Awareness of Behavioral Tracking and Information Privacy Concern in Facebook and Google

Emilee Rader
Department of Media and Information
College of Communication Arts and Sciences
Michigan State University
emilee@msu.edu

ABSTRACT

Internet companies record data about users as they surf the web, such as the links they have clicked on, search terms they have used, and how often they read all the way to the end of an online news article. This evidence of past behavior is aggregated both across websites and across individuals, allowing algorithms to make inferences about users' habits and personal characteristics. Do users recognize when their behaviors provision information that may be used in this way, and is this knowledge associated with concern about unwanted access to information about themselves they would prefer not to reveal? In this online experiment, the majority of a sample of web-savvy users was aware that Internet companies like Facebook and Google can collect data about their actions on these websites, such as what links they click on. However, this awareness was associated with lower likelihood of concern about unwanted access. Awareness of the potential consequences of data aggregation, such as Facebook or Google knowing what other websites one visits or one's political party affiliation, was associated with greater likelihood of reporting concern about unwanted access. This suggests that greater transparency about inferences enabled by data aggregation might help users associate seemingly innocuous actions like clicking on a link with what these actions say about them.

1. INTRODUCTION

In February 2012, the New York Times published an article describing how the Target Corporation uses "predictive analytics" to find patterns in personal information about customers and their behavior, that has been collected first-hand by Target or purchased from third parties [10]. The article continues to be frequently mentioned because of a (perhaps apocryphal) anecdote about a father who found out that his teenage daughter was pregnant, by looking through the coupons she received from Target via the US postal service. Over the past few years, this example has been used by many as a warning about the future of information privacy, because it illustrates how behavioral data that is collected without a person's knowledge as they interact with systems in their daily lives (here, purchase records from Target) can be used to infer intimate details that one might prefer not to disclose.

Most web pages include code that users cannot see, which collects data necessary for making predictive inferences about what each individual user might want to buy, read, or listen to[1]. This data ranges from information users explicitly contribute, such as profile information or "Likes" on Facebook, to behavioral traces like GPS location and the links users click on, to inferences based on this data such as gender and age [15], sexual orientation [18], and whether or not one is vulnerable to depression [7].

Whether or not users explicitly intended to provide the information, once it has been collected it is not just used to reflect users' own likes and interests back through targeted advertisements. Algorithms use this data to turn users' likenesses into endorsements—messages displayed to other users that associate names and faces with products and content they may not actually want to endorse [31, 32]. Algorithms make inferences about who we are, and present that information on our behalf to other people and organizations.

Internet users express discomfort with data collection that enables personalization. For example, a recent Pew survey found that "73% of search engine users say they would NOT BE OK [sic] with a search engine keeping track of searches and using that information to personalize future search results, because it is an invasion of privacy" [28]. Eighty-six percent of Internet users have taken some kind of action to be more anonymous when using the web—most often, clearing cookies and browser history [30].

Nevertheless, people use search engines and social media on a daily basis, and simple browser-based strategies like deleting cookies and browsing history are not enough to protect one's information online. For example, the configuration of plugins and add-ons of a particular web browser on a specific machine comprises a unique "fingerprint" that can be traced by web servers across the web, and this information is conveyed through headers that are automatically exchanged by every web browser and web server behind the scenes [25].

It is clear that users are concerned about online privacy, and that transparency—especially regarding what can be inferred about users based on seemingly innocuous data like clicking a link in a web page—is lacking. What, then, are the disclosures that users actually do know about, and how is this awareness related to privacy concern? The goal of this research was to investigate whether users recognize that their behaviors provision information which may be used by personalization and recommendation algorithms to infer things about them, and if this awareness is associated with privacy concern.

I found that a sample of web-savvy users were resoundingly aware that Internet companies like Facebook and Google can col-

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9–11, 2014, Menlo Park, CA.

[1] https://www.eff.org/deeplinks/2009/09/online-trackers-and-social-networks

lect data about their behaviors on those websites, consisting of things like when and how often they visit those sites, and what links they click on. I refer to information like these examples as *First Party Data*, because it can be collected directly from user actions with websites. However, greater awareness of the collection of First Party Data was associated with a LOWER likelihood of concern about unwanted access to private information.

Participants were much less aware of automatic collection of personal information produced by aggregation across websites, which can reveal patterns such as one's purchase habits, or aggregation across users, which can reveal potentially sensitive information like sexual orientation. But unlike First Party Data, those users who had greater awareness of either kind of aggregation had a GREATER likelihood of concern about unwanted access. This suggests that a solution involving informed consent about collection of First Party Data would not support better boundary management online, and that different approaches are needed to make the consequences of aggregation, rather than the disclosures themselves, more transparent.

2. RELATED WORK

2.1 Boundary Management Online

People interact with one another in contexts structured by the roles they assume and the activities they engage in; by the social norms of the situation; by their own objectives and goals; and even by aspects of the architecture of the physical world [26]. Westin [42] defined privacy as "the claim of an individual to determine what information about himself or herself should be known to others", and all of these factors contribute to people's assessments of what information they want to allow others to know in what context.

While there are many structural aspects of offline physical and social contexts that help people negotiate boundaries between public and private, managing boundaries when sharing information online is more difficult. Social media systems, in particular, suffer from "context collapse": users have multiple audiences for their posts with whom they might want to share different sets of information, but it can be difficult to understand which part of one's potential audience is able to see the content [12], or is even paying attention [29]. Stutzman and Hartzog [39] conducted an interview study of users with multiple social network profiles, who used profiles on different systems to manage boundaries and disclosures. They sometimes kept the profile identities completely separate, and other times they strategically or purposefully linked them to create boundaries between audiences with which they shared different degrees of intimacy. Different systems have implemented interface mechanisms and controls for specifying the boundaries between audiences, but no industry best practices or standards seem to exist for interfaces to manage access to one's personal information [4]. For example, Bonneau and Preibusch reported that at the time of their research, only two out of 45 social network sites (Facebook and LinkedIn) offered users the capability to see what their profile looked like to users with different levels of access.

Users don't always change privacy settings and mechanisms from the defaults, and even when they do, they aren't always successful at achieving their desired result. Liu et al. [21] designed a Facebook app to collect 10 photos from participants' Facebook accounts, along with the visibility setting associated with each photo. They also asked each user to indicate who their desired audience was for each photo. They found that 36% of the photos were shared with the default—fully public—setting, while participants indicated only 20% of the photos should have been public. In an experiment, Egelman et al. [11] presented users with different information sharing scenarios in Facebook and asked to specify access control polities. They found that when users made mistakes—when their desired level of access did not match what they specified through the system—they erred on the side of revealing more broadly than they wanted to.

In systems that do not provide privacy mechanisms, users express discomfort about what others might infer about them by learning about characteristics of the content they consume. Personalized content can reveal potentially embarrassing information to others [40]. For example, Silfverberg et al. [33] studied the social music service Last.fm and found that participants reported making personal judgments about other users based on their music preferences. Music has an emotional quality, and participants worried that allowing others to know what music they were listening to might reveal information about what they were feeling that they might not want to disclose. At that time, Last.fm did not allow users to protect any of the information in their profile, so the only recourse they had was to create separate profiles for different audiences.

Some users also express concern about the possibility that behavioral advertising might reveal private information about them based on past web browsing sessions. After having behavioral advertising explained to them, 41 out of 48 participants in one study felt concerned about what they perceived as a loss of control over their information [41]. A majority of participants in another study reported that they had been embarrassed in the past by advertising that appeared on a web page they were viewing, that was also seen by another person in the vicinity (e.g., "what were you browsing last night") [1]. These examples each illustrate circumstances where data collected for personalization has made it more difficult for users to manage the boundary between information they do and do not want to reveal.

2.2 Information vs. Social Privacy

There is an important distinction between *social privacy* and *information privacy*. "Social privacy" concerns how we manage self-disclosures, availability, and access to information about ourselves by other people. "Information privacy" refers to the control of access to personal information by organizations and institutions, and the technologies they employ to gather, analyze, and use that information for their own ends [36].

Privacy settings in most online systems are designed to manage social privacy, and people are willing to take steps to enforce social boundaries online when such options are available [16]. For example, people who are more concerned about information privacy reported using privacy management tools more, according to Litt [20] who analyzed a Pew Internet & American Life data set from 2010. However, people may not perceive a connection between social privacy and threats to information privacy. Strategies such as specifying one's privacy settings and maintaining multiple profiles allow users control over social privacy, but they do not support better control over information privacy, because the architectures and algorithms that collect and make inferences from the information are mostly invisible to users. It is difficult to manage information boundaries appropriately when users are unaware of disclosures [8].

While some of the information used by personalization algorithms for tailoring content to user interests and preferences comes from information people explicitly contribute and can therefore self-censor, much of the data is collected invisibly as users surf the web. Companies are not always as transparent as they could be in their stated practices about what data they have access to, and how they will use it. For example, Willis et al. [43] conducted an investi-

gation to determine the extent of personalization in Google search results. They "induced" interests in fake profiles by doing searches with particular keywords and viewing specific videos on YouTube, expecting that this information would be used by Google to determine which ads to display. Google's policy at the time stated that ads displayed with search results would be *contextual* ads, selected only based on information in the search result page itself. The researchers found that *non-contextual* ads based on inferred interests from previous interactions appeared alongside the contextual ads, despite the policy. They also found that some of the non-contextual ads could potentially reveal sensitive personal characteristics based on the inferred interests, such as an ad which contained the question, "Do you have diabetes?"

In a different study, Korolova [17] investigated the extent to which information Facebook users specified as available to "Only me" could be used for targeted advertising. In one example, she created a series of Facebook advertisements targeted toward characteristics of a person known to the research team, who had specified that profile information about age should be hidden from everyone. The specially crafted ads differed according to only one dimension: the age of the user to whom the ads should be displayed. Using Facebook's advertiser interface, Korolova was able to infer the private age of the target person based on updates about the performance of ad campaigns—since the ads for the incorrect ages were not displayed. Her experiment demonstrates the possibility that even when users indicate they want to keep specific information private, Facebook has used that information to target advertisements in a potentially revealing way.

