node. Figure 3.5 shows the resulting network for the 15 categories in the top 500 websites in the United States. In shows the aggregate number of connections from websites in a given website category to websites in other categories. To account for different number of websites in each category, we compute percentage of total connections from a category to itself and other categories. Figure 3.6 shows the details. The percentage of connections from News category to Sports and Arts categories are 13.3% and 13.4% respectively. The percentage of connections from News category to Adult, and Kids and Teens categories is 9.4% and 9.8% respectively. This may indicate that trackers are more likely to link user activities from News, Sports and Arts categories than News, Adult and Kids categories. To improve confidence in the identified patterns, we can further refine website categories into subcategories. We can also use advanced network group detection methods such as community detection.

Table 3.3: Extent of tracking on websites in News, Adult and across all website categories.

Website	Category	#Trackers on website	#Connections to other websites
huffingtonpost.com	News	48	370
examiner.com	News	51	323
washingtonpost.com	News	48	316
nytimes.com	News	36	313
drudgereport.com	News	32	298
redtube.com	Adult	4	9
tubecup.com	Adult	12	9
youporn.com	Adult	6	8
pornhub.com	Adult	2	6
sex.com	Adult	6	6
huffingtonpost.com	News	48	2969
free-tv-video-online.me	Computers	64	2955
kohls.com	Shopping	55	2846
wikia.com	Computers	43	2764
azlyrics.com	Arts	54	2685
examiner.com	News	51	2598
thekitchn.com	Home	50	2562
cars.com	Shopping	41	2506
evite.com	Computers	50	2468
washingtonpost.com	News	48	2393

Limitations First, it is possible that trackers on websites may change over time, and, hence, the network of trackers and websites may change. A longitudinal study of websites and trackers can address this issue at least partially. By analyzing the network of websites and trackers at different points in time, we can identify patterns that occur regularly. Second, if the categories have small number of websites e.g. Kids and Teens category in our analysis, the results from these categories may not be reliable. To avoid website categories with sparse number of websites, we could analyze all top level categories and their subcategories that have at least a certain number of websites. Third, all trackers may not be equally important from a privacy perspective. Different types of trackers can combine data from websites in different ways. Further, they may use the collected data for different purposes. Hence, we could further differentiate between types of trackers by adding them as an additional node attribute. We could also add the types of data collected by trackers as attributes. Four, our analysis of linking of activities across websites does not account for data sharing between trackers and websites that occurs offline. For example, two trackers DoubleClick, Google Analytics and Google.com are owned by the same parent company Alphabet, and the parent company can aggregate data collected by the two trackers. However, if we consider DoubleClick,

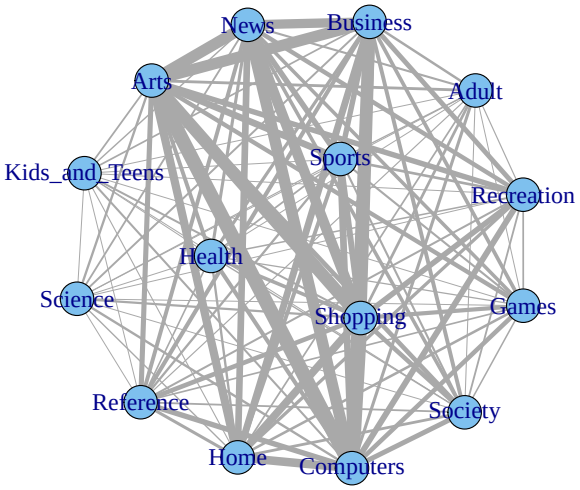


Figure 3.5: Aggregate connections among website categories.

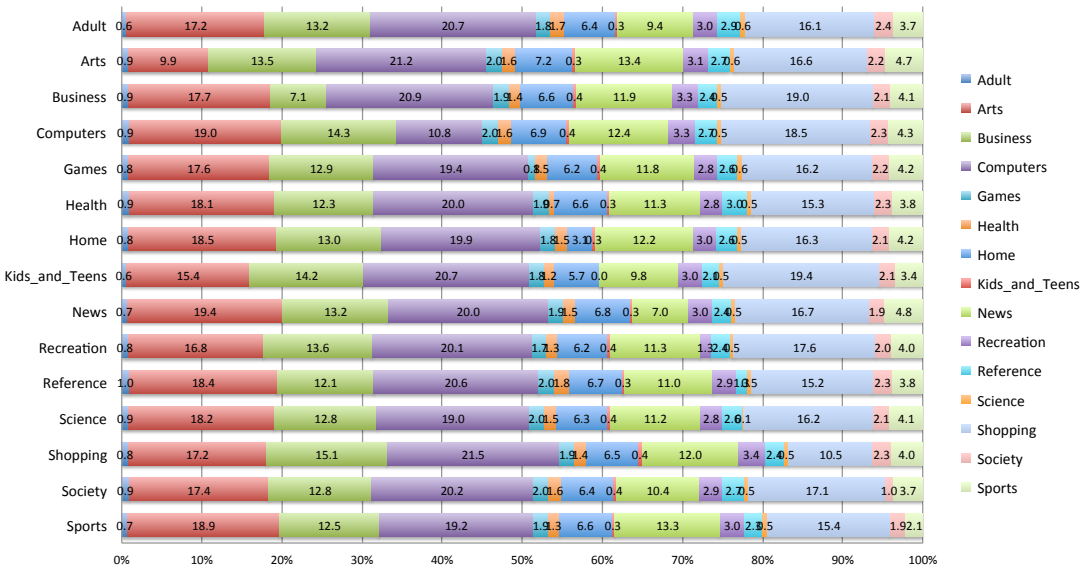


Figure 3.6: Percentage of connections from a website category to itself and to other website categories.

Google Analytics and Google.com as separate nodes, offline aggregation of data would not be visible from the network. Lastly, we could add demographic information such as age, income and education level of website users as attributes of website nodes. We could then analyze whether and how trackers target user demographics.

3.2 Impact of Tracker Data Practices

First, we discuss whether and how current tracking mechanisms can be used to siphon data from one country to another. We use the term "siphon" to indicate a one-way channel that once set up will result in a continuous flow of personal information from the source to the destination. We focus on the role of Internet tracking mechanisms in allowing Russia to siphon personal data from German users. Second, we discuss how trackers link user activities. We discuss how trackers link users' activities on adult websites with activities on other website categories. We also discuss how trackers link users' comments and discussions from different website categories.

3.2.1 Siphoning of User Data

Studying siphoning of user data is interesting for two reasons. First, in the current political climate, countries e.g. Russia have been allegedly collecting data about users from another country e.g. the United States for political gains. In this case, siphoning can facilitate data transfer to an adversarial country e.g. Russia. Second, data protection regulations in several countries e.g. the EU, Russia and China have imposed restrictions on flow of data across geopolitical borders. Siphoning could enable entities to transfer data across geopolitical boundaries violating data protection regulations.

Study Details To study the role of Internet tracking in allowing Russia to siphon personal data from German users, we analyze Russian trackers on German websites. We consider a tracker as a Russian tracker if it is either owned by a company that is primarily based in Russia or if the tracker sends data collected from the website to a server with a .ru domain.

To identify Russian trackers on a webpage, we used Ghostery browser extension and OpenWPM web measurement platform. Using OpenWPM platform we automated the process of visiting websites and collecting trackers on the websites. While visiting each website, Ghostery browser extension identified trackers on the website. For example, in Figure 3.7, Ghostery shows a Russian tracker Mail.Ru on a German news website SZ.de.

We studied four website datasets. The first data set consisted of 12 popular mainstream news websites, both national and local, in Germany (NewsDE) and was the primary data set that we analyzed. For each website in the NewsDE dataset, we collected

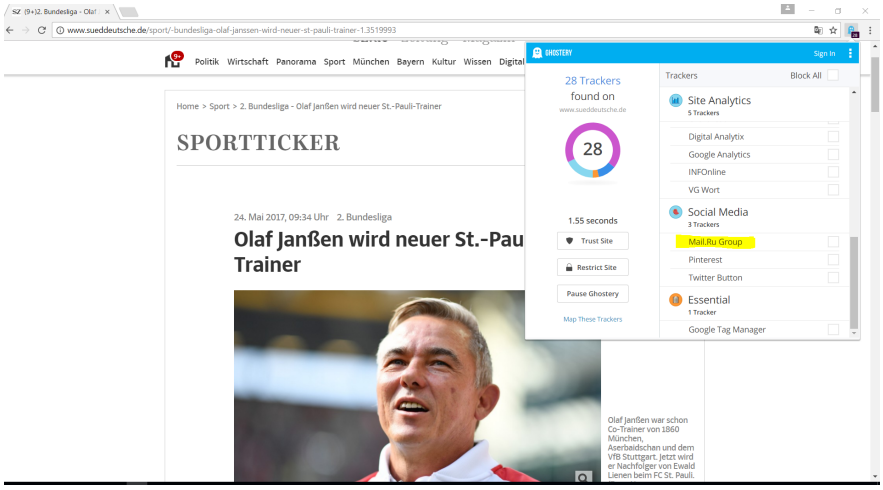


Figure 3.7: Russian tracker Mail.Ru on German news website SZ.de.

trackers on the top-level webpage and approximately 100 subpages. The OpenWPM platform first visited the top-level page and then randomly visited 100 subpages of the top-level page. We used the remaining three datasets for additional analyses. Two of the datasets contained 1000 most popular websites in Germany (TopDE) and Russia (TopRU) respectively. The last dataset contained one million most popular websites in the world (TopWW). For websites in TopDE, TopRU and TopWW datasets, we collected trackers on the top-level webpage only, but not any subpages. We determined website rank using Alexa traffic statistics.

Prevalence of Russian Trackers In the NewsDE dataset, we found Russian trackers on 10 out of 12 websites. There were five unique Russian trackers. We list the prevalence of these five Russian trackers in TopDE, TopRU and TopWW website datasets in Table 3.4. To identify the types of data collected by the Russian trackers, we analyzed the trackers' privacy policy and code. Two of the trackers, Mail.Ru and Segmento trackers do not have privacy policies. Hence, it is difficult to understand the types of data collected by them. Examining a tracker's code reveals information directly collected by the tracker, but not data collected indirectly via third-party companies. Privacy policies can disclose information about data collection from third-party companies. Further, they can disclose the purposes for which data is used, with whom data is shared and how long the data is retained. Although companies could engage in data practices not disclosed in their privacy policies, data protection regulations may force policies be accurate. The types of data directly collected by the five Russian trackers include IP address, website URL, cookies and time of accessing the website. IP address is considered personally identifiable information within European Union. Further, it is possible to map IP address to an individual user profile that includes personal information such as name and postal address [5].

Table 3.4: Prevalence of Russian trackers

	TopDE		TopRU		TopWW		
	.de	All	.ru	All	.de	.ru	All
Mail.Ru	0	48	148	198	10	6244	10004
AdRiver	8	29	126	165	97	3061	5814
Segmento	1	14	56	75	0	1073	1640
AdSniper	0	7	16	23	5	1190	2810
Facetz	0	4	23	36	13	2777	4613

Parameters of Tracking Patterns for Siphoning Data We identified several tracking patterns that can be used to siphon personal data from German Internet users. In each tracking pattern, there are two key parameters: distance to data and type of control. A tracking pattern with shorter distance to user data has better timeliness and can collect data of finer granularity with higher accuracy. We identified four components of a tracking pattern – website, website core package, tracker and data market place – that an entity can control. The type of component influences distance to user data, for example, a website interacts directly with users and, hence, is closest to user data. A data marketplace is farthest from user data because it does not allow direct interaction with users. A tracking pattern with shorter distance to user data is easier to identify and more visible to regulatory authorities.

The type of control parameter determines the level of control. Higher control over components of a tracking pattern e.g. website or tracker implies easier to access to data siphoned by the tracking pattern. Although all the five Russian trackers found on the NewsDE websites can siphon data, whether an entity such as the Russian government has access to siphoned data depends on the type of control the Russian government has over the trackers. In recent years, the Russian government has increased control through ownership of companies [83]. In addition, it has increased control via Internet data protection regulations and financial regulations [83]. We classify the way the Russian government exerts control over the Russian trackers into three categories: State-owned, State-proxy and State-regulated. Among the three categories, ownership provides highest control and regulations provide lowest control over data collected by the tracker. Trackers in the state-owned category are owned by the Russian government. One of the five Russian trackers Segmento (segmento.ru) belongs to the state-owned category because its owner is Sber Bank a Russian government bank. Trackers in the State-proxy category are not directly owned by the Russian government, but by people or companies that are closely associated with the Russian government. For example, one of the five trackers, Mail.Ru Group is owned by Alisher Usamov who belongs to President Putin's trusted network of people [83]. Since 2013, the Russian government has increased this type of soft control via a trusted network of people known as systema [83]. The Russian government in recent years has passed laws and regulations to nationalize and increase control over Russian Internet. Internet related companies

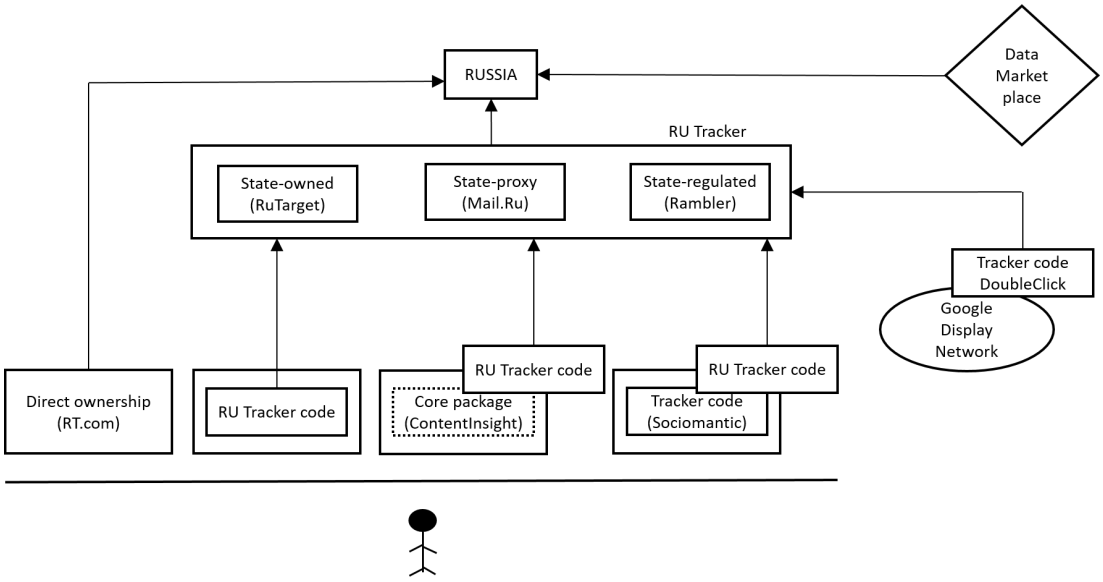


Figure 3.8: Tracking patterns for siphoning data.

based in Russia must adhere to these regulations and share data with the government at the government's request. Because of the difficulty in regulating companies provide services in Russia, but are owned by foreign-based companies e.g. Facebook and Google, the Russian government has passed financial regulations which put a cap of 20% on foreign investment in Russian Internet related businesses. We categorize trackers that are regulated by the Russian government under the State-regulated category. This category provides least control over data collected by the Russian trackers. Three out of the five trackers found on NewsDE websites belong to this category.

Tracking Patterns for Siphoning Data Below we describe tracking patterns that we identified. The following patterns are arranged in the order of increasing distance to user data. We show the patterns in Figure 3.8.

- *Direct ownership of website*: This enables minimizes distance to user data. When users directly interact with the website, it is possible to collect data in real-time and with maximum accuracy. Data collection is usually part of core functionality of the website and cannot be blocked by users. The owner of a website is susceptible to regulations including data protection regulations, which may require disclosing, usually via a website privacy policy, data collection, use and retention practices. A regulatory authority such as the EU Data Protection Agency may identify any violation of regulations and levy fines or other restrictions on the operation of the website.

  None of the mainstream news websites in NewsDE dataset are directly owned by the Russian government. However, alternative news websites such as RT Deutsch

(rtdeutsch.com) and SputnikNews (sputniknews.com) are funded by the Russian government and are popular within Germany. RT Deutsch and SputnikNews are ranked 167 and 233 respectively in Germany, and a recent study found that the percentage of German Internet news readers that visited these sites was 17% (4% frequently, 13% occasionally) and 15% (4% frequently, 11% occasionally) respectively. Because the Russian government funds these news websites, it may be able to collect personal data about German users who visit these websites.

- *Embed tracker directly within website*: Although an entity may not own a website, it can embed a tracker within the website. Compared to direct ownership of a website, this pattern increases distance to user data. However, due to the presence of two entities – website and third-party tracker – it is more difficult for a regulatory authority to attribute blame for violation of data protection regulations. The regulatory authority must decide whether to hold the website responsible or the third-party tracker responsible for data collected by the third-party tracker. Among the different ways to embed a tracker within a website, directly and statically embedding a tracker within a website minimizes distance to user data, but maximizes visibility to regulations. To embed a tracker directly within a website, the tracker owner must have influence over the website. A directly embedded tracker collects information every time a user visits the website and sends it to the entity that owns the tracker. However, compared to direct ownership pattern where data collection cannot be blocked, tracker blocking tools could block the tracker and prevent data collection.

  None of the websites in NewsDE dataset has a directly embedded Russian tracker. However, we found Russian trackers directly embedded within TopDE websites. One example is the Russian tracker LiveInternet, which was found on 46 TopDE websites, 450 TopRU websites and 34201 TopWW websites. In many instances, it is directly embedded within the website. LiveInternet collects IP address, URL, screen properties etc. when a user visits the website in which it is embedded. LiveInternet is owned by German Klimenko, Internet Adviser to Russian President Vladimir Putin, and, hence, we classify it as a State-proxy tracker. LiveInternet tracker has a large reach within Russian and is found on nearly 50% of TopRU websites. Comparatively, it is found only on 5% of TopDE websites and most of these websites are Russian domain (.ru) websites popular within Germany and none are German domain (.de) websites.

- *Embed tracker using a core package used by website*: Instead of embedding a tracker directly within a website, a tracker can be embedded in a package used by the website. A package is generally owned by a third-party company and the website uses it to implement some core functionality. In this pattern, the tracker owner needs to have influence over the package not on the website itself. Compared to directly embedding a tracker within a website, the additional level of indirection makes it is more difficult to identify and regulate the tracker. The news website Sueddeutsche Zeitung (sueddeutsche.de) in our German news

dataset uses the ContentInsight editorial analytics package that embeds a Russian tracker Mail.Ru Group. Sueddeutsche Zeitung is ranked 100 in Germany, and 48% (12% frequently, 36% occasionally) of German Internet news readers visit it. The ContentInsight package loads the Mail.Ru Group tracker frequently, but not always. The tracker collects URL, IP address, cookies etc.

- *Embed tracker using a tracker embedded within website*: A tracker embedded within a website can dynamically load additional trackers. Nine websites from the NewsDE dataset have embedded trackers that load Russian trackers. The Russian tracker AdRiver (adriver.ru) was found on seven news websites. One news website had a tracker Sociomantic (sociomantic.com) that loaded AdRiver onto the Bild (bild.de) news website. In case of the six other news websites, AdRiver was loaded by the tracker DoubleClick (doubleclick.net). Sociomantic is owned by a German company and DoubleClick is owned by Google, a company based in the United States. DoubleClick has a real-time bidding protocol where trackers can bid for getting loaded onto websites within the Google Display Network consisting of two million websites. As part of evaluating a bid, trackers can see the website URL or site id, truncated IP address and location of the website. If a tracker wins a bid, DoubleClick shares the website URL if not already shared and full IP address. Adriver for example won a bid for getting loaded onto Zeit (zeit.de) website.

- *Acquire data from third-party*: As discussed earlier, companies can buy personal data of users from third-party companies. They can buy completely identifiable profiles containing hundreds of attributes including name, postal address, email, mobile phone number, birth date etc. from data market places. The advantage of purchasing personal data is that it has least visibility and hence hard to identify and regulate. The disadvantage is that it provides lowest timeliness and accuracy of data; tracking entity cannot collect data in real-time, and data can be erroneous. Sometimes as high as 80% of the data in a profile may be incorrect [5].

