Title
KnowPrivacy

Authors
Gomez, Joshua
Pinnick, Travis
Soltani, Ashkan

Publication Date
2009-10-10

KnowPrivacy

Joshua Gomez, Travis Pinnick, Ashkan Soltani
School of Information, UC Berkeley

UC Berkeley School of Information Report 2009-037
10 October 2009

Abstract

Online privacy and behavioral profiling are of growing concern among both consumers and government officials. In this report, we examine both the data handling practices of popular websites and the concerns of consumers in an effort to identify problematic practices. We analyze the policies of the 50 most visited websites to better understand disclosures about the types of data collected about users, how that information is used, and with whom it is shared. We also look at specific practices such as sharing information with affiliates and third-party tracking. To understand user concerns and knowledge of data collection we look at surveys and polls conducted by previous privacy researchers. We look at records of complaints and inquiries filed with privacy watchdog organizations such as the Federal Trade Commission, the Privacy Rights Clearinghouse, The California Office of Privacy Protection, and TRUSTe. Finally, to gain some insight into what aspects of data collection users are being made aware of, we look at news articles from three major newspapers for topics related to Internet privacy. Based on our findings we make recommendations for website operators, government regulators, as well as technology developers.

Contents

1. Introduction

1.1 Goal

In the spring of 2009, U.S. government officials began expressing growing concern about consumer privacy on the Internet. Lawmakers and regulators are particularly concerned about "behavioral advertising," the use of internet-based technologies to collect information for purposes of targeting advertisements to individual consumers.

Federal Trade Commission (FTC) Chairman Jon Leibowitz expressed disappointment with what he characterized as the industry's inability to effectively self-regulate, and announced that the industry was near its last chance [1]. Earlier in the year, the FTC had issued revised guidance urging website operators to tell consumers that data is being collected for behavioral advertising purposes and to provide a clear and easy-to-use means to opt out [2]. In April, a House subcommittee met to discuss possible legislation to regulate the practice [3]. Congressman Rick Boucher is planning to conduct a joint hearing with the Subcommittee on Commerce, Trade and Consumer Protection in the early summer to examine online privacy, including behavioral advertising [4].

The goal of this project was to examine both the data handling practices of popular websites and the concerns of consumers in an effort to identify practices which may be deceptive or potentially harmful to users' privacy and, based on our findings, offer potential solutions that policymakers should consider when discussing any new Internet privacy regulations or that website operators could implement to potentially avert or soften such regulation.

1.2 Design

In this project we examined the common practices among website operators of collecting, sharing and analyzing data about their users. We attempted to identify practices which may be deceptive or potentially harmful to users' privacy and we make recommendations for changes in industry practice or government regulations accordingly. We compared industry practices with users' expectations of privacy, identified points of divergence, and developed solutions for them.

To make this comparison we assembled a picture of practices and perceptions through data from several sources. First, to assess users' perceptions, expectations and knowledge, we gathered data from surveys of public opinions found in previous research done by various public policy and polling organizations. Next, we analyzed which practices upset them enough to file complaints with privacy watchdog organizations such as the FTC, the Privacy Rights Clearinghouse, the California Office of Privacy Protection, and TRUSTe. Finally, we looked at popular media to get a sense of what is being discussed in stories about Internet privacy, what users are made aware of, and what they may not know about.

To get a corresponding understanding of website practices, we conducted our own survey of website privacy policies, identifying the types of data that sites collect about users, the purposes for which that data is used, and with whom that data is shared. From this general picture, we narrowed our focus to specific behaviors. We looked specifically at the use of third-party tracking beacons, which are usually excluded from the provisions laid out in a website's privacy policy. We also investigated the practice of sharing data with "affiliates."

From these various sources of data we identified points of conflict between the privacy expectations of Internet users and the actual practices of website operators.

In this report, we first provide background information that describes how Internet companies collect information about users and a brief summary of current regulations. Then we discuss our methods and findings with respect to user expectations and website practices.