In some studies, users report that they like personalized search, because personalization provides better results [27]. Likewise, many people say that they are comfortable with customized ads based on the contents of their email or Facebook profile, and also find tailored ads to be useful [1, 41]. However, when asked directly about the sensitivity of specific Google search queries, 84% of users in one study said that there were queries in their search history that they felt were "sensitive", and 92% wanted control over what Google was tracking about them as they searched the web [27]. Less than 30% of participants in another study were aware that browsing history and web searches could be used to automatically create a profile about them, and most people were unable to distinguish between the company represented by the ad content, and the company responsible for displaying the ad [41].

Altman [2] wrote, "If I can control what is me and not me; if I can define what is me and not me; if I can observe the limits and scope of my control, then I have taken major steps toward understanding and defining what I am." There are few options for users who want to manage multiple identities with respect to systems or companies, rather than self-presentation to other people, for the purpose of maintaining separate personalization experiences. The invisibility of the architectures and algorithms responsible for personalization make it difficult for users to manage boundaries appropriately with respect to information privacy [8].

2.3 Research Questions

Users may be in danger of losing control over the mechanisms by which they develop and enforce their individuality online, because they don't know and can't control who the system thinks they are, and how that identity is presented to other people and organizations. This study focused on situations people encounter in everyday web use where information disclosure boundaries are not straightforward. The purpose was to investigate (1) whether users are concerned about privacy when they engage in common behaviors on the web that can enable automated disclosures to take place; (2) whether people are aware of different types of data that can be automatically collected about them when they use Facebook and Google Search; and (3) how the perceived likelihood of automated data collection might be related to privacy concern.

3. METHOD

I conducted a 2 (*Site*: Facebook or Google Search) x 3 (*Behavior*: Link, Autocomplete or Ad) x 2 (*Sensitivity*: High or Low) between-subjects online experiment hosted by Qualtrics, in May 2013. Participants viewed a hypothetical situation that varied according to these three dimensions, which are described in detail below. This study was approved as minimal risk by our Institutional Review Board.

3.1 The Site Dimension

The two levels of the *Site* dimension were Facebook and Google Search. Interacting via social media and searching for information on the web are two very common Internet-related activities, yet they have some interesting similarities and differences. Many of the underlying web technologies, particularly related to the implementation of dynamic, interactive web pages, are the same in these two situations. However, one way in which these two sites differ is the degree to which user actions take place in a social context. Searching is typically a solitary activity, and it is reasonable to assume that people feel more like they are interacting with the search engine database than another human being when they search for something. Using social media feels like communicating, even when one is simply browsing the Facebook News Feed. This contextual difference could affect whether people feel their actions on the two sites can be observed or not. In addition, the settings and mechanisms users have to control access to their information on Facebook are all geared toward social privacy, not information privacy.

3.2 The Behavior Dimension

I chose three behaviors to include in this study: clicking a link, typing in a text box, and viewing ads in a web page. These behaviors seem on the surface like they are not directly related to disclosures of personal information, because they do not directly ask for it. However, it is possible to infer personal information from all three.

Clicking a Link: When a user clicks a link in Facebook or Google, he or she sees visual feedback that the system has registered the action when the web page changes to display new content. Clicking a link in both systems sends a request to the server that hosts the content of the page the user is navigating to. Users may already be aware of this, since it is a fundamental aspect of how the Internet works. However, both Google and Facebook can employ redirects so that they can collect data about which links users click on. So while there is visible feedback that something server-related is happening, it is less clear to users that Google and Facebook can record information about what links you click on.

Data consisting of which links users have clicked on can be used to infer the gender and age of individual users who have not revealed that information, as long as a sufficient number of other users with similar browsing patterns have provided their gender and age information. This is accomplished by first identifying the most common gender and age segment for the visitors of a set of web pages. Then, the age and gender of other visitors to those pages are inferred, whether or not they have chosen to reveal them. Gender can be inferred with 80% accuracy, and age with 60% accuracy [15].

Typing and Autocomplete: When a user types in a text box on

Facebook or Google Search, both sites send individual characters back to the server as they are typed. This real-time communication supports auto-completing search terms and the names of Facebook friends when creating a status update, without having to explicitly click the Submit button. However, the extent to which this feedback might be understood to communicate outside the web browser differs across the two sites. For example, when a user types a status update, the only visual indicator that information has been transmitted occurs when one's Facebook friends' names appear below the text box. However, Google Instant Search updates the entire web page as a search query is typed by the user. These different levels of feedback may lead to different conclusions on the part of the user about what and how much information might be going back-and-forth between themselves and the system as they are typing, before they explicitly submit the text. In reality, data is sent back to the server in both cases.

Viewing Ads in a Web Page: Ads in web pages can have a visible relationship with other information displayed at the same time in the web page (called *contextual* ads), or be based on other data available to advertising companies about the end user (confusingly called *non-contextual* ads) [43]. Therefore, different types of ads provide different kinds of feedback from the system to the user about inferences the system has made about them. Google ads in search result pages appear after the user has requested information via a search query, and tend to be contextual. This might trigger users to notice that ads are personalized, and they might therefore be more concerned about privacy. On the other hand, because Facebook ads are more likely to be based on one's profile information and "Likes" rather than information displayed in the News Feed (i.e. non-contextual), users who notice this may feel more concern about why particular ads were selected for display. However, there is invisible data collected too, that users do not receive feedback about: when an ad loads in a particular page, data is recorded about which ad loaded where.

3.3 The Sensitivity Dimension

The sensitivity of the information involved might increase overall privacy concern, and affect whether users wonder if data about their actions can be recorded. The High Sensitivity condition included ads, links to content, and search queries or posts about depression, a psychological disorder that is both common and highly stigmatized, and affects both men and women [23, 13]. The content and statements in the stimulus materials related to depression were based on research conducted by Moreno et al. [24], looking at college students' references to their own depression on social media websites. The Low Sensitivity condition consisted of content such as links to the website of the a local minor league baseball team, a technology-related article, and ads for a laptop or iPad.

3.4 The Experiment Procedure

The online experiment started by displaying a hypothetical situation that varied by condition, designed to closely resemble common experiences while using the web. Below is the text displayed to participants, corresponding with the levels of the *Behavior* dimension. Each condition was accompanied by a partial screen capture to illustrate what was happening, and the manipulation of *Site* and *Sensitivity* took place via the screen captures. All screen captures are included in Appendix A.

Link You visit Facebook and start reading posts in your Facebook News Feed. You scroll down the page, and click on a link a Facebook Friend has shared. The page changes to show the web page for the link that you clicked on.

Autocomplete You visit Google and start typing in the search box. Google makes a guess about what you might be searching for, and shows search results before you finish typing.

Ad You are viewing posts in your Facebook News Feed. As you scroll down the page, reading posts made by Facebook friends, you notice ads displayed on the right side of the screen.

Participants were asked a closed-ended and an open-ended *privacy concern* question, immediately after viewing the hypothetical situation:

1. Would you be concerned about unwanted access to private information about you in this scenario? [Yes, Maybe, No]
2. Please explain your answer to the previous question. [open-ended]

This emphasis on "unwanted access" follows from several definitions of privacy as control over access [42, 2]. Asking participants about concern over unwanted access is essentially operationalizing privacy as control over one's information. Likert scales often measure both direction and intensity at the same time (e.g., a "Very Satisfied" to "Very Dissatisfied" scale measures both whether someone was satisfied or dissatisfied, and by how much) [9]; however, the privacy concern question in this study asks about the presence or absence of concern, not how much concern. The additional *Maybe* option, rather than simply *Yes* or *No*, allows more accurate measurement of responses by not forcing participants to choose between the two extremes if they were unsure.

Asking the question in this way does not ask participants about specific things that may have caused them concern, and therefore it is not clear what they might have been thinking about when they answered the question. This phrasing of the question was intentional, in order to avoid "priming" participants to consider things they might not have thought about before when answering the question. The point of the manipulation was to trigger participants to think about a specific situation, but NOT to trigger them to think about specific *characteristics* of the situation, as a way to get as unbiased a response at possible given the study format.

After the privacy concern question, participants responded to a 16-item question that asked them to estimate the likelihood that Facebook or Google could collect different kinds of data about them: "How likely do you think it is that [Google | Facebook] can AUTOMATICALLY record each of the following types of information about you?" The motivation for asking about these items was to identify what kinds of "tracking" users think may be going on when they use the web, and through later regression analysis to identify associations between these beliefs and the likelihood of privacy concern. Participants indicated the likelihood of each statement between 0 and 100 in intervals of 10, using a visual analog scale represented as a slider. Half of the participants in the study were asked these questions about Facebook, and the other half about Google, and this depended on what *Site* condition they were randomly assigned to after they completed the consent form. The 16 items ranged from the clearly possible (which links the user clicks on), to the unlikely to be perceived as possible to collect (what the user's desktop image looks like). The question also included a few examples of information that can be inferred; for example, sexual orientation, which can be inferred from Facebook "Likes" [18]. However, few participants were expected to believe it likely that Facebook or Google could automatically detect this. See Figure 6 for the text of the items.

I included two sets of *control questions* in the survey: one to measure participants' Internet literacy (operationalized as familiarity with a set of Internet-related terms), and another to gauge the level of importance each participant placed on digital privacy. The questions that comprise the Internet Literacy index variable are based on the Web Use Skills survey reported in Hargittai and Hsieh

	Ad		Autocomplete		Link	
	High	*Low*	*High*	*Low*	*High*	*Low*
Facebook	60	60	61	56	60	60
Google	59	55	61	55	60	54

Figure 1: Number of participants in each condition. Independent variables are Site (Facebook or Google), Behavior (Ad, Autocomplete, or Link), and Sensitivity (High or Low).

(2011) [14]). This variable consists of the average of participants' assessments of their level of familiarity with the a list of Internet-related terms (*M*=3.57; *SD*=0.75, Cronbach's α=0.8).

I selected the questions that make up the Privacy Preferences index variable from two published privacy scales. The first was the "Blogging Privacy Management Measure", an operationalization of Communication Privacy Management theory applied to blogging by college students by Child et al [5]. This scale measures how bloggers think about boundaries between private and public when disclosing information online. I modified 8 items from that scale, replacing "blog" with "Facebook" where appropriate. An example item included in this study is, "If I think that information I posted to Facebook really looks too private, I might delete it." In addition, I selected four items from the "Information Privacy Instrument" developed by Smith et al [37]. This scale was designed to measure individuals' perceptions of organizational practices surrounding information privacy. An example item from this scale used in the study is, "It usually bothers me when companies ask me for personal information." Participants responded to these 12 items on a 5-point likert scale of Strongly Disagree—Strongly Agree.