  One website in our NewsDE dataset, Merkur (merkur.de) embeds a tracker called Disqus to track community discussions and comments on its website. To comment on an article on the Merkur website, a user must create an account with Disqus using his or her email address and name. When a user posts a comment, the Disqus tracker collects comments, name, email, IP address etc. Disqus tracker is present on 1.9% (18909) of top 1M websites worldwide including health, adult, religion, banking and insurance websites. Disqus creates individual user profiles by combining a user's data from the Merkur website with the same user's data from other website where the user posts comments. Further Disqus obtains additional personal data about the user from data market places and combines it with the user profile. Merkur and other companies, including companies in Russia, can purchase user profiles created by Disqus.

Impact of Siphoning Patterns Using the tracking patterns we identified, it is possible for a country e.g. Russia to siphon personal data of users in another country e.g. Germany by using existing Internet tracking mechanisms. Because of lack disclosure of data practices e.g. missing privacy policies, it is difficult to understand how the siphoned data is being used. For instance, the country that siphons data could use it to identify the news websites that citizens of another country visit, infer their political opinions from the articles they read, and show them ads influence and change their opinions. Tracking patterns provide trade-off between quality of data and visibility, and by using patterns with low visibility, it is possible to evade detection and regulation.

3.2.2 Linking of User Data

As we discussed earlier, network analysis can show how trackers link user activities from different categories. We analyzed two types of linking. First, we analyzed how trackers link comments users write on websites belonging to different categories. Second, we analyzed how user activities on adult websites are linked to user activities on websites in other categories.

Study Details We studied linking using two website datasets. The first data set contained one million most popular websites in the world (TopWW). The second data set contained 1000 most popular websites in the United States (TopUS). For each website in the two datasets, we collected trackers on the top-level webpage. To identify trackers on a webpage, we used Ghostery browser extension and OpenWPM web measurement platform. Using OpenWPM platform we automated the process of visiting websites and collecting trackers on the websites. While visiting each website, Ghostery browser extension identified trackers on the website. We determined website popularity using Alexa traffic statistics. We used Alexa website categories to classify websites into categories.

To identify a tracker as comment-related, we used Ghostery classification of trackers. It classified each tracker as follows: *Advertising*, *Pornvertising*, *Comments*, *Social_media*, *Site_analytics*, *Customer_interaction*, *Audio_video_player* or *Essential*. An Advertising tracker tracks user activities to deliver advertisements. A Pornvertising tracker tracks user activities on adult websites. A Comments tracker enables users to write comments on websites e.g. comment on a news article or product purchased. A *Social_media* tracker such as Facebook or Linkedin Widgets tracks user activities on social media sites and other websites. *Customer_interaction*, *Audio_video_player* and *Essential* trackers are required for interacting with website content.

Linking of User Comments User activities on websites can include writing comments e.g. users can write comments after reading a news article or purchasing a product. Users can also write reviews, participate in discussions etc. Companies can embed

Table 3.5: Comments trackers on top 1M websites in the world.

Comments Tracker	Count
LiveInternet	34201
Disqus	18909
Yotpo	2251
eKomi	998
Answers Cloud Service	589
Livefyre	499
HyperComments	291
GetSatisfaction	103
Unknown Advertisers	42
GetKudos	23

Table 3.6: Comments trackers on top 1K websites in the United States.

Comments Tracker	Count
Answers Cloud Service	24
Disqus	20
Livefyre	10
LiveInternet	6
GetSatisfaction	2
eKomi	1

Comments trackers on websites to track user comments, reviews and discussions. User comments/reviews are generally considered as public and companies can combine comments from multiple websites.

Table 3.5 shows the distribution of Comments trackers on websites in the TopWW dataset consisting of one million most popular websites in the world. LiveInternet tracker is the most common Comments tracker and is present on 34201 websites out of 1M websites. The second most common Comments tracker is Disqus, which is present on 18909 websites. Table 3.6 shows the distribution of Comments trackers on websites in the TopUS dataset consisting of one thousand most popular websites in the United States. Answers Cloud Service is the most common Comments tracker and is present on 24 websites. Disqus is the second most common Comments tracker and is present on 20 websites.

Let us analyze the impact of Disqus tracker, which is frequently found on websites both in the TopWW and TopUS datasets. Disqus tracker enables community discussion/comments on websites. As per Disqus privacy policy, it collects comments, users' personally identifiable information or PII (name, email, IP address etc.) and clickstream data (page view, mouse scroll, click) etc. It also collects PII from 3rd

Table 3.7: Website categories on which Disqus Comments tracker is present.

Website Category	Count
News and Media	8
Arts and Entertainment	7
Computer and Electronics	2
Career and Education	1
Internet and Telecom	1
Sports	1

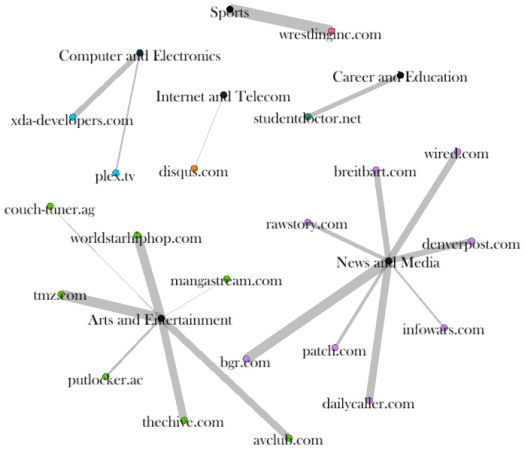


Figure 3.9: Two-mode network of website category and websites containing Disqus tracker.

party databases and combines them with data collected directly via Disqus tracker on websites.

Table 3.7 shows the categories of websites, from the TopUS dataset, on which Disqus tracker is present. Disqus is present on News and Media, Arts and Entertainment, Computer and Electronics, Carrer and Education, Internet and Telecom, and Sports website categories. Figure 3.9 shows a two-mode network of websites containing Disqus tracker and the category of the website. The website category nodes are colored black. For example, it shows "News and Media" node connected to eight websites in that website category. similarly, it shows seven websites linked to the "Arts and Entertainment" website category node. Although the website StudentDoctor.net is classified as "Career and Education," it can also be considered a health-related website.

In the TopWW website dataset, Disqus is present on website categories related to health (e.g. 4healthresults.com, DiyHealth.com), adult (e.g. FreePornaz.com, GayAsian-Porn.biz), religion (e.g. Christianstt.com, CatholicStand.com), banking (e.g. Bankgid.com,

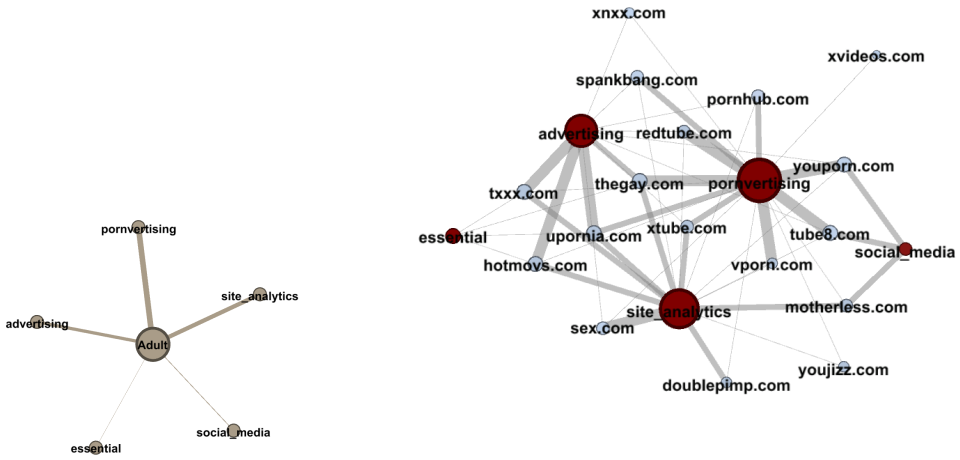


Figure 3.10: Network showing five types of trackers on Adult websites (left). Network showing connections between different types of trackers and adult websites (right).

InvestmentBank.com), insurance (FreewayInsurance.com, InsuranceHotline.com etc.) etc. Disqus can combine user comments from these website categories. Since Disqus collects user PII, it can identify and combine comments of individual users. Users may consider their comments on websites categories such as health sensitive private information. They may not want it combined with their comments on other website categories.

Linking of Activities on Adult Websites Among the top 1K most popular websites in the United States, 17 websites contain adult content. Figure 3.10 (left) shows that the 17 adult websites contain five types of trackers: Pronvertising, Social_media, Advertising, Site_analytics and Essential. Hence, trackers can collect user activities on adult websites and use it for advertising purposes. They can also combine it with users' social networking activities.

Figure 3.10 (right) shows a network consisting of two types of nodes: the type of tracker (red) and adult websites (blue). The size of the red nodes is based on the number of trackers of a given type. Pornvertising trackers occur most frequently (6) followed by Advertising trackers (5) and Site_analytics (4) trackers. There are two Social_media trackers related to Twitter social networking site.

Figure 3.11 shows connections between trackers and adult websites. Although Pornvertising trackers (red nodes) occur most frequently, individual trackers do not appear on many adult websites. Advertising trackers such as DoubleClick (green nodes) are present on a larger number of adult websites than popular Pornvertising trackers TrafficJunky and Exoclick. DoubleClick may show advertisements on based on adult

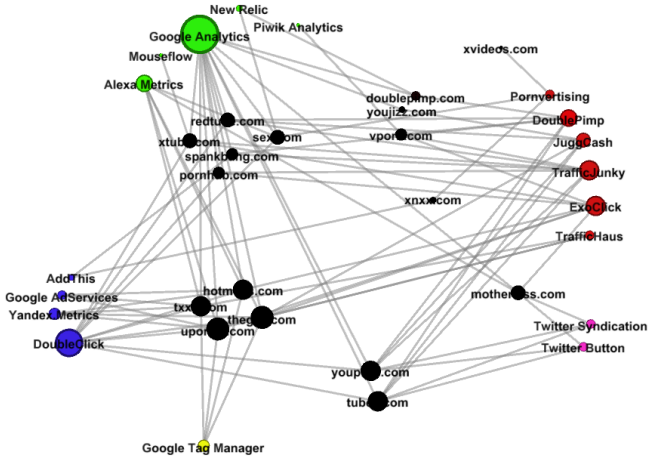


Figure 3.11: Network showing connections between trackers and adult websites.

website activities on other websites. Similarly, it is possible that Twitter Social_media tracker can show ads on Twitter website based on activities on adult websites. People may consider their activities on adult websites as sensitive private data, and they may not like advertisements based their porn related activities to show up on social media or other websites.

3.3 Summary

We discussed how we can use network analysis to identify tracker data practices. We applied network analysis to a two-mode network consisting of website and tracker nodes. We showed how to identify linking of user activities across website categories, prevalence of trackers in website categories, extent of tracking on websites and patterns of linking across website categories.

We studied the impact of linking of user data on user privacy. Our analysis showed how trackers link user comments and discussions from different website categories such as health, banking, religion and adult. It also showed how user activities on adult websites can be linked to user activities on websites in other categories such as social media.

We studied the impact of tracker data practices on siphoning of user personal data. Analyses showed that Internet tracking mechanisms can facilitate siphoning of personal data across borders while evading data protection regulations. We identified six tracking patterns that Russian trackers use to siphon data from German Internet users. We found that two key parameters, distance to data and type of control, determine timeliness, accuracy and granularity of siphoned data.

4 Privacy Impact of Aggregator Data Practices

The online services landscape is driven by a data economy in which data aggregators and service providers trade user information on data marketplaces [26, 84]. As part of the data economy, data aggregators and service providers collect extensive amount of data about individuals from multiple sources, including public, online and offline sources [26]. By combining information from multiple sources, they create behavioral profiles of individuals. The data and profiles may be used for purposes such as personalization, risk mitigation products, people search and targeted advertising [26]. The data economy benefits users by providing better products and services. It also sustains many free services such as search and social networking. However, the data economy also raises privacy concerns. For example, studies have found that users have privacy concerns when behavioral profiles are used for advertising [27, 85, 86, 28]. Using profile data for risk mitigation services such as background checks raises concerns about accuracy of data [26].

We propose a novel approach for identifying aggregator data practices [5]. We identify the types of data that aggregators collect about users by examining behavioral profiles of users and user data sold on data marketplaces. The United States Federal Trade Commission has investigated the types of data that companies may potentially use to build behavioral profiles. However, they did not look at contents of actual behavioral profiles [26]. We examine behavioral profiles of users via mechanisms that allow users to access their own behavioral profiles [87, 88, 89]. Companies may allow users to access to behavioral profiles to increase transparency of their data collection practices. We identify data sold on marketplaces from documents published by data aggregators and service providers.

We study the impact of aggregator data practices [5]. First, we study user concerns regarding data in their behavioral profiles. Prior studies have focused on user concerns and perceptions regarding use of behavioral profiles for advertising [27, 28]. We focus on user privacy concerns regarding actual contents of behavioral profiles. Our approach of using user's own behavioral profile for eliciting concerns and surprises leads to a more contextualized and nuanced understanding of user concerns regarding online behavioral profiles. Second, we estimate the extent of errors in user behavioral profiles. Third, we evaluate the extent of transparency provided by profile access mechanisms by comparing

Parts of this chapter were previously published by ASE [5].

data shown in actual profiles with user data sold on data market places. Lastly, we identify usability issues of profile access mechanisms.

4.1 Identifying Aggregator Data Practices

Figure 4.1 shows a simple conceptual model of the data economy that highlights the role of data aggregators. Users provide their personal data to public and private sector service providers when they receive products and services from these service providers. We group all entities such as websites, offline stores, advertisers and marketers under the umbrella of private sector service providers. Data aggregators collect different types of user data available from service providers and also via direct engagement with users. Private sources of information include offline and online surveys, in-store and online transactions, website and forum interactions, and social networking activity [90]. Technical measures such as browser cookies, flash cookies and Javascript enable can be used to collect user data [91, 20, 7, 8] from online interactions. Public sources of information include census data, voter registration databases, occupation data from state license boards, bankruptcy records, county deed and tax assessor records and Yellow-pages directories [90]. Data aggregators combine the data obtained from these sources and build behavioral profiles of individual users. The data and behavioral profiles are traded on data marketplaces [84]. Service providers can purchase data and profiles, and use it to enrich their knowledge about their customers, which may help them to improve their services.

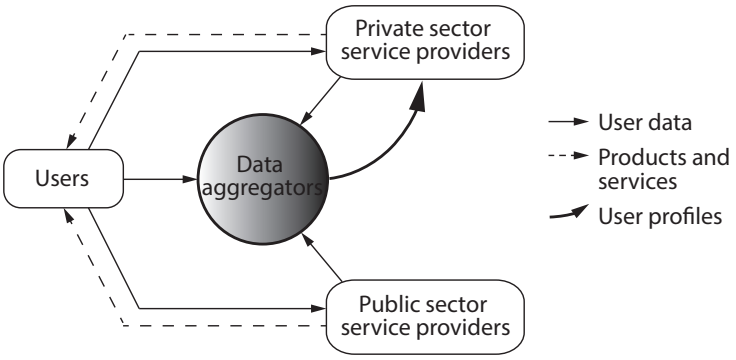


Figure 4.1: A conceptual model of the data economy.

We propose a novel approach for identifying aggregator data practices [5]. We identify the types of data that aggregators collect about users by examining behavioral profiles of users and user data sold on data marketplaces. The United States Federal Trade Commission, an agency in charge of protecting consumer privacy, investigated the activities of data aggregators (brokers) [26]. They identify how data is acquired via various data sources and collection techniques, types of data collected, potential uses

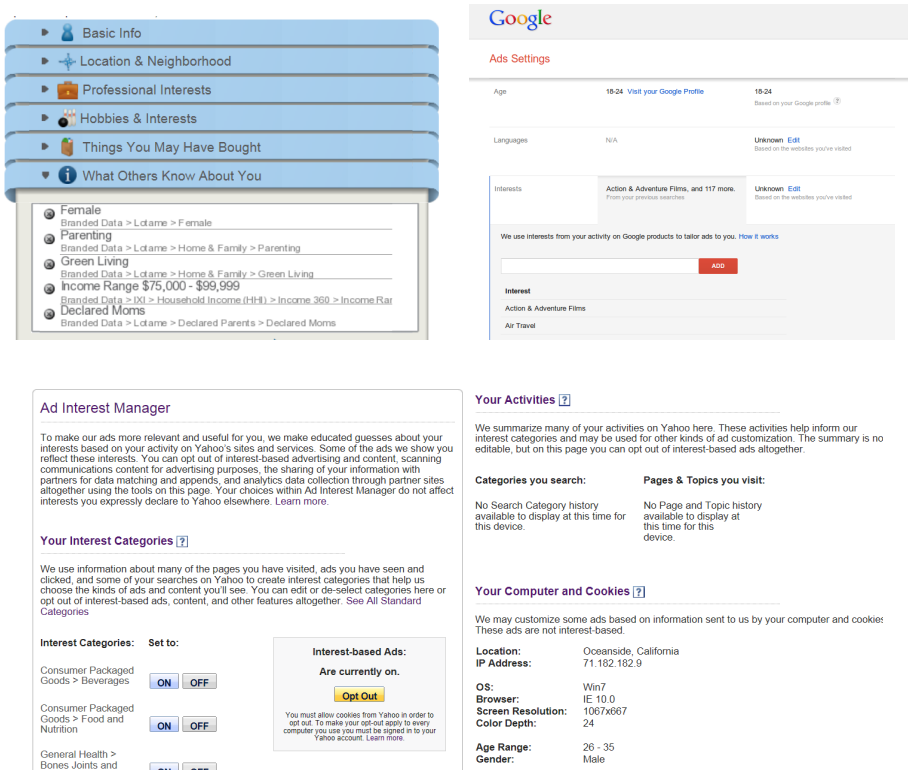


Figure 4.2: Sample profiles: BlueKai Registry (top left), Google Ad Settings (top right), and Yahoo Ad Interests (bottom)

and steps taken to maintain data accuracy. However, they do not investigate contents of actual user profiles.

We examine behavioral profiles of users via mechanisms that allow users to access their own behavioral profiles [87, 88, 89]. To increase transparency, some companies allow users to access their behavioral profiles. Companies may choose to show only some of the data that they have collected about the user [92]. In addition to looking at their data, users may be able to edit data in their profiles. Companies may use client-side or server-side validation to provide access to user profiles. For example, BlueKai [87], Google [88] and Yahoo [89] provide access to profiles based on browser cookies. Companies such as Acxiom [92] and Microsoft [93] require that users create an account with them and sign-in to access their profiles. To create an account user may have to provide details such as email address and name. Additionally, companies may request information such as full legal name, full address (street, city, state and Zip code), date of birth and last four digits of social security number to verify the identity of a user [92]. Figure 4.2 shows examples of BlueKai Registry, Google Ad Settings and Yahoo Ad Interests profiles.

4.1.1 User Behavioral Profile Analysis

We discuss the process of identifying aggregator data practices from user behavioral profiles. We study behavioral profiles from three companies: BlueKai Registry, Google Ad Settings, and Yahoo Ad Interests (see Figure 4.2). We analyze data from eight behavioral profiles ($N = 8$) including five BlueKai Registry profiles, two Google Ad Settings profiles, and one Yahoo Ad Interests profile. These are cookie-based profiles that do not require users to create accounts or sign-in on data aggregator websites. We think participants will find cookie-based profiles easier to access. We expect our results to be representative because these companies cover large number of users and the profiles contain data from diverse sources. Data in the BlueKai Registry profiles comes from nearly 30 third-party companies that participate in the BlueKai Audience Data Marketplace [84]. Currently the marketplace is the world's largest third-party data marketplace providing data on 300 million users or approximately 80% of the US population. Google Ad Settings displays interests and other information inferred from user activities on Google and more than one million partner sites [88]. Yahoo Ad Interests shows data inferred from Yahoo's sites and services [89].

Collection of Profiles

We conducted semi-structured in-person interviews with eight participants. We explained to the participants that companies may collect data about them, and may create behavioral profiles. We informed the participants that they might be able to access their profiles. We requested them to look at their profiles (BlueKai, Google and/or Yahoo), and if they felt comfortable, share information in their profile with us. Our participant pool included graduate students with engineering and/or science background. Only one participant was aware that he could access profiles created by companies. Six participants had never deleted cookies from their browser, one deleted cookies selectively and one regularly deleted cookies. From our eight participants, we collected information on eight profiles including five BlueKai Registry profiles, two Google Ad Settings profiles, and one Yahoo Ad Interests profile. All of the participants tried to access their BlueKai profile. Six were able to access their Bluekai profile, but two were not able to access it as they were using cookie blocking and/or script blocking. Of the six participants who accessed their Bluekai profile, five shared profile information with us. Three of our participants looked at an additional profile; two looked at Google profile and one looked at Yahoo profile, and they shared profile information with us.