2. Data Collection and Behavioral Profiling

2.1 Collection

2.1.1 Typical Website Tracking

When a user visits a website, the server automatically collects certain information about the visitor, such as IP address, web browser and operating system type, the page visited, the referring page, and the time of the visit. To keep track of a user while visiting various pages on a website, the operator may install a "cookie" on the user's machine. The cookie is a simple text file, usually containing a unique identifying number. Some cookies are temporary and some may be retained on the hard drive and used for multiple visits. If the website requires login or registration information, it can correlate personally identifiable information (PII) with browsing behavior.

2.1.2 Third-party Tracking

2.1.2.1 Cookies

Many websites are advertising supported, and typically, the ad images for these sites are not served directly from the main website operator. Instead they are pulled from the servers of the advertisers or an advertising network. In the process, advertisers can place cookies on the user's machine. Since the advertisers place ads on multiple sites, the cookie allows the advertiser to observe the user's browsing behavior across many websites. Large ad-serving agents, such as DoubleClick or Zedo, span significant portions of the World Wide Web and thereby acquire extensive behavioral data.

Another type of third-party tracking is completely invisible to users. Web Bugs enable third parties that do not even serve ads to place cookies on a user's browser and track the user's navigation across the web.

2.1.2.2 Web Bugs

Web bugs are embedded in a web page's HTML code, and are designed to enable monitoring of who is reading the page. Web bugs are typically a small graphic embedded in the page, usually an invisible 1-by-1 pixel, and are also called "web beacons," "clear GIFs," or "pixel tags." Other methods of creating tracking bugs exist, such as using JavaScript code. Ad networks can use web bugs to aggregate information to create a profile of what sites a person is visiting. The personal profile is identified by the browser cookie of an ad network, allowing the network to track behavior across sites over time. Information web bugs may transmit to a server include:

- The IP address of the computer that fetched the web bug
- The URL of the page where the web bug is located (which essentially reveals content)
- The time the web bug was viewed
- The type of browser that fetched the web bug image
- A previously set cookie value

Blocking web bugs is difficult. One defense is to disable third-party cookies, thereby limiting the types of information they can collect and associate with personally identifiable information. However, not all browsers have this functionality. Furthermore, blocking third-party cookies does not remove the web bug itself, since it is part of the web page and not the cookie. Removal of the cookie prevents the tracker from identifying the individual user.

However, it would still have the capability to track navigation data using IP address as an identifier. In cases where a user maintains a static IP, that may be all that is necessary match a profile to an individual user. A dynamic IP address can be linked once the user makes a conversion by logging into or making a purchase on a website that shares such information with the tracking network.

A user could install a plug-in that blocks all third-party content, including bugs. However, this solution would also remove much desirable content, such as embedded media files. Additionally, newer tracking methods, such as flash cookies are not easily controlled by the user but are increasingly used to store user identification information.

2.2 Use

Website operators can use information about user behavior for various purposes. They can use the data for the development and improvement of the website, making it easier to use. They can customize a site to fit individual users' tastes. An e-commerce site can make product recommendations based on previous purchases or they can use the information to deliver targeted ads. Many of these uses benefit the visitors to the site and are actively sought by consumers.

2.3 Sharing

Sometimes site operators will rent or sell personal and behavioral data about users to third parties. More often, the operators will share the data with marketing partners or corporate affiliates and subsidiaries, meaning that user behavior may be profiled not only by sites visited by a user, but also by any other entities with whom those sites may choose to share this information.

However, sometimes it is unclear what a website means by the terms "affiliate," "third party" and "partner." Our analysis of privacy policies found that many stated they do not share data with third parties, but they do share data with affiliates, suggesting that they only share data with companies under the same corporate ownership. However, many of these websites also allow third parties to track user behavior directly through the use of web bugs. In a conversation with one of the website's Chief Privacy Officer, he claimed that they consider the advertising serving company DoubleClick to be a "marketing partner," and not a third party.