To create the index variable, I reverse-coded where necessary and averaged across all 12 questions. The Privacy Preferences index variable therefore represents both attitudes toward individual disclosure in social media, and comfort level with the way organizations handle private user data. The mean of the privacy preferences variable was 4.003 (*SD*=0.5, Cronbach's α=0.74), which indicates that on average, participants valued online privacy, and were bothered by the idea of companies selling information about them to third parties.

3.5 Participants

I recruited participants from Amazon Mechanical Turk (MTurk), and restricted the sample to workers from the USA who had a 95% or higher approval rating after completing at least 500 tasks. MTurk workers were first required to answer an eligibility screening questionnaire. Participation was limited to MTurk workers who reported that they visited both Facebook and Google Search at least weekly, and were 18 or older. Using web-savvy MTurk workers as participants was convenient, but also purposeful: people who make money by completing tasks on the Internet are a best-case scenario for finding a population that is aware of invisible data collection and privacy risks on the Internet, compared with the usual suspects like undergraduates or a snowball sample. Participants completed the questions in an average of 7.56 minutes (*SD*=6.1 minutes) and received $2 in compensation. 748 participants started the survey; 47 were excluded because they did not finish the survey, or they failed to answer the attention check questions correctly, or they completed the survey during a Qualtrics service disruption.

After these exclusions, the number of participants remaining in each condition ranged from 54 to 61 (see Figure 1). The answers of the remaining 701 participants to the demographic questions resemble what other researchers have found about MTurk samples [3]—this sample was young (*M*=30.25 years old, *SD*=9.22), 80% white, more male (57%) than female (42%), and the majority (79%) had completed some post-high-school education or earned a 4-year college degree. Nearly all participants reported visiting Facebook (86%) and Google Search (98%) daily or more often. Finally, 97% of participants in the final sample reported having personally experienced a situation similar to the condition they were assigned to in the study.

	Estimate	*Odds Ratio*	*Std. Error*
Behavior: Autocomplete	-1.86***	0.16	0.37
Behavior: Link	-1.03**	0.36	0.35
Site: Google	-0.80***	0.45	0.35
Sensitivity: Low	-0.28	0.75	0.35
Autocomplete x Google	1.28*	3.59	0.51
Link x Google	1.03*	2.80	0.49
Autocomplete x Low	-0.01	0.99	0.54
Link x Low	-0.24	0.79	0.50
Google x Low	-0.80	0.45	0.51
Autocomplete x Google x Low	0.22	1.24	0.76
Link x Google x Low	-0.48	0.62	0.75
Internet Literacy	-0.12	0.89	0.10
Privacy Prefs	0.99***	2.71	0.17

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 1: Coefficients for the Proportional Odds Multinomial Logistic Regression. The dependent variable represents participants' level of concern over unwanted access to private information, with three levels: *Yes, Maybe,* and *No*. The Baseline condition is Facebook:Ad:High. AIC is 1309.42; McFadden's Pseudo-R^2 is 0.096.

4. RESULTS

As expected based on previous research, more people answered *No* (377 participants) and *Maybe* (173 participants) than *Yes* (151 participants) when asked if they were concerned about unwanted access to private information. What follows are several analyses that help us to better understand when participants were more likely to express concern.

4.1 Conditions and Privacy Concern

I used a Proportional Odds Multinomial Logistic Regression to evaluate the relationship between the experiment conditions (*Site* x *Behavior* x *Sensitivity*), Internet Literacy and Privacy Preferences as controls, and the dependent variable: participants' answers to a single question about whether they would feel concerned about unwanted access to private information in the condition they were randomly assigned to. Like any closed ended question having an ordinal response format, it is possible that a *Yes* from one participant might mean more concern than another participant's *Yes*. While it is impossible to objectively compare the subjective experience of concern across participants, within each individual it is reasonable to interpret *Yes* as more concern than *Maybe*, which is more concern than *No*. The results from the model are in Table 1.

The multinomial logistic regression estimates the probabilities of choosing higher levels of concern than *No*. The baseline condition is Facebook:Ad:High, and all of the coefficients must be interpreted in relation to that combination of categories. Positive coefficients indicate greater likelihood of expressing concern; coefficients around 0 mean no additional likelihood on top of the baseline, and negative coefficients indicate lower likelihood of concern. For example, the large, negative estimate for the Autocom-

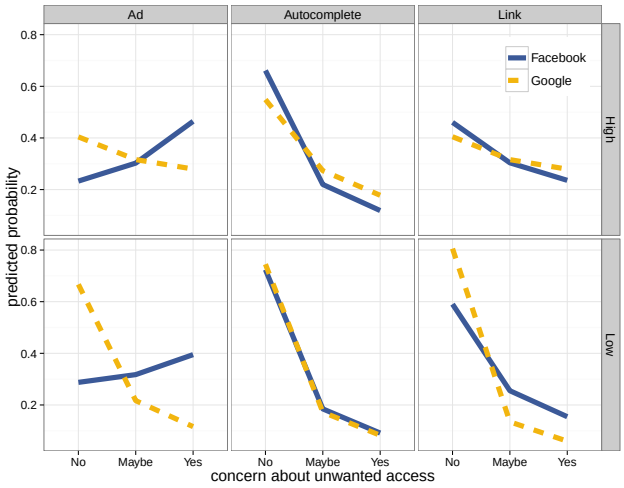


Figure 2: Predicted probabilities from the regression model presented in Table 1. The x-axis is the categorical response to the concern question, and the y-axis is the predicted probability of choosing a particular response.

plete conditions (-1.86) means that participants exposed to these conditions were much LESS likely to say they would be concerned about unwanted access to private information than participants exposed to any of the Ad conditions. Figure 2 presents the results as predicted probabilities generated from the model for a hypothetical participant who is average on the Internet Literacy and Privacy Preferences control variables.

Privacy Concern is Highest for Facebook Ads

Participants were most likely to express concern about unwanted access when they viewed the Facebook Ad conditions at both levels of Sensitivity. Participants who answered *Yes* to the concern question in the Facebook:Ad:High Sensitivity condition explained why they were concerned, by suggesting that the content of the ads makes them feel uncomfortable about what Facebook knows about them. They said things like, "Private information is being read from my posts," and "These ads seem to tell me that the computer knows about certain traits of mine due to my computer's history. I don't want Facebook to have this access." Participants in the Google:Ad:High Sensitivity condition expressed similar concerns, although fewer answered *Yes* to the concern question: "I would be concerned that someone could find out my search for depression by checking my Google search history, and that they keep a record of that when they display ads to me."

In contrast, participants in the Google:Ad:Low Sensitivity condition who said they would NOT be concerned about unwanted access said things like the following: "I think I've gotten used to having google [sic] searches causing ads to be pushed at me. In this case, nothing in the results is based on personal information–it's all from the search query just entered." This statement clearly expresses that the participant believes search results and ads are based on search queries, not personal information, implying that the participant feels the queries themselves are not personal information.

Figure 2 also clearly illustrates a statistically significant Scenario x Site interaction. Participants were more likely to say they were unconcerned than concerned about unwanted access to private information in the Google:Ad conditions. However, the opposite was true for participants exposed to the Facebook:Ad conditions. This means that web-savvy users, like Turkers, are more worried about privacy violations when they see targeted ads in Facebook than in Google Search.

Privacy Concern is Similar for Sensitive Ads and Links

The lines on the graph in Figure 2 for both Facebook and Google in the Link:High sensitivity conditions are similar to each other, and they also look very similar to the line for Google in the Ad:High condition. These predicted probabilities were indeed very similar: around 40-45% likelihood of answering *No*, 30-32% likelihood of answering *Maybe*, and 24-28% likelihood of answering *Yes*. In other words, participants were similarly likely to express concern about clicking on a "sensitive" link about depression in Facebook OR Google, as about viewing "sensitive" ads about depression in Google. Reasons they expressed for being concerned included statements focused on social, not information privacy: "Because, I just clicked on the link. I only would be concern if facebook [sic] announced on the news feed that I read the article"; and "it wouldn't bother me in the least if it was discovered that i'd [sic] been searching for information on depression". However, participants who did express concern said things that indicated they are aware of some of the data collected about them, e.g.: "I am very concerned about my search history, and specifically in this scenario I would be concerned about someone knowing I was depressed" and "Sometimes you get to stories by linking from other places online, and those could turn up in the URL of the story. Someone clicking on it could potentially figure out where I was surfing."

Privacy Concern is Lowest for Links in Google

The lowest likelihood of concern about unwanted access to private information in the experiment came from participants exposed to the Google:Link:Low Sensitivity condition. Just 6% of participants having average Internet Literacy and Privacy Preferences exposed to this condition are predicted by the model to choose *Yes*. This is clear evidence that web-savvy users view clicking on links in Google search results as an activity that does not have the potential to reveal information about them. As one participant explained, "It's just a link to a page. It's not asking for any personal information."

Autocomplete Does Not Warrant Concern

Participants in the Autocomplete conditions consistently reported that they would not be concerned about unwanted access to private information. Just 29 out of 233 participants exposed to Autocomplete conditions, across all levels of Site and Sensitivity, expressed concern. Their explanations made vague allusions to being tracked online, without being specific or technically accurate: "Nothing is every [sic] really private when online and Facebook offering suggestions when I type a status update proves I'm not just being paranoid."