Analysis of Profiles

We analyzed the eight profiles from Bluekai, Google and Yahoo for different types of data. We computed the size or the total number of data items in each profile. The two Google profiles contained $\simeq$120 items, the Yahoo profile had $\simeq$25 items, one Bluekai

profile had ≃10 items, two BlueKai profiles had ≃30 items and two BlueKai profiles had ≃570 items. Based on the number of items, we can say that we have four small-sized profiles, two medium-sized profiles and two large-sized profiles.

We organized the data from the profiles ($N = 8$) into seven categories: demographic, geographic, technical, predictive, psychographic, behavior and life event. We based our categories on the categories commonly used in data marketplaces and privacy policies to describe different types of data. We tried to choose distinct, non-overlapping categories, so that each data item fell into only one category. However, for some data items, it was difficult to choose a single category. For example, it was difficult to decide whether "Credit Card Holder" belonged to individual demographic or behavioral data category.

A challenge during analysis was to comprehend profile items. For example, the meaning of "Demographic > High Confidence" and "Credit Card Interest Score" was not clear. Does "High Confidence" imply that the user has high confidence or that the company has high confidence in the accuracy of demographic data? Does "Interest Score" mean how much interest a user is paying or how much he is interested in getting a new card? We could resolve some of the ambiguities by reading several documents published by data companies. For example, we could resolve "High Confidence" as implying data accuracy, but could not disambiguate the meaning of "Interest Score."

Geographic category was present in Yahoo and BlueKai profiles. Only Yahoo profile contained technical category. All three profiles showed individual demographic data regarding gender and age. However, BlueKai profile contained additional individual demographic data including marital status, education and occupation. It also contained demographic data related to user's household and work. The remaining categories appeared only in the BlueKai profiles.

Note that if a profile from a company does not show a certain category, it does not imply that the company does not have such information; a company may choose not to show some of the categories. Yahoo, for example, states on its Ad Interests Manager page "In addition to the information shown here, Yahoo! may use [...] information provided by partners to help customize some of the ads [...] [89]." Further, "Yahoo! may combine information, including personally identifiable information, that we have about you with information we obtain from our trusted partners," and BlueKai is one of its trusted partners [94]. In terms of improving transparency by allowing users to see the data in their profiles, BlueKai profiles are better than Google and Yahoo profiles because they provide more detailed information.

4.1.2 Aggregator Data Practices

We describe the seven categories of data types we found in actual behavioral profiles. We provide a summary with examples in Table 4.1. For demographic, geographic, technical and life event categories, we describe all the data types we found. However,

for psychographic, behavioral and predictive categories, the number of data types that we found are many, and, hence, we discuss representative examples. Further, for each category, we contrast what we found with the data that we may find if we examine more profiles.

We studied profile contents from relatively small number of profiles ($N = 8$). We looked at behavioral profiles from three data aggregators, and all of them were cookie-based profiles. If we study larger number of profiles, profiles from other companies, or server-based profiles, we may find other types of data.

Demographic Data Demographic data contains individual, household and firmographic subcategories. Companies associate individual's full name, full postal address, mobile number, email address and email activity date with both demographic and other categories discussed below [95].

- *Individual demographic*: We found gender, age (e.g. 20-24 years), marital status, education level (e.g. Some College), occupation (e.g. IT Professional), voter indicator, parent (e.g. Declared Mom), home ownership (e.g. Home owner or Renter) and languages. We found age, but companies also have date of birth [96]. In addition to voting, they have party affiliation (e.g. Democrat) and political donor (e.g. Contribute conservative) data [97, 96]. Other information include religious affiliation (e.g. Hindu), race/ethnicity (e.g. Arabs), family position (e.g. Female head of household) and summarized credit statistics including wealth rating (e.g. Decile), credit rating (e.g. High) and net worth [97, 90, 96].

- *Household demographic*: It includes details of an individual's household. We found household income (e.g. $20K-$30K), household size (e.g. 1), number of adults (e.g. 1), children in household (e.g. No), home type (e.g. Multifamily Dwelling), median home value (e.g. $0-$100K), length of residence (e.g. Less than 3 years), discretionary spending (e.g. $40K-$50K) and auto (e.g. Less than $20K). In individual demographic, we did not find individual income, but when household size or number of adults is one, then household income implies an individual's income.

  For household demographic, companies have rich set of additional attributes. In addition to knowing presence of children in a household, they know number of children, their gender and age, which can be a range (e.g. 0-3 years, 4-7 years) or month, day and year of birth [90, 96, 97]. They have indicators for the types of persons in a household, for example, presence of smoker, veteran in household, elderly parent in household [90]. Further, they have data about mortgage and refinance (amount, term, loan type, rate type) [97].

- *Firmographic*: It generally includes details about an individual's profession and organization he is affiliated with. We found type of industry (e.g. College and Universities), number of employees (e.g. 1-20 employees), and characteristics

Table 4.1: Examples of Data Types Found in User Profiles

Category	Subcategory	Examples
Demographic	Individual	Female
		Single
		20-24 years
		Some College
		IT Professional
		Voter
	Household	Income Range – $20K-$30K
		Household Size – 1
		Children in Residence - No
		Home Type – Multifamily Dwelling
		Home Value – $0-$100K
		Length of Residence – Less than 3 years
		Discretionary Spending $40,000-$49,999
		Auto – less than $20K
	Firmographic	Business Data > Micro (1-20 employees)
		Business Data > Software
Geographic		US > Pennsylvania > Pittsburgh
		Oceanside, California
Technical		IP address – 71.182.182.9
		OS – Win7
		Browser – IE10
		Screen resolution – 1067X667
Predictive		Credit Card Interest Score – 16-17%
		Credit Card App Intent Score – 10-11%
		Auto insurance online buyer – High Propensity
		Online Higher Education Enrollee – High Propensity
		In-Market – Cell-Phones and Plans
Psychographic	Interests	Health > Bones, Joints, Muscles > Pain
		Interest in Religion – Value Tiers 1-3
		Sweepstakes – Value Tiers 1-3
		Weight Conscious Code - Value Tiers 1-3
		Video Games – Ninetendo 3DS
		Travel Destinations > New York
	Attitudes	Buy American – Not Likely
		Show me the Money – Most Likely
		Aspirational Fusion - Hope for Tomorrow
Behavior	Activities	OTC Medicine > Pain Reliever
		Gastrointestinal – Tablets
		Offline CPG Purchasers > Brand > Hebrew National
		Charmin Ultra Soft
		Past purchase > ISP > Internet > Verizon
	Lifestyle	Green Living
		Owns a Regular Amex Card
		Eco Friendly Vehicle Owner
		Discount Shopper
		Prepaid Wireless Plan Subscriber
		Premium Channel Viewer
Life Event		Empty Nesters

of the profession (e.g. High Net Worth) and position (e.g. Technical Business Decision Maker). Additionally, companies have data about sales revenue, years of establishment (e.g. <2 years), domain expertise and seniority [96].

Geographic Data Geographic data includes location and neighborhood of a user. For example, we found "US > Pennsylvania > Pittsburgh," "US>Massachusetts>Boston-Cambridge-Quincy" and "Oceanside, California" for a participant that currently lived in Pittsburgh and Boston, and had lived in Oceanside about five years ago. The smallest granularity we found was at the city/county level. However, companies have geographical data at the level of full postal address, Zip code +4 (block level) and Zip code [96, 98]. For example, one company from BlueKai Marketplace claims to have 208 million postal addresses [98], and 72 million postal addresses linked to email addresses [96]. Each postal record is linked to a consumer's demographic, interests and behavioral data.

Technical Data Technical data generally includes information related to users' computers and devices used to access the Internet. We found IP address (e.g. 71.182.182.9), operating system (e.g. Windows 7), browser (e.g. IE 10), color depth and screen resolution. Interestingly, companies may use IP address to identify an anonymous consumer visiting a website in real-time. For example, they can map an IP address to a consumer's full name, full postal address, mobile number, purchases, interests and $\simeq$260 more attributes [95]. They also use IP address to infer location, for example, Yahoo states, "We use the IP address to infer your location [...]"

We did not find technical data regarding browser cookies and online activities and interactions, for example, search history, websites visited, articles read, comments, ratings and uploaded files. However, companies collect such information to derive psychographic, behavioral and predictive data. They use browser cookies to identify a website visitor's gender, presence of children (Yes or No), age (e.g. 20-29) and household income (e.g. $75,000-99,999$) [99].

Predictive Data Companies generally employ proprietary models and algorithms that combine data from multiple public, proprietary and self-reported sources, both online and offline, to make predictions about users. Predictions can be made about behavior, attitude, interest etc. For our predictive data category, we consider data that indicates user's intent to purchase, usually in the near future. We discuss other types of predictions as part of other categories discussed below.

We found examples that predicted purchases related to credit card, personal health, higher education, computers, cell phones, auto insurance, flying, hotels etc. For example, "Credit Card App Intent Score – 10-11%" indicates intent to apply for a credit card. "Personal Health – Values 70-90%" indicates future purchase propensity regarding personal health products; "In-Market – Cell-Phones and Plans" and "In-Market –

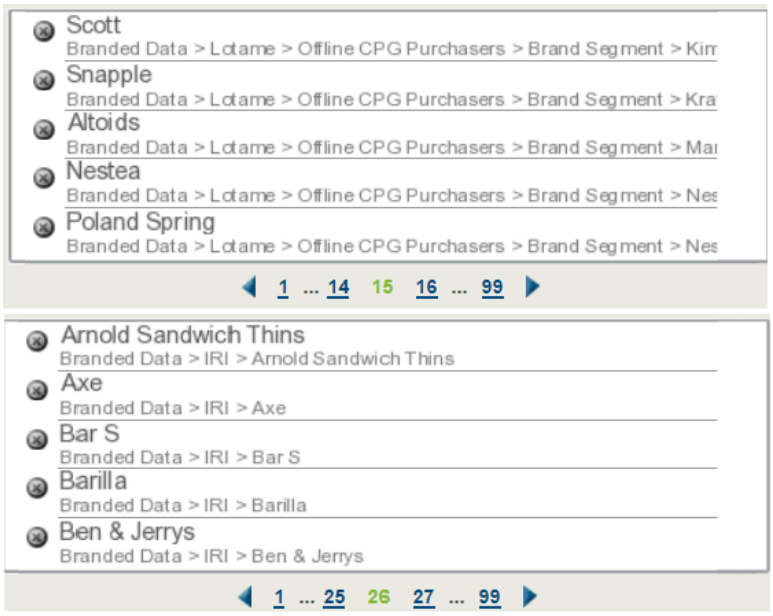


Figure 4.3: Listing of Consumer Packaged Goods in a profile

US Domestic Flyers" indicate that the user is currently shopping for cellphone plans and flights; "Auto insurance online buyer – High Propensity" and "Online Higher Education Enrollee – High Propensity" indicate users looking to buy insurance or enroll in courses. Companies have in-market data for many other areas including real estate, apartments and automotive purchases [96].

Psychographic Data Psychographic data generally includes interests and attitudes of a user. We found interests related to health (e.g. Bones, Joints, Muscles > Pain, Weight Conscious Code – Value Tiers 1-3), religion (e.g. Interest in Religion Code – Value Tiers 1-3, Christian Music Code – Value Tiers 1-3), travel (e.g. Destinations > New York, Vacation Packages), automotive (e.g. Coupe), sweepstakes, news (e.g. News and Politics > Government) etc. Companies possess additional data including gambling, lottery, alcohol and tobacco [90].

Profiles can include data on attitudes and values of users. Companies can use that information to trigger desired response from users. Some of the attitudes we found are as follows. "Buy American – Most Likely," which may indicate relatively high importance of pride in decision making. "Work Hard, Play Hard – Not Likely," which may indicate users' desire to be at the forefront of both their career and outside relative to their peers. "Stop and Smell the Roses – Most Likely," which may indicate a belief in altruism.

Behavioral Data Behavior data contains data related to users' lifestyle, activities and personality. For example, the entry, "Green Living," found in one of the profiles indicates that the user exhibits environmentally friendly lifestyle. Companies can further differentiate between users that act and those who only think (e.g. Behavioral Greens vs. Think Green), and between undecided and those who are against (e.g. Potential Green vs. True Brown) [90]. The profile containing "Green Living" also contained "Eco Friendly Vehicle Owner." Other lifestyle aspects we found include credit (e.g. Owns a Regular Amex Card), finance (e.g. Owns Mutual Funds), shopping (e.g. Discount Shopper) and travel (e.g. Theme Park Visitor).

In activities, we consider past purchases, both offline, such as stores and pharmacies, and online. One participant had over-the-counter medications (e.g. OTC Medicine > Pain Reliever, OTC Medicine > Cough and Cold) purchased at a local pharmacy listed in her profile. In addition to OTC, companies have information about medications (e.g. oral contraceptive, Lipitor, Insulin) purchased by users and any ailments they may have (e.g. Alzheimer's, clinical depression, Diabetes-2) [90].

Another participant had a list of ≃300 past consumer packaged goods (CPG) purchases in his profile. We list part of his profile in Figure 4.3. Figure 4.3 shows data collected by two aggregators Lotame (upper) and IRI (bottom). The figure shows past purchases such as iced tea, spring water, pasta, ice cream, deodorant etc. CPG entries can include brand (e.g. Ben & Jerrys, Hebrew National, Häagen-Dazs) and items (e.g. Nestea, Charmin Ultra Soft, Gastrointestinal – Tablets, General Mills > Fiber One).

Companies have other data such as purchase of alcohol and tobacco [90]. Lastly, companies have built models to predict an individual's personality type (e.g. introvert, leader). They have assigned personality types, by name and postal address, to 85% of the US adult population [100, 96].

Life Event Data Life event data indicates certain events in a user's life that may lead to changes in behavior and/or create specific needs. We found "Empty Nester," which may indicate that the user's children have left for college. Other life events that companies focus on include new movers and new parents [96, 90].

4.2 Impact of Aggregator Data Practices

To understand the impact of aggregator data practices, we elicited users' surprises and concerns regarding the data in their behavioral profiles. We also conducted an online survey to with a sample profile that we designed based on data found in user behavioral profiles. Prior research has studied user understanding, perceptions and concerns of targeted advertising, which uses behavioral profiles to personalize ads. Turow et al. surveyed Americans' attitudes towards targeted advertising that used data collected

from online websites and offline stores [21]. They used telephone interviews and closed-ended questions to understand attitudes of a representative sample of the US adult population. McDonald and Cranor studied users' understanding about targeted advertising including technical mechanisms such as cookies used for targeted advertising, and user concerns regarding targeted advertising [85]. Ur et al. studied user beliefs, attitudes and concerns regarding targeted advertising using semi-structured interviews [27]. Agarwal et al. studied users concerns regarding embarrassing and suggestive ads that may arise out of targeted advertising [28]. Gomez et al. studied user concerns regarding advertiser data practices by looking at three sources of information: consumer complains to the US Federal Trade Commission and other organizations, results from user surveys regarding privacy, and published news media articles [15]. These studies have not investigated privacy concerns regarding actual contents of behavioral profiles, and further, they have not employed user's own behavioral profile.

4.2.1 Study Details

We conducted semi-structured interviews eight participants ($N = 8$). We explained to the participants that companies may collect data about them, and may create behavioral profiles. We informed the participants that they might be able to access their profiles. We requested them to look at their profiles (BlueKai, Google and/or Yahoo), and if they felt comfortable, share information in their profile with us. We asked them to voice any concerns, surprises or questions regarding the data in their profiles. Our participant pool included graduate students with engineering and/or science background. Only one participant was aware that he could access profiles created by companies. Six participants had never deleted cookies from their browser, one deleted cookies selectively, and one regularly deleted cookies.

We conducted an online survey ($N = 100$) to validate the identified concerns with a larger and more diverse population. This survey had two purposes. First, we wanted to confirm whether a more diverse population of users agreed with the concerns that we had identified from the interviews. Second, we wanted to identify potential additional user concerns and data types that may not have been observed in the interviews. We recruited survey participants from Amazon Mechanical Turk crowd-sourcing platform [70]. We provide the survey questionnaire in Appendix A.

Survey Design To understand participant demographic, we asked them their age, gender, primary occupation and education level. To understand their technical background, we asked them whether they had a college degree or work experience in computer science, software development, web development or similar computer-related fields. We also asked them how much they liked personalization of ads on websites. We gathered information on demographic, background and liking for personalization as they may affect participant concerns. We also used demographic data to analyze diversity of our participant population.

Figure 4.4: Sample profile used in online survey.

We used a sample profile shown in Figure 4.4 to understand whether the survey participants agreed with the concerns that we identified from the interviews. We used the sample profile to understand their concerns regarding collection of sensitive data, amount of data, combining data from multiple sources, level of detail and data use. We felt that survey participants could not provide meaningful answers regarding concerns of accuracy of information and editing profile data based only on a sample profile. Hence, we did not ask them about those concerns.

We created the sample profile using data from profiles of the interview participants. To understand concerns about sensitive data collection, we added items related to credit (Credit Card Interest Score 8-9%) and health (Personal Health – Values 70-90%) both of which our interview participants had found sensitive. We also added entries related to religion (Interest in Religion Code – Value Tiers 1-3), individual demographic (Female and Declared Mom) and household demographic (Income Range $75K-$99K). To address the concern on amount of data, we ensured that the profile had data items from several categories: demographic, psychographic, behavior and predictive. Geographic category was represented by the "Location and Neighborhood" tab. To show data being combined from multiple sources, we added an offline CPG purchase (Offline CPG Purchasers > Vicks). To cover concern about level of details, we picked items that were very specific "Interest > Video Games > Sony > PlayStation 3." Further, the predictive values such as "Values Tiers 1-3" also increased the specificity of items. Lastly, we felt that it would be more realistic to show the data items as they appeared in actual profiles; a user looking at her actual profile would not have additional explanations or links to documents that could clarify her ambiguities.

Before showing the sample profile, we explained to the participants that advertisers collected data about them in order to personalize ads. Further, advertisers may create profiles about them using the collected data. We then showed them a sample profile (Fig. 4.4). To check whether participants were paying attention, we asked them to

select, from a list of six items, at least two items present in the sample profile. We then asked the participants to rate, on a 5-point Likert scale of "Strongly disagree" to "Strongly agree," how much they **agreed or disagreed** with the following list of concerns. We randomized the order in which the concern-related statements were displayed.

1. I am concerned because I believe that the profile contains sensitive data
2. I am concerned by the amount of data in the profile
3. I am concerned because my data from multiple sources (e.g. online activities, in-store, other companies) is being combined
4. I am concerned by the level of detail (e.g. specific information, not just broad categories) in the profile
5. I am concerned about how my data may be used

After the participants rated the concerns, we asked them, using an open-ended question, whether they had any other concerns regarding the sample profile. We also asked them, using a 5-point Likert scale, if their liking for personalization had decreased after seeing the types of data collected for personalization. We were interested in knowing if awareness of behavioral profiles can change participants' opinions.

Since we could not address, with a sample profile, concerns regarding accuracy of information and editing profile data, we gave participants the option of looking at their own profiles. We made this step optional, to know whether participants were really interested in looking at their own profiles. We stated that their payment and bonus were not affected if they chose not to look at their profiles. For participants who chose to look at their own profiles, we provided instructions to access BlueKai, Google and Yahoo profiles. We then gave these participants an option to describe their reactions. This also helped us identify any additional concerns or data types. Lastly, we asked all participants if they had any further comments.

Survey Participant Background We recruited participants (n=100) from Amazon Mechanical Turk crowd-sourcing platform [70]. Our participants were at least 18 years of age and located in the United States. We used the Mechanical Turk location feature to ensure that users were from the United States. We collected informed consent from our participants. We offered a payment of $0.5 for completing the survey and a $0.3 bonus for following the survey instructions correctly. We implemented our survey on the Survey Gizmo platform, and redirected participants from Mechanical Turk to Survey Gizmo.