2.4 Aggregators

Just as these site operators can sell data about users, they can also purchase more data about them to build better profiles, a process referred to as enhancement. Some companies, such as ChoicePoint, base their entire business model on the aggregation and selling of personal information. These data brokers acquire information from phone books, court documents, voter registries, and other public records. Some data brokers have websites where much of this information can be found or purchased by anyone. In our analysis of privacy policies, about a quarter of the websites expressly stated that they buy information about users from third parties to supplement data collected directly from their users.

3. Current Regulation

3.1 Federal Legislation

The United States follows a sectoral model for privacy regulation, where certain sectors or business models of the economy are regulated. This leaves significant gaps between sectors. E-commerce is largely governed by two laws—the Children's Online Privacy Protection Act of 1999 (15 U.S.C. § 6501–6506), and the growing "common law" of privacy created by Federal Trade Commission enforcement actions. Self-regulation plays a major role in US privacy protections.

3.2 Self-Regulation

The Federal Trade Commission will enforce cases against companies that fail to deliver on privacy promises, or that engage in practices that are so injurious that they arise to "unfairness." However, it does not act on individual complaints, nor has it issued rules for how companies should collect, process, and disclose personal information. It does suggest that companies adhere to commonly-accepted principles for handling personal information, known as Fair Information Practices.

3.2.1 Fair Information Practices (FIPs)

In an effort to directly address the issues of data collection, the FTC updated its list of Fair Information Practice Principles in 2007 [5], and in February 2009, revised its principles for Online Behavioral Advertising [2]. The FIPs are a set of guidelines for data collecting entities to make their practices more protective of consumer privacy. The FIPs consist of five core principles: notice, choice, access, security, and enforcement. The first four are meant to make consumers aware of data collection, enable them to control what it is used for, let them see the data that has been collected, and ensure that the data is correct and secure. Finally, the enforcement principle suggests that some method of enforcement be used: either industry self-regulation or governmental regulation through private remedies or civil/criminal sanctions.

Critics of the FIPs often point out their relative weakness when compared to rules in other regions, such as the OECD (the Organization for Economic Co-operation and Development) set of FIPs, or even in other US agencies, such as the Department of Homeland Security [6].

FIPs are intended to create rights for users and responsibilities for data collectors. However, since they are not formally codified, internet companies have a disincentive to restrict their data collection practices. Analysis of their users' behavior and preferences can help site operators make their site more appealing and therefore gain an advantage over competitors. Selling customer profiles to direct marketers is also a valuable source of revenue. If users feel that their privacy has been invaded by data collection they have little recourse, as self-regulation creates an imbalance of power between users and site operators.

3.2.2 Trust Seals

One method of self regulation is the use of trust seals, such as those offered by TRUSTe or BBBOnline. These organizations advise data collecting companies on ways to improve their practices. Once a set of standards are met, the websites are allowed to display an icon signifying their compliance, thus creating a sense of trust and security on the part of the users. It was hoped that this method of self regulation would create a market for privacy protection.

Unfortunately, the seal programs have not gained wide adoption.[*] While the goals of certification authorities are admirable, critics have pointed out that they often have not achieved their desired effect [7]. This happens because many popular websites have insufficient incentives to participate in the programs because users are already comfortable using their sites. Meanwhile, many websites that have very poor privacy protections, or even deliberately exploitative practices, are eager to get the seal to gain more users [8]. Indeed, as security expert Ross Anderson stated, "certification markets can easily be ruined by a race to the bottom; dubious companies are more likely to buy certificates than reputable ones, and even ordinary companies may shop around for the easiest deal" [9].

3.2.3 Privacy Policies

In 2003, California enacted the Online Privacy Protection Act, which requires website owners to conspicuously post a statement of their policies regarding the collection and sharing of personal information [10]. The goal of this legislation was to create some transparency in data collection practices and to help users make informed decisions. However, the legislation does not regulate the substance of websites' practices; they only need to disclose those practices.