The 155 participants in Autocomplete conditions who answered *No* to the privacy concern question gave reasons based on the *Site* they were asked about. Facebook participants in the Autocomplete condition who were unconcerned gave reasons such as, "I am not concerned about my privacy because Facebook already has my friends [sic] information. Facebook is just taking the list of my friends and presenting them in a new way." Likewise, participants exposed to both Google Autocomplete conditions said things like, "I don't really find this to be an invasion of privacy, I see it as Google thinking ahead. I would be pleased if the search that I wanted popped up before I finished typing it. It would save me some time"; and "The information that they are presenting is [the] most common used search that involves what you are beginning to

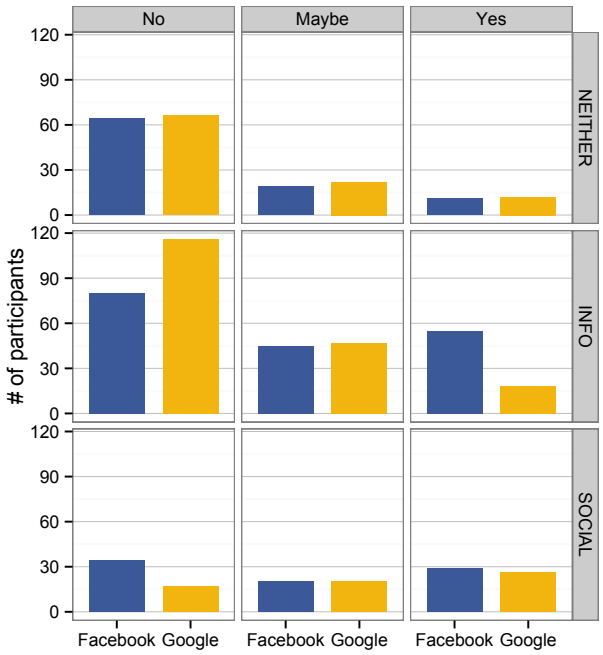


Figure 3: Number of responses coded as *Neither, Info* or *Social*, broken down by *Site* and the participant's concern response.

type. It does not contain specific information about what I have searched for."

In fact, Autocomplete works by sending keystrokes back to the servers of Facebook and Google, as they are typed, and matching them with other users' previously recorded queries. It is possible to use freely available "developer tools" for popular web browsers (e.g., Firebug, a plugin for Firefox) to see requests that pass information back and forth between the browser and Facebook's or Google's servers. On Facebook, this includes each character as it is typed in the Status box. These requests happen in the background, very quickly, and are typically not visible to end users. Features like Autocomplete further blur the line between social vs. information privacy, and recent research about self-censorship in social media [6, 35] does not take into consideration that users share ALL content they type with Facebook and Google, not just what they choose to submit or post.

"Unwanted Access" Refers to Websites, Companies

It is possible that when two different people answered *Yes* to being concerned about unwanted access to private information, they were concerned about different things. To investigate this, I analyzed participants' open-ended explanations for why they chose *Yes, Maybe* or *No* to the privacy concern question, to better understand what participants interpreted "unwanted access" to mean. A research assistant who had not previously examined data from this study used a bottom-up process to identify themes in 100 randomly selected responses, and developed the coding scheme based on those themes. The research assistant and the author then coded all 701 responses, without knowing which condition each response had come from or how the participant had answered the privacy concern question. The coders met to resolve disagreements and produce a final coding for each response. Cohen's κ was 0.82, indicating "excellent" inter-rater agreement [19].

	Estimate	*Odds Ratio*	*Std. Error*
Site: Google	0.116	1.123	0.306
Code: INFO	1.043***	2.839	0.264
Code: SOCIAL	1.136***	3.115	0.305
Google x INFO	-1.135**	0.321	0.371
Google x SOCIAL	0.374	1.454	0.437
Internet Literacy	-0.059	0.942	0.101
Privacy Prefs	0.922***	2.515	0.165

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 2: Coefficients for the Proportional Odds Mulitnomial Logistic Regression. The dependent variable represents participants' level of concern over unwanted access to private information, with three levels: *Yes, Maybe,* and *No*. The Baseline condition is Facebook:NEITHER. AIC is 1334.33; McFadden's Pseudo-R^2 is 0.070.

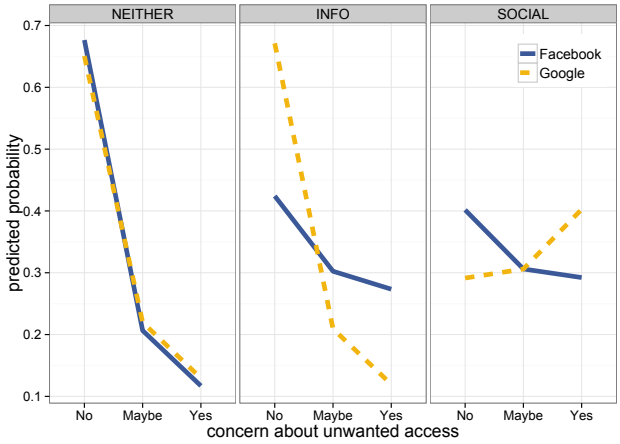


Figure 4: Predicted probabilities for the regression in Table 2. The x-axis is the categorical response to the concern question, and the y-axis is the predicted probability of choosing a particular response.

The final coding scheme had three mutually-exclusive categories, *Neither, Info* or *Social*. Responses coded as *Neither* did not provide enough evidence for coders to tell what kind of access the participant focused on when deciding whether he or she would feel concerned in the hypothetical situation. Examples of responses coded as *Neither* (n=194) include, "Nothing on the Internet is really private" and "All that appears is my name and where I am".

Responses coded as *Social* (n=146, the smallest category) included language referencing control over access by specific people, such as friends and family, social network connections, work supervisors, or being targeted by hackers. Responses coded *Social* were similar to the following: "No reason to be afraid, especially if my friend wouldn't mind it" or "I hate when previous searches pop up while someone is browsing my computer."

Finally, responses coded as *Info* (n=361, the largest category) mentioned control over access by websites, companies, governments, or other organizations. More responses were coded *Info* than *Social* or *Neither* combined. Many of these responses used passive voice and ambiguous pronouns, indicating that it may have been difficult for participants to put into words specifically when or how the unwanted access could take place. Examples of *Info* responses include, "I wouldn't really be offended by them targeting

ads towards me. That's how they make money" and "I wouldn't be 100% sure that my information was not linked to this site when I clicked the link."

In a few instances, responses contained both references to information and social privacy. If it was possible to tell which type of unwanted access the participant was more concerned about, that code was applied; otherwise, these responses were coded as *Social* (this happened only a handful of times). The number of responses coded as each category is presented in Figure 3, broken down by *Site* and the participant's concern response.

More "Info" Concern about Facebook than Google

I conducted a Proportional Odds Multinomial Logistic Regression with concern about unwanted access as the dependent variable, *Site* and *Type of Unwanted Access* (Info or Social) as regressors, and Internet Literacy and Privacy Preferences as controls. This analysis allows me to estimate, for example, the likelihood that a participant who mentioned social versus information privacy in his or her explanation would report concern about unwanted access depending on exposure to hypothetical situations involving Facebook or Google. The regression results are presented in Table 2.

The large, positive coefficients for the *Info* and *Social* categories mean that responses assigned those codes were more likely to be associated with *Yes* answers to the concern question, than responses coded as *Neither*. The large, negative coefficient for the Google x *Info* category means that information privacy concern was less likely to be associated with *Yes* answers in the Google conditions than in the Facebook conditions. All of these coefficients are also statistically significant.

The graph in Figure 4 shows the predicted probability of concern for participants with average Internet Literacy and Privacy Preferences. This graph illustrates that when participants associated "unwanted access" with privacy from websites, companies, and other institutions, those who were randomly assigned to Facebook conditions (solid blue lines in the graph) were more likely to express concern than those assigned to Google conditions (yellow dotted lines). However, this pattern was reversed for participants that associated "unwanted access" with social privacy. Participants who mentioned privacy from other people in the explanations for their answers were more likely to say they would be concerned when exposed to hypothetical situations involving Google than Facebook.

4.2 Perceived Likelihood of Data Collection

I conducted an exploratory factor analysis to identify patterns in participants' perceived likelihood that different types of data can be collected about them automatically while interacting with Facebook or Google Search. The maximum likelihood extraction with varimax rotation resulted in four interpretable factors. The factor loadings and text of the items are in Figure 6, and frequency histograms for each item are represented in Figure 5. The x-axis of each histogram in Figure 5 represents participants' assessments of the likelihood of each type of data being collected about them, ranging from 0 (Unlikely) to 100 (Likely) in increments of 10. The y-axis represents the number of subjects who chose each likelihood increment, for each variable. The gray line represents Facebook, the black dotted line in each histogram represents Google. Reliability scores (Cronbach's α) are also reported in Figure 6, for index variables created for each factor by averaging within participants across all items that comprised the factor.

OLS regressions with each factor's index variable as the dependent variable and the experiment conditions plus Internet Literacy and Privacy Preferences as controls revealed no significant interactions. This means that participants' answers on these items did not vary based on the experiment condition they were randomly assigned to. However, there was a main effect for *Site*, likely because participants were asked to estimate the likelihood of automatic data collection in Facebook OR Google. (Participants assigned to one of the Google conditions answered questions about Google throughout the entire study.)

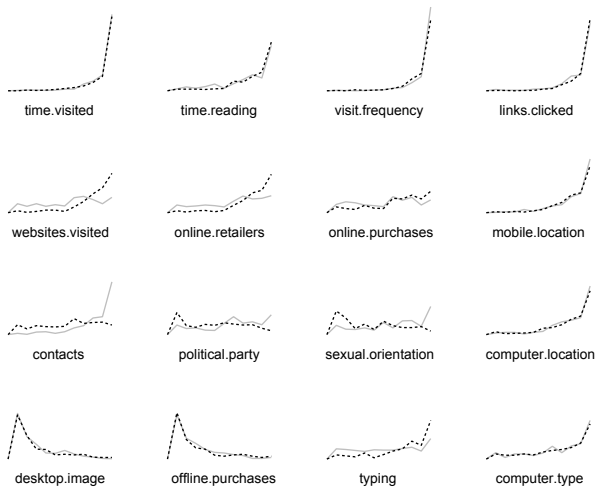


Figure 5: The x-axis of each frequency histogram represents participants' judgments of the likelihood of each type of data being collected about them, ranging from 0 (Unlikely) to 100 (Likely). The y-axis represents the number of subjects who chose each likelihood increment. The gray lines represent Facebook, the black dotted lines, Google. The questions associated with each histogram are in Figure 6.

Factor 1: First-Party Data

The questions that make up the "First-Party Data" factor are across the top of Figure 5 and down the right side. This factor includes the items *time.visited, time.reading, visit.frequency, links.clicked, mobile.location, computer.location* and *computer.type*. Each item asks about information that is available to websites directly as a result of user interaction. The pattern of these responses clearly illustrates that participants were aware that these types of information can be automatically collected. Nearly every participant felt that what time they visited Facebook or Google could be collected, for example, but there was a little bit more variance among participants about whether it is likely that Facebook or Google could figure out what type of computer they were using. It is actually possible to automatically collect this information—one's operating system and browser version are sent from the web browser to the web server when it requests a page.