The average age of the participants was 27.74 years ($SD = 7.57$) and median was 26 years. The male to female ratio was four to one. Thirty seven participants had completed a four year bachelors degree or higher. Twenty five participants had a college degree or work experience in computer science, software development, web development

or similar computer-related fields. Twenty five participants were students, and the rest had diverse occupations including administration, art, business, education, engineer, law enforcement, service, skilled labor and homemaker. Our survey participant pool was more diverse than our interview participant pool especially in education level, occupation and technical background. Thirty six participants agreed (8 strongly agree, 28 agree) that they liked personalization of ads, and 39 disagreed (12 strongly disagree, 27 disagree). Hence, the pool was balanced in its opinion of personalization.

Limitations We conducted in-person interviews with graduate students ($N = 8$) with science and engineering background. For our online survey, we recruited participants ($N = 100$) from Amazon Mechanical Turk, and they may have more technical knowledge than an average person. Further, our online survey results may contain self-selection bias. By recruiting participants from a more diverse pool, we may identify new concerns and surprises. Lastly, we can improve estimation of profile accuracy by asking participants to verify information on all entries in their profiles.

4.2.2 User Concerns

Our study shows that users have several concerns about behavioral profiles including extent of collection, collection of sensitive and confidential data, and level of detail. Our interview participants considered health and credit data sensitive, but a more diverse audience may also find other data found in profiles, e.g. religion and income types, sensitive. Our analysis of data aggregator documents shows that they have much more intrusive data including fully identifying data such as first and last names, and complete postal addresses, which can further exacerbate user concerns. We identified the following concerns from user interviews.

- *Collection of sensitive data*: Participants expressed surprise and/or concern about credit and health information. One participant was surprised by credit information "Credit Card App Intent Score – 10-11%" and "Credit Card Interest Score – 16-17%." He was concerned because he did not understand the meaning or implication of the credit information in his profile. A participant who found a over-the-counter medication "OTC Medicine > Pain Reliever" said that it scared her. She had recently purchased pain medications from pharmacy for an injury. Another participant who had "General health > bones, joints, muscles > pain" considered the data confidential and did not want it to be in his profile. As result of an injury suffered during an accident, the participant was in pain for a prolonged time. He had not shared the details with other people. In his opinion, extracting this information from a few online searches and reflecting it in his profile was akin to sharing the information with others.
- *Combining data and extent of collection*: One participant who had an extensive profile with $\simeq$570 items was surprised and concerned by the amount of data

gathered. The participant's profile contained demographic – e.g. age, gender, household income – location, past purchases including a comprehensive list of ≃300 offline consumer goods purchases etc. The participant was surprised about how all the information was obtained without his knowledge or consent. Further, he was concerned to see his data from multiple sources being combined. He explained that it is okay for individual companies to have data about his business with these companies, for example, cellphone company knowing about cell phone plans, or pharmacy knowing about consumer goods purchases. However, a third party combining data from multiple sources and building profiles was not okay to him. The participant mentioned that it was not clear how all this would affect him.

- *Granularity of data*: For some data types, the concern was regarding the granularity or level of detail. One participant was okay with broad interest categories, but not with specific categories. For example, he was not concerned to see "Web Services" listed under interests. However, he would be concerned if a specific instance such as "Pirate Bay" was listed. Another participant was concerned about granularity of retention period. He pointed out that a health condition listed in his profile was more than five years old. The participant had forgotten about it, but the information was still present in his profile. The participant's concern is similar to the "right to be forgotten" argument [101].

- *Data use*: Participants were concerned about how the data in their profiles may be used. One of the participants, who had credit scores listed in his profile, was concerned about its implications. One more participant expressed similar sentiment when he said it was not clear how the extensive collection and combining of data would affect him. Both the participants were indirectly, if not directly, thinking about the purposes for which the data may be used. Another participant was more direct: he felt that data can be used to infer actions performed by the user. He was concerned that by combining different interests, for example, Pirate Bay and Movies, one could conclude that he had downloaded movies illegally.

- *Accuracy of data*: All profiles had errors to varying degrees. In general, participants were not concerned when the data was incorrect. A participant even stated that he was happy that there were so many errors. Participants, however, became concerned when the data in the profile was correct. For example, one participant initially found many entries regarding credit and income, but was not concerned. This was because the entries consistently, but erroneously, stated that the participant was affluent with 350000+ income, had top 1% credit and owned American Express card. However, after seeing an OTC medication entry that was correct, the participant said, "Now I am scared." Later, this participant hypothesized that companies added incorrect data to profiles so users would not worry too much. A participant expressed concern when only two out of twelve entries regarding professional interests were correct. One reason that contributed to user concern was the level of detail or specificity of the correct entries. Only

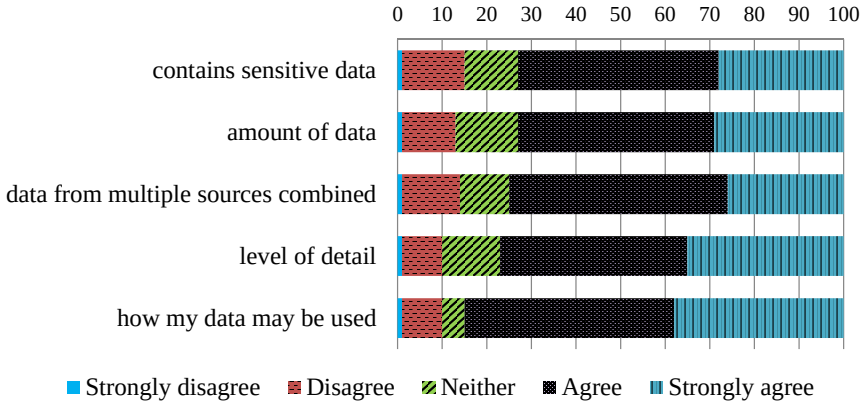


Figure 4.5: Percentage (x-axis) of survey participants ($N = 100$) who agreed with indicated concerns (y-axis)

one participant pointed out that he would be concerned about incorrect data if it was used to make adverse decisions about him. This is interesting as it highlights the importance of accuracy in behavioral profiles as perceived by users.

- *Editing profile data*: In general, participants did not try to correct erroneous data in their profiles. Two participants said that correcting the data would enable companies to track them further. A second reason was that the implications of editing data was not clear. One of the participants asked "What does edit mean? Is the data deleted from all sources?" However, we hypothesize that users may want to correct the data in their profiles if erroneous data may lead to decisions that adversely impact them.

Figure 4.5 shows how online survey participants ($N = 100$) rated concerns regarding collection of sensitive data, amount of data, combining data from multiple sources, level of detail and data use. For each of the five concerns, at least 70% of the participants either agreed or strongly agreed that they were concerned. Participants were most concerned about how their data may be used (85%), followed by level of detail (77%), aggregation (75%), amount of data (73%) and collection of sensitive data (73%). Using a MANOVA, we found that the differences among user concerns were significant ($F[4, 96] = 3.9, p < .05$). At least 70% agreement on each concern assures us that a more diverse population agrees with concerns that we identified from our in-person interviews.

We analyzed survey participant comments for additional concerns. Majority of the participants did not express new concerns. Seven participants were concerned about the security of their data; they worried that their data could be abused by hackers, criminals and identity thieves. Four participants expressed concerns that their data could be shared or sold to third parties, and accessed by the government. These are important and should be explored further.

4.2.3 Poor Data Quality

Our study shows that behavioral profiles contain large number of inaccuracies, which violates an important fair information practice principle: the data quality principle. All interview participants found varying levels of inaccuracies in their profiles. Twenty three of the 51 survey participants participants (45%) who shared information about their profiles reported inaccuracies and only three participants (6%)) reported accurate profiles. Participants reactions to inaccuracies included "blatantly incorrect," "80% inaccurate," "somewhat dated" and "hilariously overestimated." Although companies seem to be verifying the accuracy of the data that they obtain [96], it is not clear how effective their processes are. Since data is being combined from multiple companies, a few companies taking steps to ensure correctness may not be sufficient.

Some companies claim that their sources are accurate as they are "self-reported" by users and not modeled or predicted. The correctness of these self-reported sources are questionable. Users may be taking surveys or registering without being aware of implications in a different context. In fact, research has shown that people deliberately provide fake data as a way of protecting their privacy online [68]. There are many other ways in which errors may be introduced: sharing a store loyalty card with another shopper who forgot her card, browsing from a friend's account, or purchasing items for your employer.

It is also important to consider the accuracy of predictive data. It is debatable how accurate the results are when a company predicts religious affiliation, country of origin, ethnicity and languages spoken, based on an individual's name [102]. Further, desired level of accuracy would depend on the type of data (likelihood of buying toilet paper vs. median bankruptcy score) and its potential uses (advertising vs. hiring).

Interestingly, users were generally not concerned to see inaccuracies. Reactions of both interview and survey participants regarding inaccuracies were similar. Most of the participants who reported inaccuracies explained that they felt relieved and less concerned about data collection, and they did not want to correct the errors. Only two survey participants felt that inaccuracies in their profiles could adversely affect them. Three survey participants mentioned that they edited data in their profiles – one of them corrected errors and two of them deleted correct entries. Users appeared to be thinking mainly about companies tracking them, and having incorrect information about users seemed to defeat that purpose. However, users also worried over how their data may be used. Decisions based on erroneous data, for example, fraud detection based on incorrect purchases or job screening using incorrect personality type, may adversely impact users. Hence, we hypothesize that users will start caring about inaccuracies as they become more aware of its implications.

4.2.4 Insufficient Transparency

Claims of anonymity of profiles are misleading. Companies overlay anonymous data such as financial records with identifying information obtained from public, online and offline sources. This action of combining information from multiple sources not only creates a rich, 360-degree view of all aspects of life, but also associates it with a specific individual, her name, address and other personal information. Statements that imply that profile data are anonymous or pseudonymous, for example "Consumers can also control their anonymous profile [87]," are misleading.

Providing access to user behavioral profiles is an important step in the right direction. Seventy one survey participants (71%) chose to look at their own profiles even when it was optional. This indicates that people are interested in learning about their behavioral profiles. This may also indicate that many people are unaware of profile access mechanisms provided by companies. This is similar to our interview pool where only one out of eight participants was aware of profile access mechanisms.

We believe that providing access to user behavioral profiles has the potential to improve transparency of aggregator data practices. However, the information provided via these access mechanisms is incomplete and insufficient. First, our study shows a large gap between the types of data companies show in user profiles and data that they actually possess about users. For example, profiles show age, but companies also have date of birth; profiles show city, but companies also have Zip, Zip+4 and postal addresses; and companies state profiles are anonymous, but they have full names. Second, some companies that provide access are more transparent than others, for example, BlueKai vs. Yahoo or Google. Lastly, profiles show information about data types, but not about how and when they were acquired or inferred. Further, they do not show information such as frequency of purchase. These details are important to meet the goal of improving transparency into company data practices.

Clarifying data usage is essential. The biggest user concern was how their data may be used. Use of profile data is not clear. Given the variety of data present in the profiles, its uses seem limitless. Data could be used for personalization, development of better products, or fraud detection. It could also be used for hiring decisions, discreet background checks or proselytizing. For a user, the impact of using her data for the former could be quite different from that of the latter. An important underlying issue is what inferences are permissible. The richness of profile data allows one to draw all kinds of inferences about a user. If a user liked race cars on Facebook, is he likely to speed? If a user brought OTC pain medications frequently, is she addicted to pain medications? Is a user who purchases LeanCuisine brand more healthy than a user who purchases H*ä*agen-Dazs brand? Is a user who regularly buys HebrewNational brand Jewish?

4.2.5 Poor Usability of Profile Access Mechanisms

It is difficult to understand the meaning of the data displayed in the behavioral profiles. Study participants had difficulty in comprehending profile data. For example, a participant asked "What does MOB/branded data mean?" Two participants thought "High/Medium Confidence" was referring to their personality. In reality, it means that the company creating the profile has high confidence in the accuracy of the data in the profile. To understand the meaning of these and many other entries, we had to read many documents. There is a need to improve comprehensibility of behavioral profiles shown to users.

Accessing profile data is not easy. For a Bluekai profile with 99 pages (see Figure 4.2), a user has to click on each page to see its contents. Each BlueKai page has only five entries and it is not possible to display more items per page. It is not possible to save Bluekai profile contents into a text document.

Effect of editing/deleting profile data is unclear. Some study participants deleted data from their profiles to ensure that companies no longer have data about them. Are edit mechanisms meeting this expectation? There are several questions about the effect of editing or deleting profile data. Do all companies that possess a user's data honor a user's request? For example, BlueKai profile shows data that its affiliates may have about the user. Does deleting data from a BlueKai profile guarantee that the data is deleted from its affiliates databases? When a user corrects an erroneous entry in a profile, is that information propagated to companies that acquired the profile data? We need to clarify the implications of edit and delete. Otherwise, they only provide a false sense of comfort to users.

It is difficult download profile data for analysis. Six interview participants pointed out specific items from their profiles that concerned or surprised them, but could not share the entire profile with us. The primary reason for this was the time it took to share the entire profile. Due to the way the profiles were displayed, it was not possible to download entire contents of a profile into a spreadsheet or XML document. To share information, a participant had to take screen shots of each page in the profile. Individual entries in a page in a BlueKai profile were images and not text, and, hence, it was not possible to copy and paste entries into a text document. It took us over an hour to copy an large profile.

4.3 Summary

We proposed a novel approach for identifying aggregator data practices. We identified the types of data that aggregators collect about users by examining behavioral profiles of users and user data sold on data marketplaces. We studied the impact of aggregator data practices on user privacy concerns regarding data in their behavioral profiles, estimated the extent of errors in user behavioral profiles, evaluated the extent

of transparency provided by profile access mechanisms, and identified usability issues of profile access mechanisms. At least 70% of the participants expressed concerns regarding collection of sensitive information such as credit and health, level of detail, and how data may be used. We found a large gap between data shown in profiles and data possessed by companies. A large number of profiles were inaccurate with as much as 80% inaccuracy.

5 Privacy Impact of Mobile App Data Practices

Smartphones have become integral part of our lives; nearly 1.5 billion smartphones were sold worldwide in 2016 alone [103]. In 2017, there were more than 5 billion smartphone applications or "apps" that users could use on their smartphones, and users downloaded nearly 200 billion apps from mobile platforms such as Google Android and Apple iOS [104]. Smartphones contain intimate data about users e.g. contacts, call logs, GPS location and photos. They gather data using sensors e.g. sound, light and accelerometer. Given that smartphones contain highly personal and detailed information, people may even consider their phones as an extension of their selves [105]. With the growth of Internet-of-Things (IoT), smartphones are also becoming gateways to IoT physical devices [106]. For example, users can interface with smart watches, televisions, bikes and cars using apps on their smartphones. IoT smart devices can collect even more intimate user data than smartphones. Since people can also use smartphones to access the Internet, smartphones are in the unique position to enable integration of user data from mobile, IoT and Internet contexts.

In this thesis, we study how mobile apps can enable integration of user data from Mobile, IoT and Internet contexts. We identify their data practices and identify how they impact user data privacy. We use the example of smartbike apps. As of 2016, there were more than 1000 bike sharing systems in 57 countries worldwide and more than 1.2 million bikes in operation [107]. The size of the global market for bike sharing was 1.2 billion Euro and is expected to grow by 20% per year and reach 3.6 and 5.3 billion Euro by 2020 [107]. Bike sharing programs have existed for more than 50 years with many goals such as improving transit systems, promoting the environment and providing affordable and/or alternative options for users who want to share rather than own bikes. However, since the late 90's shared bikes have become highly technological and equipped with GPS, RFID and other tracking mechanisms as well as require user identification through smart cards, credit cards and other forms of identification [108]. Since then they have evolved to be used with mobile phones and integrated with other technology such as public transit infrastructure. Incorporation of technology has allowed bike sharing programs to track bicycles and user information [108]. Tracking of bike and user information has been promoted to reduce bike thefts and vandalism, and to improve bike sharing infrastructure, but its impact on user privacy has received less attention. Business models that support bike programs have evolved from non-profit to for-profit

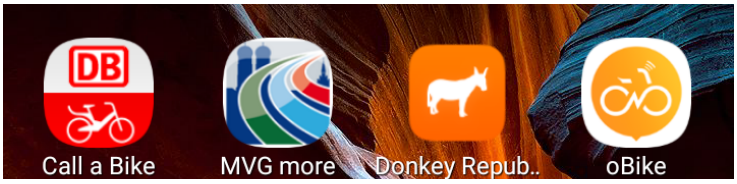


Figure 5.1: Bike sharing apps in Munich.

models supported by advertising revenue. We analyze the impact of these developments on user data privacy.

Prior work has not focused on how mobile apps can facilitate integration of user data from mobile and IoT context. Prior work has studied usability and privacy of mobile apps in general [10, 64, 109, 105, 12], but not smartbike apps in particular. Studies have examined the types of data collected from mobile apps in general [10, 110, 111], but not smartbike apps in particular. Privacy studies have looked at types of location data collected directly from smartbikes [112], but not via smartbike apps. Other studies have focused on perception of privacy e.g. feeling tracked while using bike sharing [113], perceptions of location data collected from smartbikes [114] and re-identification of users from de-identified bike sharing data [115].

5.1 Identifying App Data Practices

We discuss several techniques that can be used to identify app data practices. We apply these techniques to four bike sharing apps in the city of Munich, Germany. Bicycle traffic accounts for 20% of all commuting traffic in Munich. Given this, bike sharing programs are increasing in popularity. In 2017, two new bike sharing programs were introduced in Munich bringing the total number of bike sharing programs to four, which are Call a Bike, MVG more, Donkey Republic and oBike. All four bike sharing programs have mobile apps (see Figure 5.1) through which users can locate and reserve bikes. We study the bike apps from the Google Android platform available for download via Google Play app store.

5.1.1 App Analysis

We can analyze app behavior using static and dynamic analysis techniques such as permission analysis, static code analysis, taint analysis and network call analysis. We can also analyze privacy policy of a app, if available, to understand the app's data practices. We can compare whether actual app behavior aligns with app data practices disclosed in the app's privacy policy.

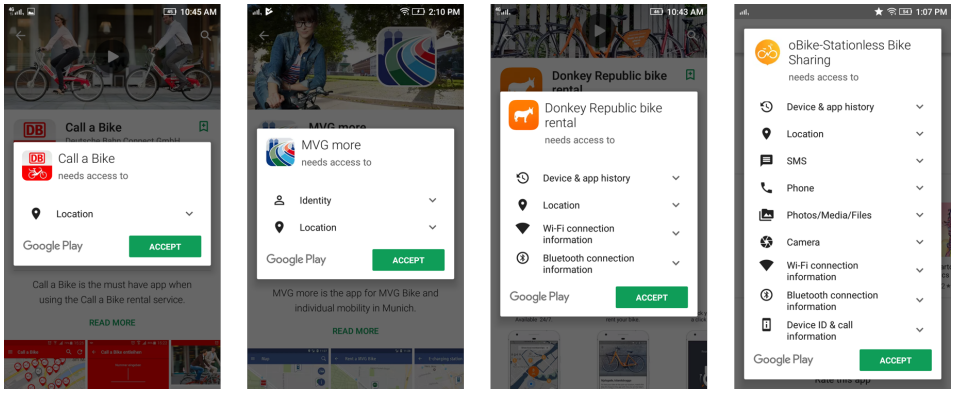


Figure 5.2: Bike share app install time permissions.

App Permission Analysis Mobile apps need different types of "permissions" to run on smartphones. Apps can request permissions either during the initial installation or later software updates. Prior research shows that apps can request permissions that may or may not be necessary to provide app functionality [116]. Further, users generally do not pay attention to permissions requested at installation [109].

We identify permissions requested by the four bike sharing apps using three methods. First, we consider the permissions requested during initial installation from the Google Play app store. Figure 5.2 shows the privacy notice for each app displayed at installation time. The privacy notices displays permissions requested by the apps. All apps have some permissions in common e.g. "Location," but some apps request more permissions than others ($min = 1$, $max = 9$). The app oBike in particular requests permissions such as SMS, Phone and Photos/Media/Files that may be unnecessary for providing core smartbike app functionality. Table 5.1 compares the install time permissions of the four apps. It highlights some of the permissions that may impact user privacy in red. Other permissions such as Location, Device and app history, Device ID and call info may also allow apps to collect private data. Information in the install time privacy notice may be insufficient to understand why apps request the permissions.