Like the trust seal programs, the requirement of privacy policies was an attempt to create a marketplace for privacy. By creating transparency, it was hoped that users could make informed decisions about which sites they use based on the site's data collection practices. However, most users do not even read privacy policies [11], and therefore little change has been made in data collection practices.

There are several reasons that privacy policies are ineffective:

Privacy policies are difficult to read. Most privacy policies are written in legal jargon that is difficult for an average person to understand [12]. Because they cannot understand the policies, most users do not even bother to read them.

Framing: privacy policies lead consumers to believe that their privacy is protected. In fact, a 2008 study found that "they do not read privacy policies because they believe that they do not have to; to consumers, the mere presence of a privacy policy implies some level of often false privacy protection" [13].

Even if they could understand them, *the amount of time required to read privacy policies is too great.* A 2008 study estimated that if users actually read privacy policies, it would take approximately 200 hours a year to read the policy for every unique website visited in a year, not to mention updated policies for sites visited on a repeating basis [14].

Even if they could understand and had the time to read policies, *there is not enough market differentiation for users to make informed choices.* A 2006 analysis of trends in privacy policies found that a strong majority of websites collect both computer and click stream data, as well as contact and uniquely identifying information. It also found that popular websites are more likely than randomly selected sites to collect more types of data [15]. Furthermore, many website policies are vague about what information they collect and how it is used.[†] Because they are all equally poor, users have no viable alternatives. This is a market failure.

Finally, even if there was market differentiation, it is not clear that users would protect themselves. The *potential dangers are not salient* to most users. And even when they are salient, *they are difficult to evaluate* against the benefits of using a particular website. Thus, *most users*

[*] In our own survey of privacy policies, only 30 of the 100 most-visited websites displayed a TRUSTe or BBBOnline seal.

[†] Results from our research confirm the homogeneity of website practices: 72% allow third-party tracking and 88% share data with affiliates.

rely on heuristics and suffer from cognitive biases, such as anchoring, hyperbolic discounting and valence effect [16,17,18,19].

Ultimately, the privacy policy solution suffers from the same problems of misaligned incentives as the trust seal programs. There is an incentive for websites to collect and share data about their users. This incentive should be balanced by the market and consumer choice, but users are unable to make informed decisions.

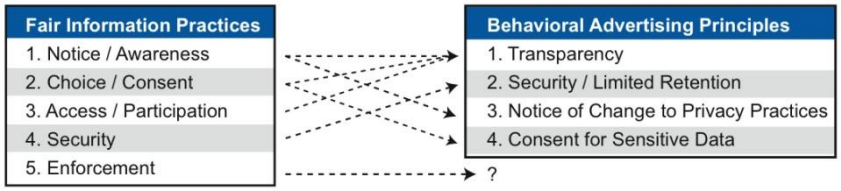
3.2.4 P3P

The Platform for Privacy Preferences (P3P) is an attempt to make the use of privacy policies easier for consumers by encoding policies into a standard machine-readable format. [20] P3P-enabled search engines, such as www.privacyfinder.org, have been built that filter out search results based on the privacy preferences of the user. Browser plug-ins that enable users to set their privacy preferences have also been developed. These plug-ins read the P3P files of visited sites and indicate whether or not it meets the user's preset criteria.

Although P3P was created to make privacy choices easier for users, some critics claim that the technology is too difficult for most users [21]. To date, the adoption rate of P3P has been fairly low. Our analysis of the top 100 websites for this project revealed that only 27 of them provided a P3P policy, and only a subset of those were valid according to the P3P standard.

3.2.3 FTC Self-Regulatory Principles for Online Behavioral Advertising

In February 2009 the FTC released a report outlining a set of self-regulatory guidelines specifically for online behavioral advertising [2]. Though similar to the original FIPs, the online behavioral advertising principles emphasize the FIPs facets differently. The FIPs Notice/Awareness principle has been combined with the FIPs Choice/Consent and FIPs Access/Participation principles to create a new principle of Transparency/Control. One component of Notice/Awareness was separated and highlighted on its own: notice of change to privacy practices. Consent for sensitive data (such as health information) was also given its own emphasis. The security principle also remains with an additional request for limited retention. Notably absent from the new set of principles is enforcement, or accountability.