Factor 2: Aggregation Across Sources

The questions making up Factor 2, "Aggregation Across Sources", are displayed in the first three histograms of the second row of Figure 5. Items *websites.visited, online.retailers* and *online.purchases* represent information about what other websites one visits and what kinds of things one shops for online. This is information Facebook and Google can only know by partnering with other websites, and associating one's profile with his or her behavior on those sites. This kind of data is similar to what one might see in a credit report that aggregates financial activity across multiple accounts, but without the score, and realize that it is possible to obtain a history

	Alpha	*Factor Loading*	*Abbreviation*	*Mean*	(*SD*)
First-Party Data	**0.78**			**84.9**	(**14.2**)
what time of day you visit [Google | Facebook]		0.817	*time.visited*	92.0	(15.6)
your physical location when using [Google | Facebook] on a mobile device		0.506	*mobile.location*	84.9	(19.9)
how much time you spend reading [Google | Facebook]		0.526	*time.reading*	80.0	(25.5)
what kind of computer you are using when you visit [Google | Facebook]		0.412	*computer.type*	71.8	(30.6)
your physical location when using [Google | Facebook] on a computer		0.501	*computer.location*	81.2	(23.9)
how often you visit [Google | Facebook]		0.756	*visit.freq*	93.2	(13.9)
what links you click on in your [Google search results | Facebook news feed]		0.712	*links.click*	91.0	(16.2)
Aggregation Across Sources	**0.87**			**67.0**	(**22.7**)
what websites you visit most often		0.764	*websites.visited*	69.6	(29.8)
which online retailers (e.g. Amazon.com) you visit most often		0.931	*online.retailers*	71.1	(29.0)
what you purchase from online shopping websites		0.689	*online.purchases*	60.1	(31.2)
Aggregation Across People	**0.80**			**57.0**	(**27.7**)
which people you communicate with online most often		0.548	*contacts*	70.0	(30.5)
your political party affiliation		0.815	*political.party*	50.8	(32.7)
your sexual orientation		0.860	*sexual.orientation*	51.0	(34.7)
"Impossible" to Collect	**0.60**			**19.4**	(**20.8**)
what the desktop image on your computer looks like		0.651	*desktop.image*	19.0	(24.0)
what you purchase from a brick-and-mortar store		0.477	*offline.purchases*	19.7	(25.1)
Not part of any factor					
what you are typing in the [search | Post or Comment] box before you submit		*n/a*	*typing*	65.0	(32.9)

Figure 6: Items measuring participants' beliefs about the likelihood that different types of data can be collected about them automatically by Facebook or Google [0 (Unlikely) to 100 (Likely)]. These items were presented in random order to each participant; here they are grouped and labeled according to the results of an exploratory factor analysis. Cronbach's α reliability scores are presented for each factor.

of one's activity that would be difficult to reconstruct from memory.

Participants were more divided in their judgments about the likelihood that Facebook and Google can know things about them that require this kind of aggregation. Participants assigned to Google thought it was more likely that information about what websites they visit and where they shop online could be collected, than participants assigned to Facebook. Interestingly, the technology and business partnerships with data aggregators that are necessary to collect this kind of data are feasible and practiced by practically all websites that use advertising. The variability in these responses indicates that participants estimations of likelihood are not likely to be based on knowledge about what is technically possible.

Factor 3: Aggregation Across People

Participants asked about Facebook vs. Google diverged the most on the items that make up the "Aggregation Across People" factor. The histograms for these questions are represented in the third row of Figure 5. This factor consists of one's *contacts, political.party,* and *sexual.orientation*: information that can be inferred through comparing patterns of behavior across people. For example, if some people disclose their sexual orientation directly in their profile, others with similar behavior patterns that did not choose to reveal this information may still be labeled the same. This kind of data is like the score or rating part of one's credit report, in that it provides information about how the system evaluates one's activity in the context of other people.

Participants asked about Google were spread across the range of responses for these questions, but tended toward thinking that it was unlikely Google could automatically collect information about their political party affiliation or sexual orientation, or the people they communicate with online. Participants who answered the questions about Facebook reported higher estimates of likelihood that this information could be automatically collected. All three of these types of information can actually be inferred from information users disclose online.

Factor 4: "Impossible" to Collect

Factor 4 consists of only two questions, that stand out in the bottom left corner of Figure 5 as the only two questions that skew toward the left or "unlikely" end of the range of possible responses, indicating that most participants believed it is not likely that Facebook or Google can automatically collect this information. This factor includes questions about the desktop image on one's computer and purchases in brick-and-mortar stores (*desktop.image, offline.purchases*). In fact, through partnerships with data aggregators it is possible that web companies can access data about users' buying habits in brick-and-mortar stores [34]. However, while it is technically possible for a web company to detect what a computer's desktop image looks like, it would be difficult to accomplish without compromising the security of the computer. I included the *desktop.image* question as a way to anchor the interpretation of users' responses to the awareness questions; if many participants thought this was possible, all responses to questions in this section of the survey would be suspect.

Typing

Finally, one question was not part of any factor: the likelihood that Google and Facebook can automatically collect "what you are typing in the [search | Post or Comment] box before you submit". Participants who answered questions about Facebook were fairly evenly spread across the range of responses (M=55.24, SD=33.7), indicating that participants varied in their beliefs about whether Facebook can record users' keystrokes as they are typing. However, the pattern is different for Google: more participants who answered the version of the question about whether Google can automatically collect information about what they are typing before

they submit the text reported feeling that this data collection was likely (*M*=75.17, *SD*=28.66).

Responses to this question are an indication that the nature of the interaction, and the type of visual feedback, may be important for understanding what is going on "under the hood". Google Instant Search provides search results as users type, and the entire web page updates to reflect search results. This seems to convey to at least some web-savvy users that information they are typing is been sent to Google in real-time. However, the information Facebook displays as users are typing consists of the names of one's friends that match the characters that have been typed. It was less clear to participants in this study whether it might be necessary to transmit those characters back to Facebook in order to make those suggestions.

4.3 Awareness and Privacy Concern

I ran a third Proportional Odds Multinomial Logistic Regression to evaluate the relationship between awareness (perceived likelihood) of automatic data collection and privacy concern. I used *Site* and three of the index variables created from the exploratory factors, described above as regressors. These variables represent participants' perceptions of the likelihood that Google or Facebook can collect First Party Data (*first.party.data*), data from Aggregation Across Sources (*source.aggregation*), or data from Aggregation Across People (*people.aggregation*). The dependent variable was the same privacy concern variable as the previous multinomial regressions: whether participants would be concerned about unwanted access to private information in the hypothetical situation they were exposed to (Yes, Maybe or No). I also included the two continuous controls, Internet Literacy and Privacy Preferences, in the model. The purpose of this analysis was to identify whether a relationship exists between participants' beliefs about how likely it is that their behaviors online are recorded, whether inferences based on that data are possible, and their concern about privacy.

I generated three sets of predicted probabilities from this model to help with interpretation. First, I held the values of all regressors at their means except for *first.party.data*, for which I generated predicted probabilities at 10-point increments between 0 and 100. I did the same for *source.aggregation* and for *people.aggregation*, holding all other regressors at their means. This allows for comparison of the effects of increasing awareness of these three types of information on the predicted probability that a participant would report *Yes*, they would be concerned about unwanted access to private information. Figure 7 depicts these results graphically. Each line in the graph represents one set of predicted probabilities. The predicted probabilities for Facebook and Google are presented separately due to the statistically significant effect of *Site* in this regression. Predicted probabilities of concern are higher for Facebook than for Google.

Figure 7 illustrates that an increase in the perceived likelihood that First Party Data can be collected automatically was associated with a DECREASE in the predicted probability of a participant expressing privacy concern. The more a participant was aware of automatic First Party Data collection, the less concerned he or she was about unwanted access to private information. The open-ended explanations indicated that many participants felt things like what time of day they visit or what links they click on did not need to be kept private. However, as the perceived likelihood of inferences enabled by Source or Person aggregation increase, the predicted probability of of concern about unwanted access to private information also INCREASES. The more a participant believes these inferences are possible, the more likely he or she was to express privacy concern.

	Estimate	*Odds Ratio*	*Std. Error*
Site: Google	-0.498*	0.608	0.197
first.party.data	-0.007	0.993	0.006
source.aggregation	0.011**	1.011	0.004
people.aggregation	0.007*	1.007	0.004
internet.literacy	-0.047	0.955	0.103
privacy.prefs	0.930***	2.535	0.165

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3: Coefficients for the Proportional Odds Mulitnomial Logistic Regression. The dependent variable represents participants' level of concern over unwanted access to private information. The Baseline condition is Facebook. AIC is 1364.8; McFadden's Pseudo-R^2 is 0.0471.

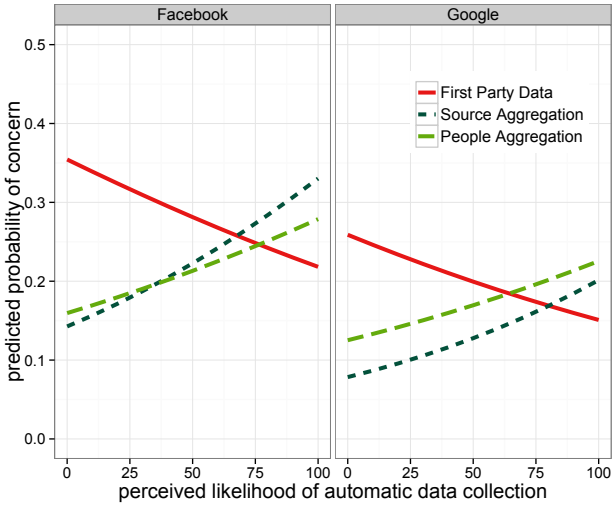


Figure 7: Predicted probabilities from the model in Table 3. The x-axis represents participants' perceived likelihood that Facebook or Google can automatically collect data about them, and the y-axis represents predicted probability of answering *Yes* to the question about privacy concern.