Other interfaces available on Android phones to examine app permissions may display more detailed information about app permissions. For e.g. some phones have "Settings > Permission Control > Apps" interface. This interface provides different information, which is slightly more detailed information than that provided by the install time privacy notice. Table 5.2 shows the information displayed by the Permission Control interface. Note that it displays "Initiate multi-party calls" instead of "Phone." Other interfaces may display even more detailed information about app permissions. For e.g. the "Settings > Apps Management" interface available on some Android phones displays information shown in Table 5.3. For e.g. it shows whether "Location" permission required is "Approximate location (network-based)" or "Precise location (GPS and network based)." This interface also groups related permissions together. For ex-

ample, it groups "Directly call phone numbers" and "Read phone status and identify" together.

Table 5.1: Smartbike app permissions displayed by Google Play mobile platform privacy notice at installation time.

Call a Bike	MVG more	Donkey Republic	oBike
Location	Location	Location	Location
		Device & app history	Device & app history
		Wi-Fi conn info	Wi-Fi conn info
		Bluetooth conn info	Bluetooth conn info
	Identity		
			Camera
			Device ID & call info
			Photos/ Media/ Files
			SMS
			Phone

Permissions as of April 23, 2018. Some of the permissions that may impact user privacy are shown in red.

Table 5.2: Smartbike app permissions shown in the "Settings > Permission Control > Apps" interface on Android platform.

Call a Bike	MVG more	Donkey Republic	oBike
Get Position	Get Position	Get Position	Get Position
Send MMS	Send MMS	Send MMS	Send MMS
Send Email	Send Email	Send Email	Send Email
		Turn on Bluetooth	Turn on Bluetooth
			Use camera
			Make calls
			Initiate multi-party calls
			Send SMS

Permissions as of April 23, 2018.

Static Code Analysis We can use static analysis tools to understand the purposes for which apps request permissions [10, 117]. For example, we could examine why a bike sharing app requests permissions such as "Precise Location," "Modify System Setting," "Access SMS" or "Read Internal Storage." Static analysis tools analyze how app functions use permissions [117]. For example, they can identify whether "Precise Location" permissions is used for the purpose of locating bikes or for the purpose of

Table 5.3: Smartbike app permissions shown in the "Settings > Apps Management" interface on Android platform.

Call a Bike	**MVG more**	**Donkey Republic**	**oBike**
Approximate location (network-based) Precise location (GPS and network based)	Approximate location (network-based) Precise location (GPS and network based)	Precise location (GPS and network based)	Approximate location (network-based) Precise location (GPS and network based)
Full network access View network connections	Full network access View network connections Receive data from Internet	Full network access View network connections View WLAN connections	Full network access View network connections View WLAN connections Receive data from Internet Control Near Field Communication
Prevent phone from sleeping Control vibration	Prevent phone from sleeping	Control vibration	Prevent phone from sleeping Control vibration Control flashlight
		Access Bluetooth settings Pair with Bluetooth devices	Access Bluetooth settings Pair with Bluetooth devices
	Mock location sources for testing	Mock location sources for testing	Modify system settings
	Add or remove accounts Create accounts and set passwords Find accounts on the device Use accounts on the device		
		Read Google service configuration	
			Directly call phone numbers Read phone status and identity
			Send SMS messages
			Take pictures and videos
			Modify or delete the contents of your SD card Read the contents of your SD card

Permissions as of April 23, 2018.

Table 5.4: Third-party libraries used by bike sharing apps.

Third-party Library	Privacy Impact	Call a Bike	MVG more	Donkey Republic	oBike
Development Aid	Low	Y	Y	Y	Y
Mobile Analytics	High		Y	Y	Y
Social Network	High		Y	Y	Y
Map/LBS	High		Y	Y	
GUI Component	Low		Y	Y	
Utility	Low		Y	Y	
Payment	High				Y

Development Aid, GUI Component and Utility libraries provide core functionality needed by apps. Maps/LBS and Payment provide specific services such as location and payment that apps may need. Apps may use Mobile Analytics and Social Network to track user activities on the app, social networking sites and other websites.

displaying ads. Static analysis tools analyze how apps process user data [117]. They can identify third-party libraries that apps use [118]. Presence of third-party libraries can indicate how apps process permissions and user data e.g. apps can use advertising-related third-party libraries to sharing users' precise location with advertisers.

We used LibRadar Android app analysis tool to detect third-party libraries in the four bike sharing apps [118]. Table 5.4 shows and compares the types of third-party libraries used by the four apps. For example, the table shows whether the bike sharing apps use social networking libraries. Call a Bike does not use any, but the other three do; MVG more and Donkey Republic use Facebook, and oBike uses WeChat. Facebook is primarily based in the United States and WeChat is based in China.

The "Privacy Impact" column in Table 5.4, indicates whether the potential impact of third-party libraries on user data privacy. We classify privacy impact as either High or Low. We classify the impact of Development Aid, GUI Component and Utility as Low because these third-party libraries are likely to use user data for providing core functionality of the mobile app. We classify the privacy impact of Map/LBS and Payment third-party libraries as High. These libraries do not collect lot of user data, but collect specific types of sensitive data e.g. current location and financial user data. Although, they may use it only to provide core functionality of the app, there is a chance that the third-party libraries may use data for other purposes. For example, smartbike apps may use Map/LBS library to locate bikes in the neighborhood, and they may use Payment library to process user payments. However, Google Maps may combine user location information with other data collected by Google services and use the combined data for advertising purposes. Given the possibility of misuse of sensitive data, we classify the privacy impact as High. We classify the privacy impact of Mobile Analytics and Social Network as High. They collect a lot of user data including sensitive data and are likely to use it for purposes not required to provide core functionality of the app.

Dynamic Behavior Analysis We can analyze the behavior of an app at run time. First, dynamic analysis can identify the data sent over the network. For example, apps have been found to upload user contacts on the phone to an external server [10]. We can also examine whether apps collect personally identifiable data. Second, dynamic analysis can identify the services that an app contacts. For example, apps could send data to advertisers, aggregators or social networking services. App behaviors could vary under different scenarios e.g. while using a bike or not using a bike. To analyze app behavior, we examine data sent and calls invoked under the following scenarios:

- App installation time
- First use without logging in to an account
- Subsequent use without logging in to an account
- While creating an account with a bike service provider
- First time logging into an account
- First use with login
- Subsequent login
- Subsequent use with login
- App behavior while user is using a bike

We used two tools to analyze the data sent over the network by bike sharing apps at runtime. Both tools intercept and log data sent from a bike sharing app. They use a man-in-the-middle attack to capture the data. The first tool, SSL Packet Capture app, runs on a smartphone. The second tool, Mitmproxy [119] runs on a laptop. Both tools install trusted SSL certificates, either on the phone or on the laptop, to decrypt SSL traffic from the app. Figure 5.3 shows data captured from Donkey Republic bike sharing app using the SSL Packet Capture app. In the case of Mitmproxy tool, we force all network traffic from the smartphone to go through the laptop. We used the packet capture app tool in the scenario where a user is riding a smartbike, and first use of the app. We used the Mitmproxy tool to capture data from all other scenarios.

Figure 5.3 shows data captured using SSL Packet Capture app (left) and Mitmproxy (right) tools. It highlights the data sent from the Donkey Republic app to a third-party Branch.io. SSL Packet Capture log shows that the app sends unique device identifiers such as *device_fingerprint_id*, *identity_id*, *hardware_id*, and *google_advertising_id* that can personally identify users. The app also sends *local_ip*, which can be used to uniquely identify a user behind a network with a public IP address. Mitmproxy log shows that the Donkey Republic app sends personally identifiable information – full name, email address and mobile phone number – to a third-party Segment.com.In the picture, the name and parts of the email and phone number have been whited out for privacy reasons.

Table 5.5 shows the entities contacted by the four bike sharing apps while a user is using the app for the first time. The user is using the app without logging in to an account. Call a Bike app contacts the least number (2) and oBike app contacts the most number (5) of entities. All apps contact Google for maps and other services.

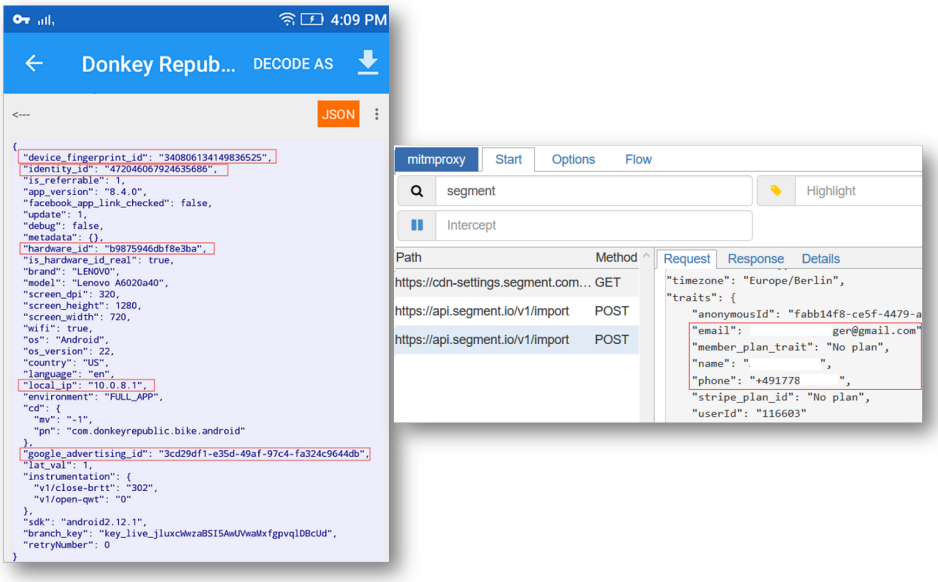


Figure 5.3: Dynamic behavior analysis using SSL Packet Capture (left) and Mitmproxy (right).

Donkey Republic and oBike contact Facebook even when user is not logged in to Facebook or is a member of Facebook. MVG more, Donkey Republic and oBike contact third-parties that enable tracking users activities, creating user behavioral profiles and aggregating user data; MVG more uses Etracker.de, Donkey Republic uses Branch.io and Segment.com, and oBike uses Appsflyer.com. The oBike app contacts a payment service called Bluepay.asia located in Thailand. They also contact third-parties that analyze app-related crashes; MVG more and Donkey Republic use Crashlytics, and oBike uses Bugly.

Biking apps collect personally identifiable information both when user has an account and user does not have an account. They share it with third-parties such as trackers and aggregators. When user does not have an account, apps collect unique device identifiers such as IMEI and IP address, which are considered to be personally identifiable information in the EU. For example, oBike sends device IMEI information to Bluepay.asia. Donkey Republic sends identifiers such as *device_fingerprint_id*, *identity_id*, *hardware_id* and *google_advertising_id* to Segment.com.

Biking apps collect personally identifiable information such as name, home address, mobile phone number and email address when a user registers and creates an account with the app service provider. They can share users personal information with third-parties. For example, MVG more shares users' email address with Etracker.de; Donkey Republic shares name, email and mobile phone number with Segment.com.

Table 5.5: Dynamic analysis of app calls to to third-party websites/services during first use (no login).

Call a Bike	**MVG more**	**Donkey Republic**	**oBike**
Google.com	Google.com	Google.com	Google.com
Gstatic.com	Gstatic.com	Gstatic.com	Gstatic.com
		Facebook.com	Facebook.com
	Crashlytics.com	Crashlytics.com	
2denker.de			
	Etracker.de		
		Branch.io	
		Segment.com	
			Appsflyer.com
			Bluepay.asia
			Bugly.qq.com

First use indicates that the user is interacting with the app for the first time after installing the smartbike app. User has not created an account with the smartbike app and has not logged in to the account.

Even when users do not have an account, biking apps collect geo-location of users. Prior research demonstrates that human mobility patterns are quite unique, and it is possible to identify individual users from their mobility patterns [120]. Hence, geo-location information is personally identifiable information. When location service is enabled on the smartphone, apps can get users' current location information. For example, when location services are enabled, Donkey Republic collects precise location (latitude and longitude) information from the smartphone.

- When location service is enabled

```
1 properties:{
2   user_location_available:true,
3   user_location:48.1469192, 11.5630183,
4   user_location_latitude:48.1469192,
5   user_location_longitude:11.5630183,
6   user_location_accuracy:19.037
7 }
```

- When location service is not enabled

```
1 properties:{
2   user_location_available:false
3 }
```

Biking apps use geo-location information for different purposes such as locating bikes, advertising, marketing, optimization, crash analysis etc. For example, Call a Bike app uses location information to locate bikes in the vicinity of the user. The MVG more app allows Crashlytics to collect user location information for processing errors and crashes related to the app. The oBike app collects and processes latitude and longitude information through its advertising API. Using the API, the app could display advertisement based on users' current location. As shown below, during our analysis, oBike app invoked the advertising API, but the response did not contain any advertisements in the "advertisement: []" field.

- oBike advertising API request

```
1 GET https://mobile.o.bike/api/v1/advertisement?
2     latitude=48.1468744013432&
3     longitude=11.562939696013927&
4     countryCode=49&
5     type=4
6     [...]
```

- oBike advertising API response

```
1 [...]
2 {
3     data: {
4         advertisements: []
5     },
6     errorCode: 100,
7     success: true
8 }
```

Privacy Policy Analysis A mobile app may have a natural language privacy policy that describes data practices of the app. However, many apps may not have such a policy; Zimmeck et al. studied 17991 apps and found that only 9295 apps (∼50%) had a privacy policy [121]. When a policy is available, we can extract app data practices from the policy using techniques discussed in Chapter 2. This can provide additional information about data practices of the apps. We can also identify whether there are mismatches between what apps state in the policy and what apps actually do. For example, Zimmeck et al. compare app data practices identified by analyzing app code with data practices extracted from policies by using machine learning classifiers trained on a human-annotated policy corpus [121]. Apps may have different data practices for different countries e.g. oBike has a specific data practices for Germany [122].

When we analyze the privacy policy of the MVG more biking app, we find that it contradicts itself and is not aligned with actual data practices of the app. First, the privacy policy initially states that third-party service provides process user data within the European Economic Area (Europäischen Wirtschaftsraums), but later states that

the Crashlytics app sends user data to the United States. Initially the privacy policy contains the following statement: "MVG has commissioned service providers for the provision of individual services in connection with MVG Rad and the provision of this app. These service providers process your data exclusively within the European Economic Area." However, the statements "Our app uses Crashlytics, a service of Crashlytics Inc., to analyze flaws in the app and fix problems. To do this, real-time crash reports are sent to Crashlytics Inc. in the USA [...]" that appear later in the policy contradict the earlier statements regarding data being processed within the EU. Crashlytics collects unique device identifiers and current location information, both of which can identify users. Second, the privacy policy states that the app uses two third-party service provides Etracker.de and Crashlytics. However, dynamic network analysis shows that the app also uses Nextbike.net and Worldpay.com, which are not disclosed in the privacy policy. MVG more app shares personal information (name, home address, mobile number and email address) with Nextbike.net. It uses Worldpay.com for processing user transactions, and shares name, home address and payment information with with Worldpay.com. Lastly, the app collects mobile phone number while a user registers for an account, and it states that "MVG will never call you for advertising purposes!" However, this information is not stated in the privacy policy.

In addition to privacy policies, apps may disclose data practices through other information sources such as FAQs. For instance, oBike has an FAQ that discusses data practices regarding processing and storage of data from German users, and states that "The data are kept in France and are therefore subject to European data protection law." [122] This information is not available in its privacy policy. We know from static and dynamic analysis of oBike app that it sends users' personally identifiable information to services in China and Thailand. Hence, user data is not solely stored in France, and there is a mismatch between oBike's stated app data practices and actual app data practices.

Limitations While a user is riding a smartbike, bike sharing programs can collect data both from the bike sharing app and the bike itself. For example, programs can collect location data using GPS on the smartphone as well as GPS on the smartbike. Obike states in its FAQ that "While you borrow an oBike, we collect your movement data and that of the bike." Smartbikes can directly transfer user data from the bike to an external service via mobile telephone communication networks [123]. Capturing data sent and received from the smartphone, e.g. using SSL Packet Capture app, will not capture data sent from the bike to the external service. Tools can be used to capture and analyze data sent from the smartbike to external services.

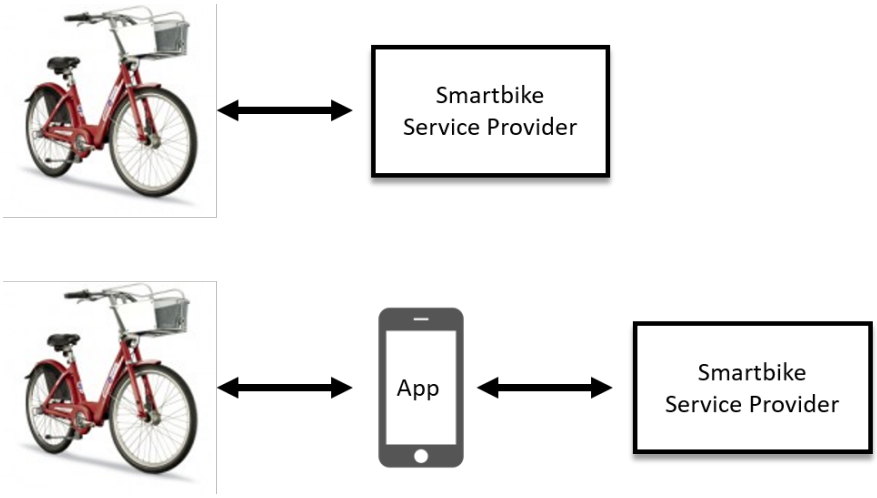


Figure 5.4: Collecting data directly from the smartbike (top) vs. collecting data via smartbike app (bottom)

5.2 Impact of App Data Practices

Figure 5.4 compares two scenarios: users interact with smartbikes without a smartbike app (top figure) and with a smartbike app (bottom figure). In the absence of a smartbike app, bike service providers collect user data directly from the bike. When an app is present, service providers can collect data through the app as well as directly from the bike. Additionally, service providers can collect data through websites if users interact with the website of service provider e.g. booking a smartbike via the website. Apps can also facilitate interaction through a service provider's website, for example, redirecting a user to the website for creating an account or displaying app privacy policy. Below, we discuss how the presence of an app and its data practices impacts user data privacy.

We discuss the impact of bike sharing app data practices on user data privacy using results from two studies. In the first study, we analyzed four bike sharing apps available in Munich. In the second study, participants ($N = 12$) analyzed bike sharing apps and shared their opinions with us. Participants were graduate students enrolled in a privacy and big data course in a Technology and Policy masters program. Participants analyzed the apps as part of a graded exercise for the class. Survey data was collected between January 18,2018 and February 4,2018. The exercise questionnaire is available in Appendix A. From the study, we gathered data about 64 bike sharing apps (30 unique) from 21 cities in the European Union. Some of the smartbike apps updated their app permissions after survey responses were collected, for example, oBike app added SMS and Phone permissions. Hence, participants could not comment on such app permissions in their responses.

5.2.1 Increase in Collection of User Data

Smartbike service providers can collect more data about users when users interact with smartbikes using smartbike apps. The increase in data collection is related to *type*, *granularity* and *temporal* aspects of user data.

Type of Data: Service providers can collect more types of user data. When data is collected solely from smartbikes, it is generally limited by the types of sensors mounted on the bike. Smartbikes generally have embedded GPS sensors that can collect location information while the bike is in use [114, 112]. However, smartphones have many more sensors such accelerometer, microphone, camera, proximity, compass, light etc. in addition to GPS sensor. Service providers can use the smartbike app to collect data from the additional sensors. For example, service providers can collect not only the location of the bike user, but also collect data regarding the speed at which the bike is traveling using the accelerometer sensor. Using smartbike apps, service providers can collect personally identifiable information that would not be available otherwise. Apps collect unique device identifiers such as IMEI number, hardware_id, device_fingerprint_id etc. that can personally identify users. Because apps request permissions such as $Read/Modify/DeleteStorage$, $Identity$, $Contacts$ and $ModifySystemSettings$, they can collect data regarding user contacts, photos, phone status and identity etc. Users find such permissions to be sensitive and/or unusual.

Apps can request varying number of permissions, for example, 64 apps (30 unique) from 21 EU cities requested access to $Min = 1$, $Mean = 8.01$, $SD = 5.11$ and $Max = 27$ number of permissions. Hence, several apps request excessive number of permissions that are unnecessary to provide core bike sharing service. This violates the data minimization principle, which requires service providers to collect and process only the data that is required to successfully provide a service. Participants in our study ($N = 12$) opine that the number of permissions requested by the apps is influenced by the business model of the bike sharing app. They think that apps run by non-profit companies and those that receive government subsidies generally request fewer permissions, but apps run by for-profit companies request more permissions.