4. Methods

4.1 User Expectations/Knowledge

To determine users' expectations of privacy we examined three types of data: surveys, complaints, and popular media.

4.1.1 Surveys

We aggregated various surveys of users conducted by academic researchers, corporate researchers, and public opinion polling companies. These organizations include:

The Annenberg Public Policy Center at the University of Pennsylvania

The Samuelson Law, Technology & Public Policy Clinic at UC Berkeley
The PEW Internet & American Life Project
The Harris Poll
Consumer Reports National Research Center
TRUSTe

4.1.2 Complaints

To determine what types of practices are invasive enough to compel users to complain we requested data from several outlets for users' complaints: the Federal Trade Commission (FTC), the Privacy Rights Clearinghouse (PRC), the California Office of Privacy Protection (COPP), and TRUSTe. All four organizations gave us quantitative data for complaints made in the five year period between 2004 and 2008, inclusive.

TRUSTe gave us aggregate information such as number of complaints per year, by type. The FTC, PRC, and COPP gave us data for individual complaints, such as date, company, and type of complaint. In addition to these full data sets, we also received random samples from the FTC and PRC that also included the free text fields in which the users explain why they are complaining. The FTC and PRC removed any personally identifiable information before disclosure.

4.1.2.1 FTC Data

The FTC receives complaints for various consumer issues, such as false advertising, unfair practices, and fraud. It has the authority to enforce regulations concerning these issues from section 5 of the FTC Act (15 U.S.C. §§ 41-58, as amended). Its authority is extended through other statutes such as the CAN-SPAM Act (15 U.S.C. 7701, et seq.) or the Fair Credit Reporting Act (15 U.S.C. § 1681 et seq.). Consumer complaints filed at the FTC are categorized with codes relating to the various statutes it enforces. See Appendix A for the list of all statute codes.

We made a request to the FTC under the Freedom of Information Act (FOIA) for all complaints filed in the General Privacy (GP), Gramm-Leach-Bliley (GLB), and CAN-SPAM (CS) statute codes for the five year period between 2004 and 2008, inclusive. This query returned 51,532 records.

In addition to the statute code, each complaint is also tagged with a statute violation code. See the table below for the list of violation codes under each of the three statutes we looked at, along with the number of complaints filed in each of them (note that there are no GP1, GP2, or GP3 violation codes in the list, as they are no longer used). The FTC data is not organized hierarchically, so records that have been categorized in the General Privacy category may have a violation code from one of the GLB violations, such as GLB2. The records may also be double coded with multiple statutes or violations.

Violation Code	Violation Description
GLB1	Company does not provide any opportunity for consumer to opt out of information sharing
GLB2	Company fails to honor request to opt out/ opt-out mechanism does not work
GLB3	Company is violating its privacy policy
GLB4	Privacy policy is misleading, unclear, or difficult to understand
GLB8	Other GLB violation
GP4	Company does not have adequate security
GP5	Other Privacy Violation
CS1	SPAM: "Remove Me" is missing, broken, or ignored
CS2	SPAM: Spam shows pornographic image
CS3	SPAM: Spam led to suspect information collection practices
CS4	Subject or From line is false or misleading
CS5	SPAM: Spammer misuses computer resources
CS6	SPAM: Other\general annoyance

Table 1 Selected FTC complaint violation codes

To get a better understanding of the user complaints we sent more FOIA requests for the free text fields of a sample of complaints within the GP5 violation code. One request sought the free text for a random sample of 200 complaints marked with the GP5 violation code in which the website complained about was in the top 10 of our list of most visited websites.[*]

Our analysis of the quantitative data revealed a significant number of complaints about data brokers and websites that serve as portals to them, such as ZabaSearch. Therefore, we also requested free text fields for a random sample of 200 complaints with the GP5 violation code in which the company complained about was one of the following: ZabaSearch.com, intelius.com, whitepages.com, addresses.com, or anywho.com.