5. DISCUSSION

The data collection technologies and algorithms supporting personalization and behavioral advertising have developed quickly and invisibly, and for web users it is increasingly hard to avoid this "surveillance by algorithm"[2]. Using the web discloses information simply by virtue of interacting with web pages, and then once the information is out of users' control, they have little choice but to trust companies and other people to protect the information the same way they would [22]. Not every user will feel great risk of harm by having their sexual orientation inferred. But, some users might want to keep information like this private, and they presently have no control over it if they want to use the web. They cannot effectively manage that boundary without withdrawing from the Internet altogether. This paper shows that users' perceptions about what unwanted access looks like have very little resemblance to the actual ability of personalization and advertising algorithms to make inferences about them, and this problem will only grow as networked sensors (and the efficiencies and conveniences they provide) become more integrated in our daily activities.

[2] https://www.schneier.com/blog/archives/2014/03/surveillance_by.html

The high-level question that motivated this research project is, when do users currently feel like their actions online are being observed—not necessarily by other people, but recorded by the system—and aggregated to make inferences about them? This is an important question, because if we know more about what situational characteristics are already cause for concern from the user's perspective, we might be able to create systems that are more transparent in the right places about what the system can infer about them.

The results of this study reflect the general trend that participants who were asked about Facebook were more likely to report concern about unwanted access than participants asked about Google. After controlling for participants' level of Internet Literacy and Privacy Preferences, participants were most likely to express concern in the Facebook:Ad conditions, while participants in the Google:Link:Low Sensitivity condition were the least likely group to express concern in the entire study. There is also some evidence in participants' explanations to suggest that they believed clicking a link in Facebook discloses information about them, but that if the same action is part of a Google Search it is not a disclosure. For example, a participant in the Facebook:Link condition wrote, "I hate that facebook knows what im interested in especially when I don't consent it [sic]," indicating that he or she believes Facebook learns about users' interests from what links they click on in the News Feed. In contrast, a participant in the Google:Link condition wrote, "I would not be concerned. I clicked the link and it took me to the place that I wanted" which reflects the perception that links in search results are for navigation only.

Ads in Facebook were more a source of concern for participants than ads in Google, because they perceived that Google ads were associated with search queries (that participants just wouldn't enter if they were sensitive), while Facebook ads were associated with personal characteristics (that participants might not want to reveal). Ads on Facebook contain evidence of aggregation. They're like little windows, not into what the system has collected about users, but into what the system has inferred about them. However, even targeted ads on Google were perceived to only reveal information that the user already gave to Google: the search query. Google may simultaneously provide both a greater feeling of control (over what search terms are entered and what happens when links are clicked), and less feedback that data aggregation is taking place (via the perception that ads are only related to search terms, not profiles).

The main difference between social versus information privacy is the behind-the-scenes aggregation and analysis that is pervasive when interacting with systems, but that does not take place when interacting with other people. The individual bits of information we reveal mean something different, in isolation, than they do as part of a processed aggregate. The invisibility of the infrastructure, from the users' perspective, is both blessing and curse: personalization holds the promise of better usability and access to information, but at the same time the fact that we can't see it makes it harder for us to understand its implications [8].

Most design and policy solutions for privacy issues assume a boundary management model, either by creating mechanisms for specifying what information should be revealed to whom, by providing information about what will be collected and how it will be used and allowing users to opt in or out (notice and choice), or by describing who has rights to ownership and control of data and metadata. The regulatory environment surrounding digital privacy relies on stakeholders to report violations [38], but this is not possible if users cannot tell violations are happening, nor are there laws and mechanisms in place for users to correct mistaken inferences that a system has made about them. Boundary management solutions rely on knowledge and awareness on the part of the user that data is being collected and used.

This study highlights a challenge for privacy research and system design: we must expand our understanding of user perceptions of data aggregation and when feedback about it triggers information privacy concern, so that we might design systems that support better reasoning about when and how systems make inferences that disclose too much. If users are presently unable to connect their behaviors online with the occurrence of unwanted access via inferences made by algorithms, then the current notice and choice practices do not have much chance of working. However, if there are cues in particular situations that users are already picking up on, like ads in Facebook that allow users a glimpse of what the system thinks it knows about them, perhaps the research community can build on these and invent better ways to signal to users what can be inferred rom the data collected about them.

6. ACKNOWLEDGMENTS

Thank you to the BITLab research group at MSU for helpful discussions about this project, and to Paul Rose for assisting with the content analysis. This material is based upon work supported by the National Science Foundation under Grant No. IIS-1217212. The AT&T endowment to the TISM department at MSU also provided support for this project.

7. REFERENCES

[1] L. Agarwal, N. Shrivastava, S. Jaiswal, and S. Panjwani. Do Not Embarrass: Re-Examining User Concerns for Online Tracking and Advertising. In *SOUPS 2013*, pages 1–16, July 2013.

[2] I. Altman. Privacy: A Conceptual Analysis. *Environment and Behavior*, 8(1):7–29, Mar. 1976.

[3] a. J. Berinsky, G. a. Huber, and G. S. Lenz. Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis*, 20(3):351–368, 2012.

[4] J. Bonneau and S. Preibusch. The Privacy Jungle: On the Market for Data Protection in Social Networks. In *Workshop on the Economics of Information Security (WEIS)*, May 2009.

[5] J. T. Child, J. C. Pearson, and S. Petronio. Blogging, Communication, and Privacy Management: Development of the Blogging Privacy Management Measure. *JASIST*, 60(10):217–237, 2009.

[6] S. Das and A. Kramer. Self-Censorship on Facebook. In *ICWSM 2013*, 2013.

[7] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting Depression via Social Media. In *ICWSM '13*, July 2013.

[8] R. de Paula, X. Ding, P. Dourish, K. Nies, B. Pillet, D. Redmiles, J. Ren, J. Rode, and R. S. Filho. Two Experiences Designing for Effective Security. In *SOUPS 2005*, pages 25–34, 2005.

[9] D. A. Dillman, J. D. Smyth, and L. M. Christian. *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. Wiley, Hoboken, NJ, 3 edition, 2009.

[10] C. Duhigg. How Companies Learn Your Secrets. *New York Times*, Feb. 2012.

[11] S. Egelman, A. Oates, and S. Krishnamurthi. Oops, I Did It Again: Mitigating Repeated Access Control Errors on Facebook. In *CHI '11*, pages 2295–2304, 2011.

[12] E. Gilbert. Designing social translucence over social networks. In *CHI '12*, pages 2731–2740, New York, New York, USA, 2012. ACM Press.

[13] M. J. Halter. The stigma of seeking care and depression. *Archives of Psychiatric Nursing*, 18(5):178–184, Oct. 2004.

[14] E. Hargittai and Y. P. Hsieh. Succinct Survey Measures of Web-Use Skills. *Social Science Computer Review*, 30(1):95–107, 2011.

[15] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic prediction based on user's browsing behavior. *WWW '07*, page 151, 2007.

[16] S. Kairam, M. Brzozowski, D. Huffaker, and E. H. Chi. Talking in Circles: Selective Sharing in Google+. *CHI 2012*, pages 1065–1074, 2012.

[17] A. Korolova. Privacy Violations Using Microtargeted Ads: A Case Study. *Journal of Privacy and Confidentiality*, pages 27–49, 2011.

[18] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *PNAS*, 110(15):5802–5805, 2013.

[19] J. R. Landis and G. G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, Mar. 1977.

[20] E. Litt. Understanding social network site users' privacy tool use. *Computers in Human Behavior*, 29(4):1649–1656, 2013.

[21] Y. Liu, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Analyzing Facebook Privacy Settings: User Expectations vs. Reality. In *IMC 2011*, pages 1–7, 2011.

[22] S. T. Margulis. Three theories of privacy: An overview. In *Privacy Online: Perspectives on Privacy and Self-Disclosure in the Social Web*, pages 9–18. Springer Verlag, 2011.

[23] L. A. Martin, H. W. Neighbors, and D. M. Griffith. The Experience of Symptoms of Depression in Men vs Women: Analysis of the National Comorbidity Survey Replication. *JAMA Psychiatry*, Aug. 2013.

[24] M. a. Moreno, L. a. Jelenchick, K. G. Egan, E. Cox, H. Young, K. E. Gannon, and T. Becker. Feeling bad on Facebook: depression disclosures by college students on a social networking site. *Depression and Anxiety*, 28(6):447–455, 2011.

[25] N. Nikiforakis, A. Kapravelos, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna. Cookieless Monster: Exploring the Ecosystem of Web-based Device Fingerprinting. In *IEEE Symposium on Security and Privacy*, pages 1–15, 2013.

[26] H. Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford Law Books. Stanford Law Books, 2009.

[27] S. Panjwani and N. Shrivastava. Understanding the Privacy-Personalization Dilemma for Web Search: A User Perspective. In *CHI 2013*, pages 3427–3430, 2013.

[28] K. Purcell, J. Brenner, and L. Rainie. *Search Engine Use 2012*. Pew Research Center's Internet & American Life Project, Washington, D.C., Mar. 2012.

[29] E. Rader, A. Velasquez, K. D. Hales, and H. Kwok. The gap between producer intentions and consumer behavior in social media. In *GROUP '12*. ACM Request Permissions, Oct. 2012.

[30] L. Rainie, S. Kiesler, R. Kang, and M. Madden. *Anonymity, Privacy, and Security Online*. Pew Research Center's Internet & American Life Project, Washington, D.C., Sept. 2013.

[31] S. Sengupta. On Facebook, 'Likes' Become Ads. *New York Times*, May 2012.

[32] A. Sharma and D. Cosley. Do Social Explanations Work? Studying and Modeling the Effects of Social Explanations in Recommender Systems. In *WWW '13*, pages 1133–1143, 2013.

[33] S. Silfverberg, L. A. Liikkanen, and A. Lampinen. "I'll press Play, but I won't listen": Profile Work in a Music-focused Social Network Service. In *CSCW 2011*, pages 207–216, 2011.

[34] N. Singer. You for Sale: Mapping, and Sharing, the Consumer Genome. *New York Times*, June 2012.

[35] M. Sleeper, R. Balebako, and S. Das. The Post that Wasn't: Exploring Self-Censorship on Facebook. In *CSCW '10*, pages 793–802, 2013.

[36] H. J. Smith, T. Dinev, and H. Xu. Information Privacy Research: An Interdisciplinary Review. *MISQ*, 35(4):989–1016, Nov. 2011.

[37] H. J. Smith, S. J. Milberg, and S. J. Burke. Information Privacy: Measuring Individuals' Concerns about Organizational Practices. *MISQ*, 20(2):167–196, 1996.