Granularity of Data: Service providers can collect data with finer granularity. Consider for example location information. When gathered solely from a smartbike, location information is limited by the GPS sensor embedded on the smartbike. However, apps can collect location information via the smartphone using GPS sensors as well as other techniques such as WiFi information and cellular triangulation. Apps can access both approximate and precise location information on the smartphone e.g. three of the four smartbike apps in Munich collect both.

Duration of Data Collection: Service providers can collect user data even when the user is not using a smartbike. For example, apps can collect location data as long as location service is available on the smartphone. If a user turns on the location service while using the app, the user has to remember to turn it off after using the smartbike. If the user forgets, apps can continue to gather user's current location information.

5.2.2 Increase in Sharing and Re-purposing of User Data

Because of smartbike apps, user data is shared with more entities. Apps share PII, precise location and other user data with third-parties, trackers and aggregators. Many apps use Google map service, and share users location with Google. Third-parties that receive user data can facilitate sharing of data with a large number of trackers and aggregators. The oBike app shares user PII with a third-party Appsflyer, which provides easy integration with a data ecosystem consisting of aggregators and trackers who can use the data for advertising, marketing and other purposes [124]. The oBike app does not prevent Appsflyer from sharing user data. While sharing user PII with the third-party Branch, Donkey Republic app opts to be tracked by advertisers. It could restrict sharing of data for advertising by setting the $limit\_ad\_tracking : true$ option, but it does not. Both oBike and Donkey Republic share PII with Facebook, which uses data for advertising purposes. Android users have to explicitly opt out of behavioral tracking by changing default phone setting. If users do so, Facebook will not use user data for behavioral advertising, but continue to use it for non-behavioral advertising.

To allow apps to get location information, Android users have to turn on the location service feature on Android smartphones. Enabling the feature allows not only the bike sharing app, but also other apps to collect user location information. Google and other apps may collect location information even when a user is not using them. Users consider access to location information as reasonable when used to locate bikes. However, apps share data with third-parties that can re-purpose location and user data for advertising, marketing etc.

5.2.3 Increase in Aggregation of User Data

Smartbike apps facilitate aggregation of user data in the following ways.

- **User Activities:** Apps use third-parties to track user activity on smartbike apps. For example, oBike with Appsflyer, MVG more uses Etracker.de and Donkey Republic uses Segment.com. Each time user uses the app, apps send information to third-parties regarding the time of use,user location, how long the user engaged with the app etc. Third-parties combine user activities across multiple uses to create user behavioral profiles.

- **Data from Mobile Apps:** Apps enable combining data from smartbike app with all other apps on the smartphone. Both smartbike and other apps share user *advertsing_id* with third-parties such as Facebook. Third-parties can uniquely identify users and combine user data from multiple apps.

- **IoT and Mobile Context:** Apps can combine data from physical IoI context and mobile context. Because of the apps, data collected from the smartbike can be combined with data from users' smartphone.

- **Mobile and Website Context:** Apps can combine data from mobile app and website contexts. For example, Call a Bike redirects a user from the app to its website for registering an account. MVG more redirects a user from within the app to its website for displaying the privacy policy. Donkey republic app does not provide a link to its privacy policy. To access the policy, one has to go to Donkey Republic's website.

- **Biking and Social Media:** Apps can combine data from biking context with social media context. For example, Donkey Republic and oBike share information regarding user activities to Facebook. They also allow users to log in to the app via Facebook login, which enables Facebook to combine user activities on smartbike apps with social media activities on Facebook.

- **Mobile and Offline Context:** Apps send data to data aggregators e.g. Segment and Appsflyer who combine user data from app context with data obtained from offline contexts such as in-store purchases [5].

- **Multiple modes of transportation:** Apps can combine user activities related to different modes of transportation such as bike, car, bus and train services. For example, Call a Bike and MVG more have a single app for renting shared bikes and cars. Users also have to use the same app for buying tickets for bus and train services.

5.2.4 Decrease in Transparency

Decrease in transparency is related to two issues: consent-related and purpose-related. Smartbike apps use default opt-out consent as opposed to default explicit opt-in consent. With opt-out consent, users are implicitly subscribed to less privacy protective data practices. Users are generally neither aware of the setting not understand what the setting does. Such default settings may not be easy to identify and modify. With opt-in, users have to explicitly consent to the data practice. Although, users may not understand what they are consenting to, they may at least become aware of the setting. Using opt-out settings do not adhere to the Privacy by Design and Privacy by Default recommendations.

- The MVG more app users two trackers using opt-out default settings. Etracker.de tracker collects user activities and creates behavioral profiles. CrashAnalytics tracker collects crash information and user location. Unless users read the privacy policy of the app, they are not aware of the default settings. Further, changing the default settings is not straightforward.

- Donkey Republic and oBike apps use the Google Advertising ID *google_advertising_id* to determine whether users consent to behavioral advertising. On the Android platform, Google sets the value of *google_advertising_id* to true by default i.e. opt-out. The apps send user data and *google_advertising_id* value to third-parties such as Facebook. Third-parties use user data for behavioral advertising if *google_advertising_id* is true. Otherwise they use user data for non-behavioral advertising.

Compared to MVG more and Call a Bike, users think that Donkey Republic and oBike request excessive number of permissions that are sensitive, unusual and unnecessary to provide core bike sharing service. For example, apps request permissions such as $Read/Modify/DeleteStorage$, $Identity$, $Contacts$ and $ModifySystemSettings$ that users consider sensitive. Users do not completely understand the purposes for which apps request those permissions. Hence, there is decrease in purpose-related transparency.

5.2.5 Decrease in User Data Protection

Apps send user personally identifiable information to countries that may have not have data protection regulations comparable to the country of data origin. For example, in our study data of German users was sent to China, Thailand and the United States where data may not be protected on par with the EU General Data Protection Regulation (GDPR). For instance in the United States, unique device identifiers are not treated as PII. Disclosure of app data practices can be inconsistent with actual app data practices. For example, oBike states that it stores user data in France, but sends user data to China, Thailand and the United States. Similarly, MVG more states that user data is processed solely within the European Economic Forum, but in reality user data is also processed in the United States.

5.3 Summary

Using the example of smartbike apps, we studied the impact of app-in-the-middle scenario i.e. how mobile apps can enable integration of user data from mobile and IoT contexts. We discussed techniques to identify app data practices. We analyzed four smartbike apps – Call a Bike, MVG more, Donkey Republic and oBike – available in

Munich, Germany. We identified smartbike app data practices and assessed their impact on user data privacy. Our analysis illustrated how the app-in-the-middle scenario increases data collection, sharing and re-purposing of user data, and decreases transparency and data protection. Several data practices of smartbike apps are inconsistent with European data protection regulations (GDPR) either directly or indirectly and can circumvent GDPR.

- Smartbike apps collect personally identifiable information (PII) both when user has an account and user does not have an account. When user does not have an account, apps collect and share unique device identifiers such as IMEI and IP address, which are considered to be PII in EU. When users create an account, apps collect and share PII such as email address and mobile phone number. They share PII (e.g. email address, mobile phone number, IMEI, IP address etc.), precise location and other user data with third-parties such as trackers and data aggregator companies. They track user activities, create user behavioral profiles, and use user data for advertising purposes
- Smartbike apps request excessive number of sensitive and unusual app permissions (e.g. SMS, Identity, Contacts, Modify System Settings etc.) that are unnecessary for providing core bike sharing service thereby violating the data minimization principle, which requires service providers to collect and process only the data that is required to successfully provide a service. The extent of permissions requested is influenced by the business model e.g. for-profit or non-profit
- Smartbike apps use default opt-out consent as opposed to default explicit opt-in consent. Hence, do not adhere to the Privacy by Design and Privacy by Default recommendations
- Smartbike apps send user PII to countries such as Thailand, China and USA that may have not have data protection regulations comparable to EU GDPR
- Lastly, smartbike app privacy policies are incomplete, inconsistent and contradictory. App data practices disclosed in the privacy policy do not match actual app data practices

6 Discussion and Conclusion

In this thesis, we used qualitative and quantitative social research methods to identify data practices of websites, trackers, aggregators and mobile apps. We assessed their impact on user data privacy. Below we discuss the implications of our results for privacy research, public policy and regulations, and development of privacy enhancing technologies.

Privacy Expectations We proposed a conceptual model of privacy expectation with Desired, Deserved, Predicted and Minimum types, and ordering among the types in terms of levels of user privacy. Our conceptual model contributes to privacy theory. Results from our empirical study supported the existence of multiple types of privacy expectations and their ordering. This implies that to measure user privacy expectations precisely, studies have to differentiate among the types. Otherwise they may inadvertently use questions such as "would you expect it to access your precise location?" to measure what users think (Predicted type) [10] and "how much did you expect this app to be accessing this resource?" to measure what users want (Desired type) [64]. Studies may also or ask users to rate the statement "This application meets my privacy expectations" to measure expectations [11, 12] that are related to the Deserved or Minimum types.

By largely studying user preferences, research in the privacy domain has implicitly focused on the Desired type. It is important to study other types of privacy expectations. For instance, Turow et al. found that 66% of Americans do not want websites to show them ads tailored to their interests [21]. However, our study shows that even in a privacy sensitive-scenario, Americans' Desired level is different than their Minimum level. Hence, Americans may not desire tailored ads, but such ads may meet their minimum expectations. Studies measuring the privacy paradox, gap between intended behavior and actual behavior, may find gaps when they measure the Desired level, but not when they measure the Minimum level. We hope our results will foster more research into privacy expectation types.

Research in the Consumer Satisfaction/ Dissatisfaction (CS/D) domain shows that the gap between expected performance, based on a type, and perceived actual performance can significantly predict satisfaction. In the privacy domain, gap between expected privacy, based on a type, and perceived actual privacy may predict satisfaction and privacy concern. For example, a gap between the Minimum level and reality may better predict privacy concern than a gap between the Desired level and reality. However, a

gap between the Desired level and reality may better predict satisfaction than a gap between the Minimum level and reality.

Differences among expectation types and reality can impact personal and public policy decisions regarding user privacy. For example, it may be more important to take regulatory action when reality fails to match the Minimum level rather than the Desired level. People may take different actions based on the differences. For instance, when banks do not meet desired privacy expectations, people may be unsatisfied but still remain a customer. However, they may switch banks if minimum privacy expectations are not met. People may not take action if their Predicted level does not match reality. In our scenario, nearly 80% of the participants do not predict that banks may collect health-related browsing activities. Hence, if banks collected it, the Predicted level would not match reality, and, because of lack of awareness, people would not take action even when minimum privacy expectations are not met.

Top banks in the United States may collect health-related browsing activities of users [55, 56]. In our scenario of banks collecting health-related browsing activities, at least half of the participants did not desire it, predict it, feel that they deserve it or tolerate it under any circumstances. Studying multiple expectation types provides a more nuanced view of users' privacy expectations. When a bank collects customers' health-related browsing activities, it shows that reality does not match any type of privacy expectation of a majority of the population. It enables us to differentiate between scenarios where reality matches none, some or all of the users' privacy expectations. This can help consumer protection agencies in identifying scenarios that have a stronger need for intervention.

Mismatched Privacy Expectations We identified mismatches in user expectations regarding online data practices. Further, we identified factors that impact such mismatches. We believe that emphasizing such mismatches in privacy notices could help users make better privacy decisions. Simplified privacy notices [36] that complement comprehensive privacy policies could highlight mismatched expectations. Current simplified privacy notices, for example privacy nutrition labels [125], although an improvement over privacy policies, are themselves too complex. By identifying mismatches in users' privacy expectations, one could selectively highlight or display elements of a privacy nutrition label or other notice formats relevant to users. Our results suggest that the number of mismatches is small compared to the total number of website data practices. Hence, emphasizing unexpected data practices could reduce the amount information in the notice that users have to comprehend.

Although website operators could themselves generate simplified notices, the low adoption of simplified and standardized notice mechanisms [126] indicates that many website operators may not do so. An alternative approach is for a third-party to highlight unexpected data practices based on mismatched expectations. For example, a browser extension could generate and display a simplified notice [43, 41]. Such a notice

could highlight snippets of text from the natural language privacy policy, corresponding to unexpected data practices. Currently third-party browser extensions, such as Ghostery (www.ghostery.com) and Privacy Badger (www.eff.org/privacybadger) generate and display information regarding online tracking practices. Similarly, a third-party browser extension could display information regarding unexpected data practices. Extensions could use just-in-time notifications or static icons that users can click to gain more information. At installation time, the extension could gather user characteristics such as privacy knowledge, concerns and demographics in order to tailor which practices are emphasized to individual users.

Organizations could also use our approach to obtain a competitive advantage by making their website data practices and privacy policies easier to understand. In the past, organizations such as Google, have tried to organize information within their policy along dimensions that are important to people, with the intent of making information easier to access. Mismatches in expectations are important, and highlighting them can aid in such efforts. Regulatory agencies such as the United States Federal Trade Commission work on protecting users' privacy, and mismatched expectations could indicate to them important public policy issues that need attention.

Tracker Data Practices We showed how we can use network analysis to identify data practices that enable linking user activities across different contexts. We investigated the role of Internet tracking in siphoning users' personal data from one country to another. We identified several tracking patterns that can siphon user data. Two key parameters of the tracking patterns, distance to data and type of control, determine timeliness, accuracy and granularity of siphoned data. Tracking patterns provide trade-off between quality of data and visibility, and by using patterns with low visibility, it is possible to evade detection and data protection regulations.

Aggregator Data Practices Using behavioral profiles analysis, we studied the impact of aggregator data. We studied user concerns regarding data in their behavioral profiles, estimated the extent of errors in user behavioral profiles, evaluated the extent of transparency provided by profile access mechanisms, and identified usability issues of profile access mechanisms. Companies that create behavioral profiles have to provide better notice to users about collection, combining and potential uses of user data. Notice includes improving awareness of access mechanisms among users. Currently, there seems to be little awareness, for example, only one out of eight interview participants in our study, knew about access mechanisms. Currently, data from very different contexts are being combined. It is important to obtain users' consent in this matter. There is a need to disclose what inferences are drawn and how prediction models are implemented. It is not sufficient for a company to state that they use proprietary models, for example, "developed a proprietary algorithm that utilizes a consumer's name, mailing address and 320 different data points to accurately assign a personality type to 85% of US adult

consumers [96]." Companies have to ensure accuracy in profile data, and address the issue of accountability for adverse impact arising from errors in profiles.

Bibliography

[1] C. Bennett. Privacy in the political system: perspectives from political science and economics. Report for Ethical, Legal and Social Issues (ELSI) component of the Human Genome Project, U.S. Department of Energy, 2001.

[2] A. M. McDonald and L. F. Cranor. The cost of reading privacy policies. *ISJLP*, 4, 2008.

[3] C. Jensen and C. Potts. Privacy policies as decision-making tools: An evaluation of online privacy notices. In *Proc. CHI '04*, pages 471–478. ACM, 2004.

[4] I. Pollach. What's wrong with online privacy policies? *Commun. ACM*, 50(9):103–108, September 2007.

[5] A. Rao, F. Schaub, and N. Sadeh. What do they know about me? contents and concerns of online behavioral profiles. In *Proc. PASSAT '14*. ASE, 2014.

[6] A. Rao, F. Schaub, N. Sadeh, A. Acquisti, and R. Kang. Expecting the unexpected: Understanding mismatched privacy expectations online. In *Symposium on Usable Privacy and Security (SOUPS)*, pages 77–96. USENIX Association, 2016.

[7] J. R. Mayer and J. C. Mitchell. Third-party web tracking: Policy and technology. In *Proceedings of the Conference on Security and Privacy*. IEEE, 2012.

[8] F. Roesner, T. Kohno, and D. Wetherall. Detecting and defending against third-party tracking on the web. In *Proceedings of the Symposium on Network Systems design and Implementation (NSDI)*. USENIX, 2012.

[9] Evidon. Company Database [online]. 2018. URL: `https://www.evidon.com/resources/company-database/` [last checked 2018-02-20].

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

[11] K. Martin and K. Shilton. Why experience matters to privacy: How context-based experience moderates consumer privacy expectations for mobile applications. *Journal of the Association for Information Science and Technology*, 67(8):1871–1882, 2016.

[12] K. Martin and K. Shilton. Putting mobile application privacy in context: An empirical study of user privacy expectations for mobile devices. *The Information Society*, 32(3):200–216, 2016.

[13] G. R. Milne and S. Bahl. Are there differences between consumers' and marketers' privacy expectations? a segment and technology level analysis. *Public Policy & Marketing*, 29(1), 2010.

[14] J. B. Earp, A. I. Antón, L. Aiman-Smith, and W. H. Stufflebeam. Examining internet privacy policies within the context of user privacy values. *Transactions on Engineering Management,*, 52(2):227–237, 2005.

[15] J. Gomez, T. Pinnick, and A. Soltani. Know privacy. Technical report, UC Berkeley School of Information, 2009. `http://knowprivacy.org/report/KnowPrivacy_Final_Report.pdf`.

[16] Y. Liu, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Analyzing facebook privacy settings: User expectations vs. reality. In *Proc. IMC '11*, pages 61–70. ACM, 2011.

[17] I. Altman. The environment and social behavior: Privacy, personal space, territory, and crowding. 1975.

[18] K. Martin. Understanding privacy online: Development of a social contract approach to privacy. *Journal of Business Ethics*, 137(3):551–569, 2016.

[19] H. Nissenbaum. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press, 2009.

[20] B. Krishnamurthy and C. Wills. Privacy diffusion on the web: A longitudinal perspective. In *Proceedings of the World Wide Web Conference(WWW)*. ACM, 2009.

[21] J. Turow, J. King, C. J. Hoofnagle, A. Bleakley, and M. Hennessy. Americans reject tailored advertising and three activities that enable it. *Available at SSRN 1478214*, 2009.

[22] N. Nikiforakis, A. Kapravelos, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna. Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. In *Security and Privacy (SP), 2013 IEEE Symposium on*, pages 541–555. IEEE, 2013.

[23] A. Soltani, S. Canty, Q. Mayo, L. Thomas, and C. J. Hoofnagle. Flash cookies and privacy. In *AAAI Spring Symposium: Intelligent Information Privacy Management*, 2010.

[24] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz. The web never forgets: Persistent tracking mechanisms in the wild. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pages 674–689. ACM, 2014.

[25] B. Krishnamurthy, K. Naryshkin, and C. Wills. Privacy leakage vs. protection measures: the growing disconnect. In *Proceedings of the Web 2.0 Security and Privacy Workshop*, volume 2, pages 1–10, 2011.

[26] Federal Trade Commission. Data brokers: A call for transparency and accountability. `http://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf`, May 2014.

[27] B. Ur, P. G. Leon, L. F. Cranor, R. Shay, and Y. Wang. Smart, useful, scary, creepy: Perceptions of online behavioral advertising. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*. ACM, 2012.

[28] L. Agarwal, N. Shrivastava, S. Jaiswal, and S. Panjwani. Do not embarrass: Re-examining user concerns for online tracking and advertising. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*. ACM, 2013.

[29] J. R. Reidenberg, N. C. Russell, A. J. Callen, S. Qasir, and T. B. Norton. Privacy harms and the effectiveness of the notice and choice framework. *ISJLP*, 11, 2015.

[30] J. Reidenberg, A. M. McDonald, F. Schaub, N. Sadeh, A. Acquisti, T. Breaux, L. F. Cranor, F. Liu, A. Grannis, J. T. Graves, et al. Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Technology Law Journal*, 30(1):39–88, 2015.

[31] F. Cate. The Limits of Notice and Choice. *IEEE Security & Privacy*, 8(2):59–62, March 2010.

[32] President's Council of Advisors on Science and Technology. Big data and privacy: A technological perspective. Technical report, Executive Office of the President, May 2014.

[33] R. M. Hogarth. *Judgement and Choice: The Psychology of Decision*. John Wiley & Sons, 1987.

[34] Council of The European Union. General Data Protection Regulation (GDPR). `http://data.consilium.europa.eu/doc/document/ST-5419-2016-INIT/en/pdf`, 2016.

[35] The Federal Trade Commission. FTC Announces Agenda for PrivacyCon 2017 [online]. 2017. URL: `https://www.ftc.gov/news-events/press-releases/2016/12/ftc-announces-agenda-privacycon-2017` [last checked 2017-09-08].