All personally identifiable information was stripped from these records before disclosure.

4.1.2.2 PRC Data

The PRC also categorizes the complaints it receives from users. They have 40 different categories, such as Collection Agencies, Genetics, and Wiretapping. We requested the records of all complaints made within the same five year period, from 2004-2008, in the two categories most pertinent to our research: Cyberspace and Database/Info Broker. We received 2202 records from this request. These records did not include any fields containing information about the user.

We also requested the free text fields for a sample of complaints from the PRC within the same two categories. We received 250 records. These free text fields were stripped of all personally identifiable information before disclosure.

4.1.2.3 Coding

We categorized the free text complaints using a set of tags that matched the concern of the user as well as the type of data involved and the type of company. We ran through a pilot set of the complaints with a limited set of tags, discussed our findings, and then developed a revised set of tags that better captured the types of data involved and the concerns of the users (these "concern" category tags included user control, public display of personal data, data aggregation, potential for physical harm, security, fraud, third-party sharing, identity theft, and excessive information requested for a given transaction). A detailed list of these tags is in Appendix B. The revised set of tags was then applied to all the complaints. Two people did the coding, with 10% overlap. Within this overlap, we had an average agreement of 92% across all the tags, which is evidence of a high degree of inter-coder reliability.

[*] See section V. Methods - B. Website Practices for a description of our determination of the top 10 websites.

4.1.3 News Stories

To get a sense of what users are made aware of we looked at media discourse. We sampled news stories containing the words "internet" and "privacy" within the same paragraph over the past two years. Our sample was pulled from three major newspapers: The New York Times, The Washington Post, and the San Jose Mercury News. We chose these papers because the Times has a very wide general distribution, the Post is located in the major policy hub of the country, and the Mercury-News is located in the major Internet and technology hub of the country.

We created a set of coding tags and had an undergraduate assistant code the sampled news stories. After her initial coding of the 2008 sample, we reviewed her work, revised the codes, recoded and had her code a sample from 2007.

As we found various topics of interest in the other data sets we conducted deeper searches in Lexis/Nexis to find the volume of occurrence each topic garnered in news sources.

4.2 Website Practices

To get an understanding of the other side of the data collection interaction, we analyzed the privacy policies of the most visited websites. We also looked at the prevalence of web bugs on these sites and tried to determine with how many potential "affiliates" they could share users' data. Our determination of the most visited websites is based on the top 100[*] US website list published by web traffic monitor Quantcast (as of March 1, 2009). Quantcast's list is based on their direct measurement and estimation of unique U.S. website visitors per month to the listed sites [22].

4.2.1 Privacy Policies

Much research has been done in the area of privacy policy analysis [23] and readability [16]. Some of these projects aim to develop methods of visualizing policies in ways that are simpler for user's to understand [24]. Our approach to policy analysis differs from these previous projects. Much of the prior work focused on granular details of privacy policies. Since this project's goal is a comparison of user expectations and actual practices, our analysis of privacy policies was simplified, using only high level concepts.

To analyze the privacy policies, we created another set of coding tags, or facets. Each policy received an evaluative code of YES, NO, or UNCLEAR for each category. YES and NO codes were only assigned if the distinction could clearly be made based on the wording of the site's privacy policy. UNCLEAR was given if the given information was not specified or was to too nuanced or vague to be determined. See Appendix C for detailed definitions of our coding facets.

Finding that understanding a privacy policy is sometimes a matter of interpretation, we sent copies of our analysis to each website operator, explaining our research project and requesting verification or corrections. To date, we have received responses from seven companies, representing twelve websites from our list.