[38] D. J. Solove. Introduction: Privacy self-management and the consent dilemma. *126 Harvard Law Review*, pages 1880–1903, 2013.

[39] F. Stutzman and W. Hartzog. Boundary Regulation in Social Media. In *CSCW 2012*, pages 769–778, 2012.

[40] E. Toch, Y. Wang, and L. F. Cranor. Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems. *User Modeling and User-Adapted Interaction*, 22(1-2):203–220, 2012.

[41] B. Ur, P. L. Leon, L. F. Cranor, R. Shay, and Y. Wang. Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising. In *SOUPS '12*, 2012.

[42] A. F. Westin. Social and Political Dimensions of Privacy. *Journal of Social Issues*, 59(2):431–453, Apr. 2003.

[43] C. E. Wills and C. Tatar. Understanding What They Do with What They Know. In *WPES 2012*, pages 13–18, 2012.

APPENDIX

A. SURVEY QUESTIONS

Data collected: May 10 – 16, 2013

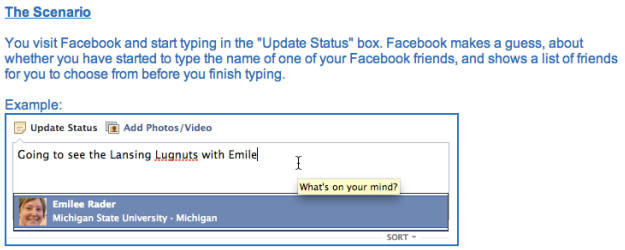
Sample: 701 Amazon Mechanical Turk workers who were 18 or older, had a 95% or higher approval rating after completing at least 500 tasks, and reported in the screening questionnaire that they visited both Facebook and Google Search at least weekly.

A.1 The Scenarios

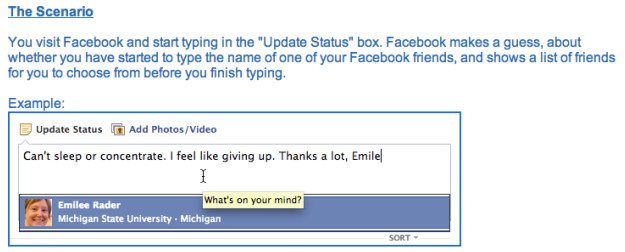
In this section of the survey, you will be shown an example of a scenario people often encounter when using Facebook or Google Search.

As you read the scenario, please think about what it would be like for you to experience something like it.

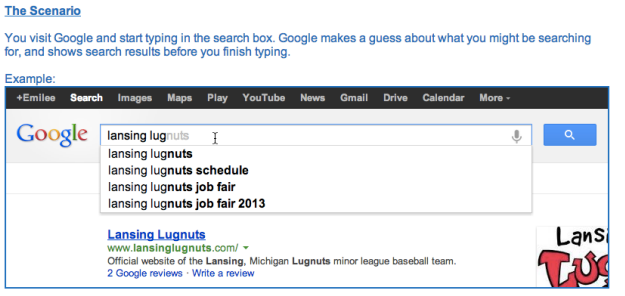
Autocomplete, Facebook, Non-Sensitive.



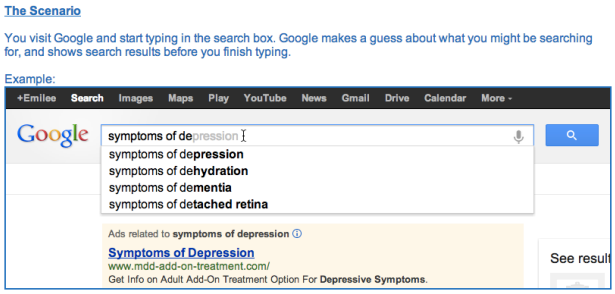
Autocomplete, Facebook, Sensitive.



Autocomplete, Google, Non-Sensitive.

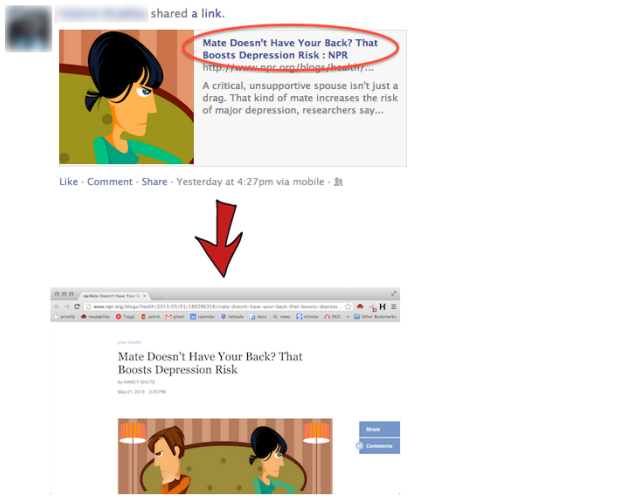


Autocomplete, Google, Sensitive.



Link, Facebook, Non-Sensitive.

Link, Facebook, Sensitive.

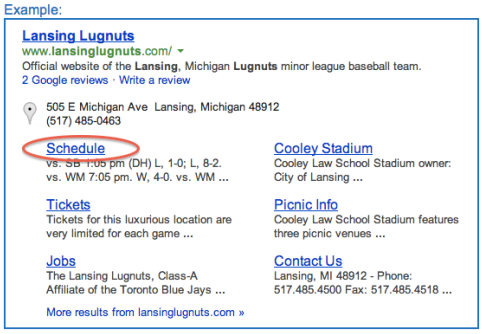


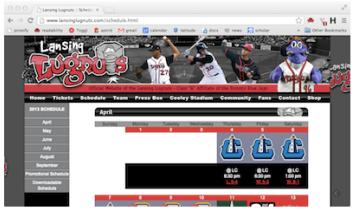
Link, Google, Non-Sensitive.

The Scenario

You are viewing the results of a Google search. You scroll down the page to find the information you are looking for, and then click on the link. The page changes to display the web page for the link that you clicked on.

Example:



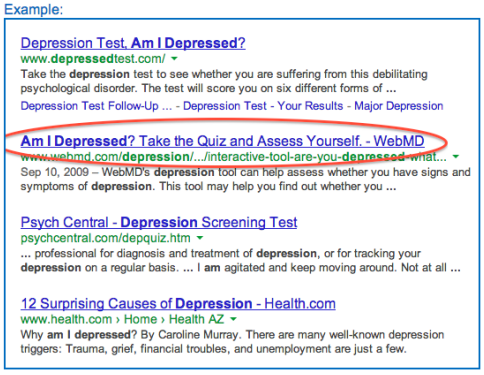


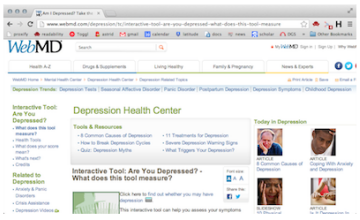
Link, Google, Sensitive.

The Scenario

You are viewing the results of a Google search. You scroll down the page to find the information you are looking for, and then click on the link. The page changes to show the web page for the link that you clicked on.

Example:



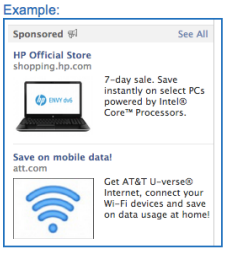


Ad, Facebook, Non-Sensitive.

The Scenario

You are viewing posts in your Facebook News Feed. As scroll down the page, reading posts made by Facebook friends, you notice ads displayed on the right side of the screen.

Example:

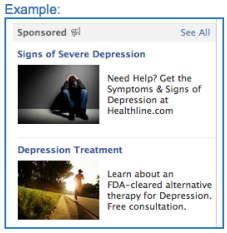


Ad, Facebook, Sensitive.

The Scenario

You are viewing posts in your Facebook News Feed. As scroll down the page, reading posts made by Facebook friends, you notice ads displayed on the right side of the screen.

Example:

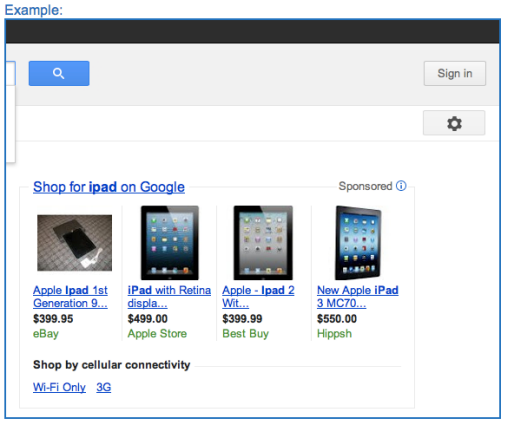


Ad, Google, Non-Sensitive.

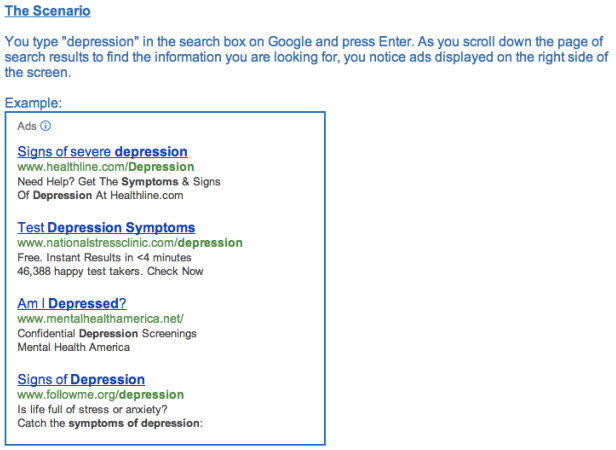
The Scenario

You type "ipad" in the search box on Google and press Enter. As you scroll down the page of search results to find the information you are looking for, you notice ads displayed on the right side of the screen.

Example:



Ad, Google, Sensitive.



A.2 Concern

Q1 Would you be concerned about unwanted access to private information about you in this scenario? (Yes=151, Maybe=173, No=377)

Q2 Please explain your answer to the previous question. (open-ended)

Q3 What would you tell someone else about how to control private information in the above scenario? Please describe what you would say, below. (open-ended)

A.3 Information Types

AWARENESS How likely do you think it is that [Google | Facebook] can AUTOMATICALLY record each of the following types of information about you? Please indicate below how likely you believe each example is on a scale from 0-100, where 0 means Unlikely, and 100 means Likely.