[36] F. Schaub, R. Balebako, A. L. Durity, and L. F. Cranor. A design space for effective privacy notices. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, pages 1–17. USENIX, 2015.

[37] Federal Trade Commission. Internet of things: Privacy & security in a connected world. FTC staff report, January 2015.

[38] Amazon. Alexa website rankings [online]. 2015. URL: `http://www.alexa.com` [last checked 2018-02-18].

[39] S. Wilson, F. Schaub, R. Ramanath, N. Sadeh, F. Liu, and N. A. Smith. Crowdsourcing annotations for websites' privacy policies: Can it really work? In *WWW*, 2016.

[40] J. Bhatia, T. D. Breaux, and F. Schaub. Mining privacy goals from privacy policies using hybridized task recomposition. *ACM Trans. Softw. Eng. Methodol.*, 25(3):22:1–22:24, May 2016. URL: `http://doi.acm.org/10.1145/2907942`, `doi:10.1145/2907942`.

[41] S. Zimmeck and S. M. Bellovin. Privee: An architecture for automatically analyzing web privacy policies. In *Proc. USENIX Security '14*, 2014.

[42] F. Liu, R. Ramanath, N. Sadeh, and N. A. Smith. A step towards usable privacy policy: Automatic alignment of privacy statements. In *Proc. COLING 2014*, pages 884–894, 2014.

[43] N. Sadeh, A. Acquisti, T. D. Breaux, L. F. Cranor, A. M. McDonald, J. R. Reidenberg, N. A. Smith, F. Liu, N. C. Russell, F. Schaub, et al. The usable privacy policy project. Technical report, CMU-ISR-13-119, Carnegie Mellon University, 2013.

[44] M. C. Gilly, W. L. Cron, and T. E. Barry. The expectations-performance comparison process: An investigation of expectation types. In *Proc. Conf. Consumer Satisfaction, Dissatisfaction, and Complaining Behavior*, pages 10–16, 1983.

[45] J. A. Miller. Studying satisfaction, modifying models, eliciting expectations, posing problems, and making meaningful measurements. *Proc. Conceptualization and Measurement of Consumer Satisfaction and Dissatisfaction*, pages 72–91, 1976.

[46] V. A. Zeithaml, L. L. Berry, and A. Parasuraman. The nature and determinants of customer expectations of service. *Academy of Marketing Science*, 21(1):1–12, 1993.

[47] J. S. Olson, J. Grudin, and E. Horvitz. A study of preferences for sharing and privacy. In *Proc. CHI '05*, pages 1985–1988. ACM, 2005.

[48] P. G. Leon, B. Ur, Y. Wang, M. Sleeper, R. Balebako, R. Shay, L. Bauer, M. Christodorescu, and L. F. Cranor. What matters to users? factors that affect users' willingness to share information with online advertisers. In *Proc. SOUPS '13*. ACM, 2013.

[49] A. Acquisti, L. Brandimarte, and G. Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, January 2015.

[50] Y. Wang, H. Xia, and Y. Huang. Examining american and chinese internet users' contextual privacy preferences of behavioral advertising. In *Proc. CSCW '16*. ACM, 2016. doi:10.1145/2818048.2819941.

[51] L. Palen and P. Dourish. Unpacking "privacy" for a networked world. In *Proc. CHI '03*, pages 129–136. ACM, 2003. doi:10.1145/642611.642635.

[52] T. Dinev and P. Hart. An extended privacy calculus model for e-commerce transactions. *Information systems research*, 17(1):61–80, 2006.

[53] K. Martin. Transaction costs, privacy, and trust: The laudable goals and ultimate failure of notice and choice to respect privacy online. 2013.

[54] SurveyMonkey. SurveyMonkey Audience [online]. 2017. URL: www.surveymonkey.com/mp/audience/ [last checked 2017-09-08].

[55] Bank of America. U.S. Online Privacy Notice [online]. 2017. URL: https://www.bankofamerica.com/privacy/online-privacy-notice.go [last checked 2017-09-08].

[56] PNC Bank. PNC Privacy Policy [online]. 2017. URL: https://www.pnc.com/en/privacy-policy.html [last checked 2017-09-08].

[57] G. B. Willis. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Sage Publications, 2004.

[58] E. Anduiza and C. Galais. Answering without reading: Imcs and strong satisficing in online surveys. *International Journal of Public Opinion Research*, 29(3):497–519, 2016.

[59] A. J. Berinsky, M. F. Margolis, and M. W. Sances. Separating the shirkers from the workers? making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, 58(3):739–753, 2014.

[60] S. Clifford and J. Jerit. Do attempts to improve respondent attention increase social desirability bias? *Public Opinion Quarterly*, 79(3):790–802, 2015.

[61] N. M. Bradburn, S. Sudman, and B. Wansink. *Asking questions: the definitive guide to questionnaire design–for market research, political polls, and social and health questionnaires*. John Wiley & Sons, 2004.

[62] S.-O. Leung. A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point likert scales. *Journal of Social Service Research*, 37(4):412–421, 2011.

[63] W. J. Conover. Practical nonparametric statistics. 1999.

[64] P. Wijesekera, A. Baokar, A. Hosseini, S. Egelman, D. Wagner, and K. Beznosov. Android permissions remystified: A field study on contextual integrity. In *USENIX Security Symposium*, pages 499–514, 2015.

[65] P. A. Norberg, D. R. Horne, and D. A. Horne. The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs*, 41(1):100–126, 2007.

[66] Office of the Australian Information Commissioner. Community attitudes to privacy survey, 2013.

[67] R. Kang, N. Fruchter, L. Dabbish, and S. Kiesler. "my data just goes everywhere:" user mental models of the internet and implications for privacy and security. In *Proc. SOUPS '15*. USENIX, 2015.

[68] L. Rainie, S. Kiesler, R. Kang, and M. Madden. Anonymity, privacy, and security online. *PEW Research Center*, September 2013.

[69] N. K. Malhotra, S. S. Kim, and J. Agarwal. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research*, 15(4):336–355, 2004.

[70] Amazon. Mechanical turk [online]. 2015. URL: `https://www.mturk.com/` [last checked 2018-02-18].

[71] G. Paolacci and J. Chandler. Inside the turk understanding mechanical turk as a participant pool. *Current Directions in Psychological Science*, 23(3):184–188, 2014.

[72] G. Paolacci, J. Chandler, and P. G. Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419, 2010.

[73] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5, 2011.

[74] M. S. Ackerman, L. F. Cranor, and J. Reagle. Privacy in e-Commerce: Examining User Scenarios and Privacy Preferences. In *Proc. EC '99*, pages 1–8. ACM, 1999.

[75] A. N. Joinson, U.-D. Reips, T. Buchanan, and C. B. P. Schofield. Privacy, Trust, and Self-Disclosure Online. *Human-Computer Interaction*, 25(1):1–24, February 2010.

[76] H. Nissenbaum. Privacy as contextual integrity. *Washington Law Review*, 79:119, 2004.

[77] L. F. Cranor, K. Idouchi, P. G. Leon, M. Sleeper, and B. Ur. Are they actually any different? comparing thousands of financial institutions' privacy practices. In *Proc. WEIS 2013*, 2013.

[78] E. M. Rogers. *Diffusion of innovations*. Simon and Schuster, 2010.

[79] S. Iyengar, K. S. Hahn, J. A. Krosnick, and J. Walker. Selective exposure to campaign communication: The role of anticipated agreement and issue public membership. *The Journal of Politics*, 70(1):186–200, 2008.

[80] W. P. Eveland and D. V. Shah. The impact of individual and interpersonal factors on perceived news media bias. *Political Psychology*, 24(1):101–117, 2003.

[81] B. Fung, C. Timberg, and M. Gold. A Republican contractor's database of nearly every voter was left exposed on the Internet for 12 days, researcher says [online]. 2017. URL: `https://www.washingtonpost.com/news/the-switch/wp/2017/06/19/republican-contractor-database-every-voter-exposed-internet-12-days-researcher-says/` [last checked 2018-02-19].

[82] S. Englehardt and A. Narayanan. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of ACM CCS 2016*, 2016.

[83] C. Vendil Pallin. Internet control through ownership: the case of russia. *Post-Soviet Affairs*, 33(1):16–33, 2017.

[84] BlueKai. Audience Data Marketplace [online]. 2014. URL: `https://www.bluekai.com/audience-data-marketplace.php` [last checked 2015-10-17].

[85] A. McDonald and L. Cranor. Beliefs and behaviors: Internet users' understanding of behavioral advertising. In *Proceedings of the Research Conference on Communication, Information and Internet Policy*. TPRC, 2010.

[86] T. Research. Consumer research results: Privacy and online behavioral advertising. July 2011.

[87] BlueKai. The BlueKai Registry [online]. 2014. URL: `http://bluekai.com/registry/` [last checked 2018-02-18].

[88] Google. Google ad settings [online]. 2014. URL: `http://www.google.com/settings/ads` [last checked 2018-02-18].

[89] Yahoo. Yahoo ad interests [online]. 2014. URL: `https://info.yahoo.com/privacy/us/yahoo/opt_out/targeting/details.html` [last checked 2018-02-18].

[90] Experian Marketing Services. List services catalog. `http://www.experian.com/assets/data-university/brochures/ems-list-services-catalog.pdf`, 2012.

[91] C. J. Hoofnagle, A. Soltani, N. Good, and D. J. Wambach. Behavioral advertising: The offer you can't refuse. *Harvard Law and Policy Review*, 2012.

[92] Acxiom. About the data [online]. 2014. URL: `https://aboutthedata.com/` [last checked 2018-02-18].

[93] Microsoft. Microsoft personalized ad preferences [online]. 2014. URL: `https://choice.microsoft.com/en-US/opt-out` [last checked 2018-02-18].

[94] Yahoo. Yahoo appended and matched data [online]. 2014. URL: `https://info.yahoo.com/privacy/us/yahoo/appenddata/`.

[95] V12group. Linkage database: Anonymous consumer recognition. http://www.v12groupinc.com/wp-content/uploads/2013/03/Linkage-Database-2014-.pdf, 2014.

[96] BlueKai. Little blue book: A buyers guide. http://www.BlueKai.com/bluebook/BlueKai-little-blue-book.pdf, 2013.

[97] V12group. Consumer selects. http://www.v12groupinc.com/consumer-selects/, 2014.

[98] V12group. Consumer in U.S. with postal address. http://listselector.com/docs/learnmoreConsumer%20in%20U.S.%20with%20Postal%20Address.pdf, 2014.

[99] BlueKai. Premium demographic data. http://www.BlueKai.com/bluebook/BlueKai-little-blue-book.pdf, 2013.

[100] V12group. PYCO personality database. http://www.v12groupinc.com/data/personality-data/, 2014.

[101] J. Rosen. The right to be forgotten. *Stanford law review online*, 64:88, 2012.

[102] E. M. Services. Ethnic insight. http://www.experian.com/assets/data-university/brochures/ems-list-services-catalog.pdf, 2012.

[103] Number of smartphones sold to end users worldwide from 2007 to 2016 (in million units) [online]. 2016. URL: https://www.statista.com/statistics/263437/global-smartphone-sales-to-end-users-since-2007/ [last checked 2018-02-25].

[104] Number of mobile app downloads worldwide in 2016, 2017 and 2021 (in billions) [online]. 2016. URL: https://www.statista.com/statistics/271644/worldwide-free-and-paid-mobile-app-store-downloads/ [last checked 2018-02-25].

[105] J. King. "how come i'm allowing strangers to go through my phone?" smartphones and privacy expectations. In *Workshop on Usable Privacy and Security for Mobile Devices (U-PriSM)*, 2012.

[106] R. Want, B. N. Schilit, and S. Jenson. Enabling the Internet of things. *Computer*, 48(1):28–35, 2015.

[107] Roland Berger. Bike Sharing 4.0 Study. Technical report, June 2016.

[108] S. Shaheen, S. Guzman, and H. Zhang. Bikesharing in europe, the americas, and asia: past, present, and future. *Transportation Research Record: Journal of the Transportation Research Board*, (2143):159–167, 2010.

[109] A. P. Felt, E. Ha, S. Egelman, A. Haney, E. Chin, and D. Wagner. Android permissions: User attention, comprehension, and behavior. In *Proceedings of the eighth symposium on usable privacy and security*, page 3. ACM, 2012.

[110] M. Van Kleek, I. Liccardi, R. Binns, J. Zhao, D. J. Weitzner, and N. Shadbolt. Better the devil you know: Exposing the data sharing practices of smartphone apps. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5208–5220. ACM, 2017.

[111] J. Zang, K. Dummit, J. Graves, P. Lisker, and L. Sweeney. Who knows what about me? a survey of behind the scenes personal data sharing to third parties by mobile apps. *Technology Science*, 2015.

[112] G. Romanillos, M. Zaltz Austwick, D. Ettema, and J. De Kruijf. Big data and cycling. *Transport Reviews*, 36(1):114–133, 2016.

[113] B. Berendt, O. Günther, and S. Spiekermann. Privacy in e-commerce: stated preferences vs. actual behavior. *Communications of the ACM*, 2005.

[114] K. Flüchter and F. Wortmann. Implementing the connected e-bike: challenges and requirements of an iot application for urban transportation. In *Proceedings of the First International Conference on IoT in Urban Space*, pages 7–12. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2014.

[115] A. Hasan, Q. Jiang, and C. Li. An effective grouping method for privacy-preserving bike sharing data publishing. *Future Internet*, 9(4):65, 2017.

[116] A. P. Felt, E. Chin, S. Hanna, D. Song, and D. Wagner. Android permissions demystified. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 627–638. ACM, 2011.

[117] S. Arzt, S. Rasthofer, C. Fritz, E. Bodden, A. Bartel, J. Klein, Y. Le Traon, D. Octeau, and P. McDaniel. Flowdroid: Precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for android apps. *Acm Sigplan Notices*, 49(6):259–269, 2014.

[118] Z. Ma, H. Wang, Y. Guo, and X. Chen. Libradar: fast and accurate detection of third-party libraries in android apps. In *Proceedings of the 38th International Conference on Software Engineering Companion*, pages 653–656. ACM, 2016.

[119] A. Cortesi, M. Hils, T. Kriechbaumer, and contributors. mitmproxy: A free and open source interactive HTTPS proxy, 2010–. [Version 3.0]. URL: `https://mitmproxy.org/`.

[120] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3:1376, 2013.

[121] S. Zimmeck, Z. Wang, L. Zou, R. Iyengar, B. Liu, F. Schaub, S. Wilson, N. Sadeh, S. M. Bellovin, and J. Reidenberg. Automated analysis of privacy requirements for mobile apps. In *Proceedings of the Network and Distributed System Security (NDSS) Symposium*, 2017.

[122] oBike FAQ [online]. 2018. URL: `https://www.o.bike/de/faqs/` [last checked 2018-02-26].

[123] Verizon IoT, Swiftmile pilot smart bike-sharing in Santa Clara [online]. 2016. URL: `https://www.computerworld.com/article/3143132/internet-of-things/verizon-iot-swiftmile-pilot-smart-bike-sharing-in-santa-clara.html` [last checked 2018-02-27].

[124] We Are Here to Help You Connect with an Ecosystem [online]. 2018. URL: `https://www.appsflyer.com/product/mobile-ecosystem/` [last checked 2018-02-28].

[125] P. G. Kelley, J. Bresee, L. F. Cranor, and R. W. Reeder. A nutrition label for privacy. In *Proc. SOUPS '09*. ACM, 2009.

[126] L. F. Cranor. Necessary but Not Sufficient: Standardized Mechanisms for Privacy Notice and Choice. *Journal on Telecommunications and High Technology Law*, 10:273, 2012.

A Survey Questionnaire

A.1 Survey: Types of Privacy Expectations

Questionnaire starts on the next page.

[Informed Consent]

This is a research study conducted by an academic university. The purpose of this study is to understand your opinions regarding websites. You must be 18 years or older to participate, and your participation is voluntary.

As part of the study, you will take a survey that will last about 10 minutes. It contains questions regarding your opinions and usage of websites, and your background.

Your answers are anonymous. **We will NOT collect personal information, IP address or other personally-identifiable information**. We receive the following demographic information from SurveyMonkey: age range, gender, type of device, annual household income and region. The risks associated with participation in this study include boredom and fatigue, and are not greater than those ordinarily encountered in daily life or online activities.

Do you understand the information above, are 18 years or older, and want to continue with the survey?
☐ Yes, I understand the information above, am 18 years or older, and want to continue with the survey
☐ No, do not continue with the survey

[Survey Instructions]

Thank you for your interest in our survey. Your answers are important to us.

Please read the following instructions carefully:
Take your time in reading and answering the questions.
Answer the questions as accurately as possible.
It is OK to say that you don't know an answer.

[Pre questionnaire]
In this survey, we would like to understand your opinions regarding websites.

First, we will ask a few questions about websites that you might use.

[To minimize order effect, randomize the order of health and banking questions]

Some people use websites such as WebMD, MedlinePlus or MedicineNet to find information on health conditions, symptoms or treatments. Other people do not do so.
As far as you can recall, have you ever used websites such as WebMD, MedlinePlus or MedicineNet to find information on health conditions, symptoms or treatments?
() Yes
() No
() Don't know/Not sure
() Decline to answer

[Omit next question if not answered "Yes" to the previous question]

Please think about the last time you used a website such as WebMD, MedlinePlus or MedicineNet to find information on health conditions, symptoms or treatments. Try to recall the information that you were trying to find.
Do you recall what information you were trying to find?
() Yes
() No
() Don't know/Not sure
() Decline to answer

Some people use websites to check their Checking/Savings account balance. Other people do not do so.
As far as you can recall, have you ever used a website to check your Checking/Savings account balance?
() Yes
() No
() Don't know/Not sure
() Decline to answer

Some people have a Checking/Savings account. Other people do not.
Do you currently have a Checking/Savings account?
() Yes
() No
() Don't know/Not sure
() Decline to answer

[Omit next question if not answered "Yes" to the previous question]

To understand whether opinions regarding banking websites vary based on the length of time a person has had a Checking/Savings account, we would like to know the length of time you have had a Checking/Savings account.

As far as you can recall, approximately in which year did you open the Checking/Savings account that you currently have? If you currently have multiple accounts, consider the Checking/Savings account that you opened earliest.

() Decline to answer
Year of opening the account (4-digit, yyyy format): _____

[Main questionnaire]

Now, we would like to understand your opinions regarding websites.

Imagine a scenario where you are a customer of a bank, and you have a Checking/Savings account with the bank. In this scenario, tell us how much you **agree or disagree** with the statements below. Use a scale from 0 to 10, with 0 indicating "strongly disagree" and 10 indicating "strongly agree."

Below, *health-related browsing activity* refers to browsing activities on websites such as WebMD, MedlinePlus or MedicineNet, which you might use/visit to find information on health conditions, symptoms or treatments.

[To minimize order effect, reverse the order of the 4 statements for half of the participants]

I want my bank to collect my health-related browsing activity to identify my financial needs and provide service relevant to me.

strongly disagree										**strongly agree**	Don't know/ Not sure	Decline to answer
0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**		

I think my bank will collect my health-related browsing activity to identify my financial needs and provide service relevant to me.

strongly disagree										**strongly agree**	Don't know/ Not sure	Decline to answer
0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**		

I deserve that my bank collect my health-related browsing activity to identify my financial needs and provide service relevant to me.

strongly disagree										**strongly agree**	Don't know/ Not sure	Decline to answer
0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**		

I would tolerate if my bank must collect my health-related browsing activity to identify my financial needs and provide service relevant to me.

strongly disagree										**strongly agree**	Don't know/ Not sure	Decline to answer
0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**		

[Post questionnaire]

[We get age range, household income, gender and US location for each panelist from SurveyMonkey]

You are almost done. Please tell us briefly about your background.

What is the highest level of school you have completed or the highest degree you have received?
() Less than high school (Grades 1-8 or no formal schooling)
() High school incomplete (Grades 9-11 or Grade 12 with NO diploma)
() High school graduate (Grade 12 with diploma or GED certificate)
() Some college, no degree (includes some community college)
() Two-year associate degree from a college or university
() Four-year college or university degree/Bachelor's degree (e.g., BS, BA, AB)
() Some postgraduate or professional schooling, no postgraduate degree (e.g. some graduate school)
() Postgraduate or professional degree, including master's, doctorate, medical or law degree (e.g., MA, MS, PhD, MD, JD, graduate school)
() Decline to answer

Please give us your feedback (optional).