4.2.2 Web Bugs

Our investigation of web bugs was based on data from Ghostery, which is an add-on for the Firefox web browser. The data was generously provided by David Cancel, creator of the software and co-founder of Compete, Inc. Ghostery identifies and informs the user of hidden

[*] Throughout this paper when we refer to the "top 100," "top 50," or "top 10" websites, it will refer to this ranking prepared by Quantcast as of March 2009, unless specifically stated otherwise.

web bugs on the pages a user visits. The software has an optional 'GhostRank' feature that allows the software to report the web bugs found on each site visited to a central database operated by Ghostery.

Mr. Cancel provided us with data from GhostRank so that we could determine how many web bugs have been identified on each of the top 100 websites as well as how many of those sites each tracking company is present on. The data provided to us was for the entire month of March, 2009. During the month of our analysis there were approximately 300,000 users. Of those who downloaded the software, approximately 10-15% (30,000-45,000 users) participated in the GhostRank reporting feature. During the month of March these users reported on 393,829 unique domains.

While data from this source cannot comprehensively cover the entire Internet, the potential for a self-selection bias is mitigated by the large sample size and large set of unique domains reported. Furthermore, we primarily sought to examine the use of web beacons on the top 100 sites, and since this data set did cover each of those sites, it was well-suited for our purposes.

The GhostRank data does not include every unique domain the users visited, only those which have web bugs on them that triggered the Ghostery plug-in. The Ghostery plug-in identifies web bugs by matching certain lines of code in the HTML. For a list of the signatures included in the Ghostery source code, see:
http://code.google.com/p/ghostery/source/browse/trunk/firefox/ghostery-statusbar/ghostery/chrome/content/db.js?r=112.

Prior work in web bugs analysis has been done using different methods. A 2007 report used a web crawler, seeded with a selected set of popular websites, to build a dataset of approximately 240,000 web pages, on 24,000 domains [25]. A 2009 study looked at 1200 sites, based on web traffic monitor Alexa's list of most popular sites [26]. This study also combined numbers for different "families" of web trackers and investigated the use of first party cookies by third party trackers, thus getting a better understanding of the breadth of coverage each tracking company has. Both of these approaches have the advantage of including in their data sets websites which have no web bugs, thus gaining insight into the percentage of the entire web a particular bug server covers. Due to the time and budget constraints of this project, we could not afford to develop or use a web crawler to gather information. The GhostRank data was a unique crowd-sourced data set that allowed us to analyze a much larger number of web domains in a very brief time and adds a valuable new perspective to the study of web beacons.

4.2.3 Affiliate Investigation

Most websites state or imply in their privacy policies that they share data with affiliates. However, they do not specify who these affiliates are. We sent messages to each of the top 50 websites and asked for this information. Expecting that most of them would not respond with this information, we also searched for affiliate information in the Mergent Online database to discover parent companies and all the subsidiaries listed under that parent. Based on these lists we got a general sense of the potential number of affiliates each company could share data with according to their policies. This data could only be found for publicly-traded companies. Privately-controlled companies do not have to disclose this kind of corporate information.

5. Findings

5.1 User Expectations/Knowledge

5.1.1 Previous Survey Data

There is overwhelming evidence from various surveys to show that users are concerned about the collection of data by websites. These surveys also show that users desire control of who can collect or see data about them and for what purposes. However, despite these concerns and desires, the studies also show that users are often ignorant of how data collection works, whether it is within the scope of the law, and how to stop it.

USERS CONCERNED WITH COLLECTION AND PROFILING

Each of the studies we looked at showed overwhelming concern by users about the collection of personal information and behavioral profiling. A Consumer Reports poll found that "72 percent are concerned that their online behaviors were being tracked and profiled by companies" and "54 percent are uncomfortable with third parties collecting information about their online behavior" [27].

A Harris Poll found that "a six in ten majority (59%) are not comfortable when websites like Google, Yahoo! and Microsoft (MSN) use information about a person's online activity to tailor advertisements or content based on a person's hobbies or interests" [28]. Supporting data comes from a TRUSTe survey, which found that "57 percent of respondents say they are not comfortable with advertisers using their browsing history to serve relevant ads, even when that information cannot be tied to their names or any other personal information" [29].