M	*SD*	
92.0	15.6	what time of day you visit [Google | Facebook]
84.9	19.9	your physical location when using [Google | Facebook] on a mobile device
65.0	32.9	what you are typing in the [search | Post or Comment] box before you submit the [search terms | post]
80.0	25.5	how much time you spend reading [Google | Facebook] status updates
71.8	30.6	what kind of computer you are using when you visit [Google | Facebook]
81.2	23.9	your physical location when using [Google | Facebook] on a computer
19.7	25.1	what you purchase from a brick-and-mortar store
60.1	31.2	what you purchase from online shopping websites
69.6	29.8	what websites you visit most often
69.5	30.5	which people you communicate with online most often
50.8	32.7	your political party affiliation
93.2	13.9	how often you visit [Google | Facebook]
50.6	34.7	your sexual orientation
19.1	24.0	what the desktop image on your computer looks like
71.1	29.0	which online retailers (e.g. Amazon.com) you visit most often
91.0	16.2	what links you click on in your [Google search results pages | Facebook news feed]

A.4 Privacy Preferences

PRIVACY PREFS Here are some statements about personal information. From the standpoint of personal privacy, please indicate how much you agree or disagree with each statement below. [Strongly Disagree (1) Disagree (2) Neutral (3) Agree (4) Strongly Agree (5)]

M	*SD*	
4.36	0.82	If I think that information I posted to Facebook really looks too private, I might delete it.
4.08	4.27	I don't post to Facebook about certain topics because I worry who has access.
2.93	1.20	I use shorthand (e.g., pseudonyms or limited details) when discussing sensitive information on Facebook so others have limited access to know my personal information.
4.03	0.90	I like my Facebook status updates to be long and detailed. REVERSE CODE
4.17	0.95	I like to discuss work concerns on Facebook. REVERSE CODE
4.36	0.81	I have limited the personal information that I post to Facebook.
3.81	1.05	When I face challenges in my life, I feel comfortable talking about them on Facebook. REVERSE CODE
3.71	1.05	When I see intimate details about someone else on Facebook, I feel like I should keep their information private.
4.33	0.88	When people give personal information to a company for some reason, the company should never use the information for any other reason.
3.99	0.96	It usually bothers me when companies ask me for personal information.
4.42	0.90	Companies should never sell the personal information in their computer databases to other companies.
3.83	1.01	I'm concerned that companies are collecting too much personal information about me.

A.5 Scenario Realism

AUTOCOMPLETE only Search engines and social media websites can make a guess about what you are about to type, while you are typing, and provide you a list of suggestions – like in the scenario displayed at the beginning of this survey. Have you ever used a website that has this "autocomplete" functionality? [Yes=227, No=6]

LINK only Search engines and social media websites provide links (URLs) to content on other websites containing information that is interesting, entertaining, etc. – like in the scenario displayed at the beginning of this survey. Have you ever clicked on a link in a search engine or social media website that took you to content on some other website? [Yes=224, No=10]

AD only Search engines and social media websites can display personalized or "targeted" advertising – like in the scenario displayed at the beginning of this survey. Have you ever noticed "targeted" advertising when surfing the web? [Yes=228, No=6]

A.6 Internet Literacy and Experience

INTERNET LITERACY How familiar are you with the following Internet-related terms? Please rate your familiarity with each term below from None (no understanding) to Full (full understanding): [None (1) Little (2) Some (3) Good (4) Full (5)]

	None	Little	Some	Good	Full
Wiki	1	23	52	187	438
Netiquette	129	61	121	175	215
Phishing	18	48	92	225	318
Bookmark	4	7	22	146	522
Cache	11	44	137	236	273
SSL	171	159	136	113	122
AJAX	409	131	83	37	41
Filtibly (FAKE WORD)	587	85	29	0	0

E1 Have you ever worked in a Òhigh techÓ job such as computer programming, IT, or computer networking? [Yes=115, No=586]

E2 How often do you visit Facebook?

Once a Week or less	6
2-3 Times a Week	88
Daily	246
Many times per day	361

E3 How often do you search the web using Google? [Once a Week or less, 2-3 Times a Week, Daily, Many times per day]

Once a Week or less	1
2-3 Times a Week	15
Daily	137
Many times per day	548

E4 Do you use ad blocking software when you browse the web? [Yes=536, No=144, Don't Know=21]

E5 Have you ever had one of the following experiences? Please check all that apply:

No	Yes	
89	612	Received a phishing message or other scam email
34	667	Warning in a web browser that says "This site may harm your computer"
57	644	Unwanted popup windows
154	547	Computer had a virus
646	55	Someone broke in or "hacked" the computer
503	198	Stranger used your credit card number without your knowledge or permission
687	14	Identity theft more serious than use of your credit card number without permission
691	10	None of the above

A.7 Demographics

D1 How old are you? Please write your answer here: [*M*=30.2, *SD*=9.22]

D2 What is the last grade or class you completed in school?

0	None, or grades 1-8
2	High school incomplete (grades 9-11)
71	High school graduate (grade 12, GED certificate)
20	Technical, vocational school AFTER high school
285	Some college, no 4-year degree
241	College graduate (B.S., B.A., 4-year degree)
27	Post-graduate
3	Other
0	I Don't Know

D3 What is your gender? [Man=398, Woman=297, Prefer not to answer=6]

D4 What is your race?

American Indian or Alaska Native	4
Asian or Pacific Islander	63
Black or African-American	41
Hispanic or Latino	26
White	560
Other	7

D5 Which of the following BEST describes the place where you now live?

A large city	155
A suburb near a large city	256
A small city or town	211
A rural area	78
Other	0
Don't know	1

D6 Most people see themselves as belonging to a particular class. Please indicate below which social class you would say you belong to:

Lower class	41
Working class	173
Lower middle class	141
Middle class	276
Upper middle class	69
Upper class	1
Other	0

D7 Are you now employed full-time, part-time, retired, or are you not employed for pay?

Employed full-time	310
Employed part-time	94
Retired	6
Not employed for pay	77
Self-employed	85
Disabled	11
Student	104
Other	14

B. CONTENT ANALYSIS

Respondents were asked to explain why they answered (Yes, Maybe, or No) to a question that asked, "Would you be concerned about unwanted access to private information about you in this scenario?"

The purpose of this coding scheme is to differentiate between two potential themes that appeared in many respondents answers. These themes are informed by the distinction in the literature between "social" privacy – or control over information in relation to other people, and "informational" privacy, or control over information in relation to technologies, organizations or the government.

Each answer should be coded "INFO", "SOCIAL" or "NEITHER".

Step 1. Determine whether the response contains an explicit reference to a potential third party accessing/obtaining information related to the respondent.

If the answer contains no clear reference to a third party, or does not implicate accessing/obtaining respondent info, or does not provide evidence that the coder can use to tell whether the third party access is "social" or "informational", code as NEITHER. Otherwise, proceed to Step 2

In general, responses with ambiguous pronouns without an explicit referent (e.g. "they", "them", "it") should be coded as NEITHER, because without more information from the respondent, it is impossible to tell whether the referent is a person, organization, government, or website. For example, "Really depends on exactly what kind of information they gathered. I am OK with just basic information".

Likewise, the presence of passive voice (e.g. "Private information is being read from my posts"), should be coded as NEITHER, because these responses typically do NOT constitute an explicit reference that allows the coder to differentiate who or what the third party is.

However, there are exceptions to the above. To proceed to Step 2 with a response that contains ambiguous pronouns or passive voice, the response must contain some other evidence that allows the coder to determine whether the potential for unwanted access is SOCIAL- or INFO-related.

This evidence often comes in the form of mentioning ads, IP addresses, databases, or some other technology or feature as if it is involved in information collection, access, or processing. For example, "It would really depend on what kind of information. Not much I can do about them using my IP address to localize the type of ad"; or, "I'm aware that certain things about me are known and will be used to select ads, and I don't mind that".

Step 2. Determine whether the explicit reference to third party access in the response includes evidence that the third party is a human being, or a group of people.

This could include language like "other people", "employers", "friends", "others", "anyone", etc. Pronouns such as "it" and "they" should NOT be treated as SOCIAL, unless the referent is present in the response. If the answer contains evidence that the third party is clearly a person or group of people, code as SOCIAL. If not, code as INFO.

Some answers might legitimately contain references to both people and organizations, governments, or websites. In these cases, try to determine from the response which aspect, SOCIAL or INFO, is causing more concern for the respondent. If it is not clear, code as SOCIAL. Example: "I wouldn't be concerned because even if google is keeping track of what all of their subscribers are looking up, there are so many people in the world that the chances of anyone looking at my individually are slim to none."

B.1 Examples, Site:Code:Concern

Facebook:INFO:Yes.

- I do not feel that ANYTHING that I say on my facebook account is private. It makes me feel strange when a computer is second guesing me before I finish typing.
- It's never comfortable for ad companies to have private information about me.

Facebook:INFO:No.

- I am posting a facebook status on facebook. I don't mind that facebook is guessing who I might be tagging in my facebook status post. All that information can be found on facebook.
- The ads seem random to me and doesn't have anything to do with me.

Google:INFO:Yes.

- I don't believe search information should be logged and associated to persons.
- Most people know that search engines, ESPECIALLY Google, collect all sorts of information about people and then pass it on to the government.

Google:INFO:No.

- I don't care if google knows what I search. I have no secrets.
- The ads are only coming up based on my search. The ads could be helpful.

Facebook:SOCIAL:Yes.

- I'm not sure I want people to know what website's I have been to.
- I am very concerned about my privacy anyway, especially when it comes to things shared on Facebook and other social networks.

Facebook:SOCIAL:No.

- Because, I just clicked on the link. I only would be concern if facebook announced on the news feed that I read the article.
- Because no one else sees me typing in the box and I already know who my friends are, can see my friends list, etc.

Google:SOCIAL:Yes.

- I would be concerned that someone could find out my search for depression by checking my Google search history, and that they keep a record of that when they display ads to me.
- I hate when previous searches pop up while someone is browsing my computer.

Google:SOCIAL:No.

- I am not sure how my provacy would be jeopardized in this scenario. Even if it were, I don't think I'd be concerned if someone were to find out I was searching for help with depression.
- Those ads are automatically displayed to anyone who enters in a particular search term. They don't have anything to do with me individually. I don't see any indications that any information was revealed to the people who placed the ads.