If you had difficulty answering any of the survey questions, briefly describe what about the question made it difficult to answer (optional).

We are done with our questions. Anything you care to add? (optional).

Thank you very much for helping us with our study.

A.2 Survey: Identifying Mismatched Privacy Expectations

Questionnaire starts on the next page.

[Interview/Survey Questionnaire]

Thank you for your interest in our study.

Your answers are important to us. Please read the instructions carefully so that you can answer our questions as accurately as possible. Take your time in reading and answering the questions.

Peoples' opinions about websites may or may not vary depending on the type of website (news, health, finance etc.) and past experience (not heard of website, heard of, not visited, visited etc.)
While answering questions about a website, **think about your interactions only with the website**. Your interactions could be through a computer, mobile phone or other device. Ignore any interactions with mobile apps, physical stores, businesses or other websites related to the website.
For each website listed below, select the option that best indicates your answer.

	I have not heard of it	I have heard of it, but not visited it	I have visited it, but not in the last 3 months	I have visited it in the last 3 months	Don't know/Not sure
[website]	○	○	○	○	○

I would like to understand your opinions regarding Internet websites. For any question, it is okay to say that you don't know the answer. If you are guessing an answer, please say so. It would be very helpful, if you explain your reasoning behind your answers.

[For each website assigned to a participant, ask the following questions]

Now, I would like your opinions regarding [website name] website. Please interact with the website (provide URL) for 2-3 minutes and get familiar with it. Please let me know when you are ready to provide your opinions.

1. As far as you can recall, have you used any websites similar to [website name]?
   Yes (please specify) / No

[Omit questions 2 and 3 if the participant has not used the website]

2. I would like you to think about the last time you visited [website name]. As far as you can recall, what did you do on the website?
3. What other things have you done on this website?

To help you answer my questions, I will explain a few terms. Please use this handout to follow along. You can refer back to the handout at any time.

[Provide handout containing definitions for contact/health/financial/current location information]
[Read definitions for contact/health/financial/current location information]

4. Consider the following scenario to answer the next question.

   Imagine that you are browsing [website name] website. You **do not have a user account** on [website name], that is, you have not registered or created an account on [website name].

   What is the likelihood that [website name] would **collect your information** in this scenario? Each row in the table below, lists a specific type of information about you. For each information type, select the likelihood that [website name] would collect that information in the scenario described above.

		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Collects your **Contact information**	Email address	○	○	○	○
	Postal address	○	○	○	○
	Phone number	○	○	○	○
	Other Please specify	○	○	○	○
		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Collects your **Health information**	Medical history	○	○	○	○
	Health insurance information	○	○	○	○
	Other Please specify	○	○	○	○

		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Collects your **Financial information**	Bank account details	O	O	O	O
	Credit or debit card number	O	O	O	O
	Credit rating	O	O	O	O
	Other Please specify	O	O	O	O
		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Collects your **Location information**	Current location (city-level or more precise)	O	O	O	O

5. What leads you to think that [website name] would collect your information when you do not have an account? Please explain.
6. Now, consider an alternate scenario.

   Imagine that **you have a user account** on [website name], and you **have logged in** to your account while browsing [website name].

   What is the likelihood that [website name] would **collect your information** in this scenario?
   Each row in the table below, lists a specific type of information about you. For each information type, select the likelihood that [website name] would collect that information in the scenario just described.

		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Collects your **Contact information**	Email address	O	O	O	O
	Postal address	O	O	O	O
	Phone number	O	O	O	O
	Other Please specify	O	O	O	O
		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Collects your **Health information**	Medical history	O	O	O	O
	Health insurance information	O	O	O	O
	Other Please specify	O	O	O	O
		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Collects your **Financial information**	Bank account details	O	O	O	O
	Credit or debit card number	O	O	O	O
	Credit rating	O	O	O	O
	Other Please specify	O	O	O	O
		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Collects your **Location information**	Current location (city-level or more precise)	O	O	O	O

Thank you. As you may know, companies that own websites may handle information collected on websites in different ways. Some companies share the collected information with other companies, and some companies do not share. Companies may have to share your information in order to provide you a service that you requested on a website.

7. In your opinion, what services can you get from [website name]? Please explain.

8. In order to provide you services, [website name] may have to share your information with other companies. In your opinion, what are those companies, if at all any? Please explain.

9. A website may share your information for purposes unrelated to providing you a service that you requested from the website. What do you think are such unrelated purposes for which [website] can share your information? Please explain.

Before sharing your information, companies may or may not ask for your permission. Some companies assume that the permission is implied because you are using the website. Other companies may explicitly ask you for permission before sharing information, for example, via an explicit written or oral consent.

10. Consider the following scenario to answer the next question.

Imagine that [website name] is sharing your information with another company, but **only for the purpose of providing you a service you requested** on [website name]. Since [website name] has to provide you a service that you requested, [website name] assumes that it has your permission to share information, that is, your permission is implied. [Website name] will share only the information required to provide you the requested service.

What is the likelihood that [website name] would **share your information** with your implied permission in this scenario?

		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Shares your **Contact information**	Email address	O	O	O	O
	Postal address	O	O	O	O
	Phone number	O	O	O	O
	Other Please specify	O	O	O	O
		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Shares your **Health information**	Medical history	O	O	O	O
	Health insurance information	O	O	O	O
	Other Please specify	O	O	O	O
		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Shares your **Financial information**	Bank account details	O	O	O	O
	Credit or debit card number	O	O	O	O
	Credit rating	O	O	O	O
	Other Please specify	O	O	O	O
		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Shares your **Location information**	Current location (city-level or more precise)	O	O	O	O

11. Consider the following alternate scenario to answer the next question.

Imagine that [website name] is sharing your information with another company for **a purpose unrelated to providing you a service you requested**. Since you are using [website name], it assumes that it has your permission, that is implied permission, to share your information for any purpose.

What is the likelihood that [website name] would **share your information** with your implied permission in this scenario?

		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Shares your **Contact information**	Email address	O	O	O	O
	Postal address	O	O	O	O
	Phone number	O	O	O	O
	Other Please specify	O	O	O	O
		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Shares your **Health information**	Medical history	O	O	O	O
	Health insurance information	O	O	O	O
	Other Please specify	O	O	O	O
		Likely	Somewhat	Somewhat	Unlikely

			likely	unlikely	
Shares your **Financial information**	Bank account details	○	○	○	○
	Credit or debit card number	○	○	○	○
	Credit rating	○	○	○	○
	Other Please specify	○	○	○	○

		Likely	Somewhat likely	Somewhat unlikely	Unlikely
Shares your **Location information**	Current location (city-level or more precise)	○	○	○	○

Thank you. As you may know, websites may allow users to delete or remove their data from the website e.g. by closing an account. Allowing users to edit or modify their data is not same as deleting data.

12. Do you think that [website name] would allow you to delete your personal data?
    - ○ Yes, it will allow me to delete all of my data
    - ○ Yes, but it will only allow me delete some of my data
    - ○ No, it will not allow me to delete my data

13. We discussed data practices such as collection and sharing of four types of information, and also deletion of information. What else would you like to know about [website name]?

[End of the interview]

Thank you. That was all I had to discuss. Would you care to add anything?

Thank you. Please take a few minutes to fill out the following questionnaire. That would be the end of our study.

Different users may have different opinions regarding websites. To help us understand how user opinions vary, please answer the following questions.

Please tell us about your experience with [website name] website.

As far as you know, do you have a user account on the website?
( ) Yes, I have an account
( ) No, I don't have an account
( ) Not sure

How many times have you visited the website in the last 30 days? Exclude the visit as part of today's study.
(Please specify a number equal to or greater than zero) _____

In your opinion, how much have you used the website in the last 30 days? Exclude use as part of today's study.

1 – Not at all	2 – Very little	3 - Somewhat	4 – Quite a bit	5 – A great deal
○	○	○	○	○

Do you know someone else who uses the website?
( ) Yes, I know someone
( ) No, I don't know anyone
( ) Not sure

In your opinion, how familiar are you with the website?

1 – Not at all	2 – Slightly	3 - Somewhat	4 – Moderately	5 – Extremely
○	○	○	○	○

In your opinion, how trustworthy is the website?

1 – Not at all	2 – Slightly	3 - Somewhat	4 – Moderately	5 – Extremely
○	○	○	○	○

As far as you know, do you have a user account on a website similar to [website name]?
( ) Yes, I have an account
( ) No, I don't have an account
( ) Not sure

Please tell us about your background.

What is your year of birth (4-digit, yyyy format)?
_____

What is your gender?
( ) Male ( ) Female ( ) Decline to answer

Which of the following best describes your primary occupation?
[List of occupations here]

Which of the following best describes your highest achieved education level?
[List of education levels here]

Do you have a college degree or work experience in computer science, software development, web development or similar computer-related fields?
( ) Yes ( ) No ( ) Decline to answer

Do you currently work or reside in the state of California?
( ) Yes ( ) No ( ) Decline to answer

While using the Internet, have you ever done any of the following things? Please check all that apply.
[ ] Used a temporary username or email address
[ ] Used a fake name or untraceable username
[ ] Given inaccurate or misleading information about yourself
[ ] Set your browser to disable or turn off cookies
[ ] Cleared cookies and browser history
[ ] Used a service that allows you to browse the web anonymously, such as a proxy server, Tor software, or a virtual private network
[ ] Encrypted your communications
[ ] Decided not to use a website because they asked for your real name
[ ] Deleted or edited something you posted in the past
[ ] Asked someone to remove something that was posted about you online
[ ] Used a public computer to browse anonymously

How would you rate your familiarity with the following concepts or tools?

	I've never heard of this.	I've heard of this but I don't know what it is.	I know what this is but I don't know how it works.	I know generally how this works.	I know very well how this works.
IP address	o	o	o	o	o
Cookie	o	o	o	o	o
Incognito mode / private browsing mode in browsers	o	o	o	o	o
Encryption	o	o	o	o	o
Proxy server	o	o	o	o	o
Secure Sockets Layer (SSL)	o	o	o	o	o
Tor	o	o	o	o	o
Virtual Private Network (VPN)	o	o	o	o	o
Privacy settings	o	o	o	o	o

Please indicate whether you think each statement is true or false. Please select "I'm not sure" if you don't know the answer.

	True	False	I'm not sure
Incognito mode / private browsing mode in browsers prevents websites from collecting information about you.	o	o	o
Website cookies can store users' logins and passwords in your web browser.	o	o	o
Tor can be used to hide the source of a network request from the destination	o	o	o
A VPN is the same as a Proxy server.	o	o	o
IP addresses can always uniquely identify your computer.	o	o	o
HTTPS is standard HTTP with SSL to preserve the confidentiality of network traffic.	o	o	o
A request coming from a proxy server cannot be tracked to the original source.	o	o	o

In order to protect your personal information, how often have you done the following?

Check that a website is secure before providing personal information.

1 - Never	2 - Rarely	3 - Sometimes	4 - Often	5 - Always
o	o	o	o	o

Ask public or private sector organizations why they need your information.

1 - Never	2 - Rarely	3 - Sometimes	4 - Often	5 - Always
o	o	o	o	o

Read privacy policies and notifications before providing personal information.

1 - Never	2 - Rarely	3 - Sometimes	4 - Often	5 - Always
o	o	o	o	o

As far as you know, have you ever had any of these bad experiences as a result of your online activities?

	Yes	No
Something happened online that led you into physical danger	o	o
Been stalked or harassed online (sexually harassed, physically threatened)	o	o
Got into trouble with local authorities, or government because of your online activities	o	o
Experienced trouble in a relationship between you and a family member or a friend because of something you posted online	o	o
Had your personal information leaked by a company	o	o
Lost a job opportunity or educational opportunity because of something you posted online or someone posted about you online	o	o
Had your reputation damaged because of something that happened online	o	o
Been the victim of an online scam and lost money	o	o
Had important personal information stolen such as your Social Security Number, your credit card, or bank account information	o	o
Something else bad happened (please explain)	o	o

You are almost done. Please share your opinion about Internet consumer experience.
Please indicate how much you agree or disagree with the following statements:

Consumer online privacy is really a matter of consumers' right to exercise control and autonomy over decisions about how their information *is collected, used, and shared.*

Strongly disagree 1 2 3 4 5 6 7 **Strongly agree**

Consumer control of personal information lies at the heart of consumer privacy.

Strongly disagree 1 2 3 4 5 6 7 **Strongly agree**

I believe that online privacy is invaded when control is lost or unwillingly reduced as a result of a marketing transaction.

Strongly disagree 1 2 3 4 5 6 7 **Strongly agree**

Companies seeking information online should disclose the way the data are collected, processed, and used.

Strongly disagree 1 2 3 4 5 6 7 **Strongly agree**

A good consumer online privacy policy should have a clear and conspicuous disclosure.

Strongly disagree 1 2 3 4 5 6 7 **Strongly agree**

It is very important to me that I am aware and knowledgeable about how my personal information will be used.

Strongly disagree 1 2 3 4 5 6 7 **Strongly agree**

It usually bothers me when online companies ask me for personal information.

Strongly disagree 1 2 3 4 5 6 7 **Strongly agree**

When online companies ask me for personal information, I sometimes think twice before providing it.

Strongly disagree 1 2 3 4 5 6 7 **Strongly agree**

It bothers me to give personal information to so many online companies.

Strongly disagree 1 2 3 4 5 6 7 **Strongly agree**

I'm concerned that online companies are collecting too much personal information about me.

Strongly disagree 1 2 3 4 5 6 7 **Strongly agree**

Thank you for participating in our study.

A.3 Survey: User Concerns Regarding Behavioral Profiles

Questionnaire starts on the next page.

Appendix A: Online Survey Questionnaire
[Consent instructions here]

Important: Please think thoroughly before answering each question. Your precise responses are very important for us. We are not interested in what someone else thinks - we want to know what you think! You may give an incomplete answer or say you do not know.

1) We are interested in understanding how you experience things online. We will start by seeking your views about website advertising. Here, "website advertising" refers to ads that are displayed on the web pages that you visit. In a sentence or two, please tell us what you think about website advertising.*

2) What is your age (in years)?*

3) What is your gender?*
( ) Male ( ) Female ( ) Decline to answer

4) Which of the following best describes your primary occupation?*
[List of options here]
( ) Other (Please specify):: _____*
( ) Decline to answer

5) Which of the following best describes your highest achieved education level?*
( ) No high school
( ) Some high school
( ) High school graduate
( ) Some college - no degree
( ) Associates/2 year degree
( ) Bachelors/4 year degree
( ) Graduate degree - Masters, PhD, professional, medicine, etc.
( ) Decline to answer

6) Do you have a college degree or work experience in computer science, software development, web development or similar computer-related fields?*
( ) Yes ( ) No ( ) Decline to answer

7) Advertisers can personalize ads on websites to ensure that the ads are relevant to you.
Please indicate how much you agree or disagree with the following statement.

I like personalization of ads on websites.*
( ) Strongly disagree ( ) Disagree ( ) Neither agree nor disagree ( ) Agree ( ) Strongly agree

Advertisers collect data about you in order to personalize ads. Advertisers may create profiles about you using the collected data.

The following is an image of a profile that shows the different types of data that advertisers collect about users like you. Please look through the entire image at your own pace, and then answer the following questions.

[Sample profile figure]

8) Please select from the list below at least two items that appear in the sample profile.*
[ ] Male
[ ] Credit Card Interest Score 8-9%
[ ] Offline CPG Purchasers > Charmin Ultra Strong
[ ] Personal Health (Values 70-90%)
[ ] Interest in Religion Code (Value Tiers 1-3)
[ ] Household Income (HHI) > Income Range $75,000 - $99,000

[Randomize Q9 – Q13]
We will ask you some questions to understand your reaction to the profile you just saw. It is important that you have looked at the different types of data in the profile before continuing. Please click next when you are ready.

9) Please describe your reaction to the profile. [Sample profile figure shown here]
Indicate how much you agree or disagree with the following statement.
I am concerned because I believe that the profile contains sensitive data.*
( ) Strongly disagree ( ) Disagree ( ) Neither agree nor disagree ( ) Agree ( ) Strongly agree

10) Please describe your reaction to the profile. [Sample profile figure shown here]
Indicate how much you agree or disagree with the following statement.
I am concerned by the amount of data in the profile.*
( ) Strongly disagree ( ) Disagree ( ) Neither agree nor disagree ( ) Agree ( ) Strongly agree

11) Please describe your reaction to the profile. [Sample profile figure shown here]
Indicate how much you agree or disagree with the following statement.
I am concerned because my data from multiple sources (e.g. online activities, in-store, other companies) is being combined.*
( ) Strongly disagree ( ) Disagree ( ) Neither agree nor disagree ( ) Agree ( ) Strongly agree

12) Please describe your reaction to the profile. [Sample profile figure shown here]
Indicate how much you agree or disagree with the following statement.
I am concerned by the level of detail (e.g. specific information, not just broad categories) in the profile. *
( ) Strongly disagree ( ) Disagree ( ) Neither agree nor disagree ( ) Agree ( ) Strongly agree

13) Please describe your reaction to the profile. [Sample profile figure shown here]
Indicate how much you agree or disagree with the following statement.
I am concerned about how my data may be used.*
( ) Strongly disagree ( ) Disagree ( ) Neither agree nor disagree ( ) Agree ( ) Strongly agree

14) Please explain if you have other concerns about the profile.

You are almost done.

We will ask you how you feel about personalized ads after seeing the profile. We will also give you a chance to look at your own profile. Please note that looking at your own profile is optional.

15) Please indicate how much you agree or disagree with the following statement.
After seeing the types of data collected for personalization, my liking for personalized ads on websites has decreased.*
( ) Strongly disagree ( ) Disagree ( ) Neither agree nor disagree ( ) Agree ( ) Strongly agree

You looked at a sample profile. Would you like to look at your own profile and learn what data advertisers have about you?

Please note that this is optional. Your payment and bonus will not be affected if you choose to skip looking at your own profile. However, what you learn may be beneficial to you.

16) Would you like to look at your own profile?*
( ) Yes ( ) No

Thank you for choosing to look at your own profile. We believe it will be beneficial to you and us.

Please copy and paste the following website link in a new tab or window to access your own profile. You should see a profile similar in appearance to the sample profile.
http://bluekai.com/registry/

Please note that the profile may not display properly if you have disabled browser cookies. You can try from a different browser if you have more than one browser installed.

If you are not able to access your profile using the above link, you can alternatively try the following websites.
https://aboutthedata.com/ (scroll to the bottom of the page to click on "See and Edit Marketing Data about You.")
http://www.google.com/settings/ads
http://info.yahoo.com/privacy/us/yahoo/opt_out/targeting/

17) Please tell us briefly what you found in your own profile and how you feel about it (optional but helpful for our research).

18) Do you have any further comments?
Thank you for taking our survey. Your response is important to us. Below is your confirmation code. You must retain this code to be paid - it is recommended that you store your code in a safe place (either by writing it down, or by printing this page).

A.4 Exercise: Privacy of Bike Sharing Apps

Questionnaire starts on the next page.

This is an **individual** exercise. Each student needs to do the exercise on their own.

Privacy and Mobile Apps (5 + 10 + 30 + 10 = 55 points)

Analyze the privacy of bike sharing apps in two major cities in the EU. You can pick any two major cities that you find interesting. For example, you can select two cities from the same country, from two different countries, from two distinct regions within a country etc. Enter the list of cities on Moodle. Each student needs to analyze a distinct set i.e. two or more students cannot analyze the same city.

a. List the two cities that you selected. Briefly explain your reasoning or hypothesis, if any, for selecting the cities.

b. For each city, identify all the bike sharing apps available in the city. List the apps in a tabular format similar to the one below. If there are any bike sharing programs that do not have a mobile app, enter them as "[Name of the program] (no app)."

City 1	**City 2**
App 1 App 2 App 3 …	App 1 App 2 …

c. For each city do the following.

   Identify and compare the permissions used by the apps by listing them in a tabular format similar to the one below.

   City X

App 1	**App 2**	**App 3**	**…**
Permission 1	Permission 1	Permission 1	…
Permission 2	Permission 2	…	…
…	…		

   Identify why the app may require each permission.

   For each app, which permissions do you find sensitive or unusual, and why?

   Do the apps have a separate privacy policy? If yes, do the app permissions align with the data practices disclosed in the privacy policy?

   What is the business model of the bike sharing program that owns the app? In your opinion, does it influence the permissions requested by the app?

d. In your opinion, are there any interesting differences between the apps from the two different cities? Explain.