Surveys from academic research also show high levels of concern. Papers from the Annenberg Public Policy Center suggest an increase in concern: in 2003, "70% of respondents agreed or agreed strongly with the statement that, 'I am nervous about websites having information about me,'" and "in 2005, the same response was reported by 79% of respondents" [30].

The Pew Internet and American Life Project asked participants the following question: "if an Internet company did track the pages you went to while online, do you think that would be...helpful because the company can provide you with information that matches your interests or harmful because it invades your privacy?" This question is interesting, as tracking could be both helpful and harmful. When asked to choose between the two words the majority of users said tracking was harmful, though a few insisted it was either both or neither: 27% Helpful, 54% Harmful, 11% Both (vol.), 4% Neither (vol.), 4% Don't know/Refused [31].

USERS DESIRE CONTROL OVER USE OF PERSONAL INFORMATION

These surveys also show that users wish to have greater control over how their information is collected and for what purposes it may be used. The Pew Internet & American Life Project asked survey participants about the importance of "controlling who has access to your personal information." 85% responded that it was very important and 9% said it was somewhat important [32].

The Consumer Reports poll found that "93 percent of Americans think internet companies should always ask for permission before using personal information," and "72 percent want the right to opt out when companies track their online behavior" [27].

TRUSTe reported that 68.4% of survey respondents "would use a browser feature that blocks ads, content and tracking code that doesn't originate from the site they're visiting" [29].

USERS LACK KNOWLEDGE ABOUT DATA COLLECTION

Despite concerns about data collection and profiling, the surveys reveal a large level of ignorance on the part of users about how data is collected. The Consumer's Report poll found that "61% are confident that what they do online is private and not shared without their permission," and "57% incorrectly believe that companies must identify themselves and indicate why they are collecting data and whether they intend to share it with other organizations" [27].

In 2003, the Annenberg surveys found that 57% of the survey participants agreed with the false statement "when a website has a privacy policy, I know that the site will not share my information with other websites or companies." Two years later 59% said the same statement was true [30].

5.1.2 Complaints

5.1.2.1 Quantitative Data

USERS CONCERNED ABOUT OPT-OUT CONTROLS

The FTC data contains several statute violation categories within the General Privacy (GP) and Gramm-Leach-Bliley (GLB) statute codes. Since the FTC data is not hierarchical, violations of the Gramm-Leach-Bliley Act may be coded in either the GP or GLB categories. Thus a record may have a GP statute code, but a GLB1 violation code. The only GLB violation code that does not appear in GP is the GLB8 – "Other" code. See the table below for the quantity of complaints, grouped by violation code, that were filed under the General Privacy statute code.

The largest group was the GP5 - "other" violation code, for which we requested free text fields and conducted qualitative analysis (see qualitative data section below). However, the table above shows that, combined, the two categories concerned with opt-out (GLB1 and GLB2) make up a significant portion of the General Privacy complaints. 39% of the total privacy complaints were tagged with one of these codes. The qualitative analysis of GP5 revealed a large portion of the complaints were concerned with control. Thus, users seem to be most concerned with their ability to control the collection and use of information about them.

Violation Code	Violation Description	Number of Complaints
GLB1	Company does not provide any opportunity for consumer to opt out of information sharing	1230
GLB2	Company fails to honor request to opt out/ opt-out mechanism does not work	1678
GLB3	Company is violating its privacy policy	534
GLB4	Privacy policy is misleading, unclear, or difficult to understand	84
GP4	Company does not have adequate security	555
GP5	Other Privacy Violation	3265

Table 2 FTC complaints categorized under General Privacy Statute Code, 2004-2008

CONCERN ABOUT ZABASEARCH AND DATA BROKERS

From the FTC, PRC, and COPP data sets, we found a similar occurrence. A significant portion of the complaints are about data brokers and online sites that act as portals to brokers, such as ZabaSearch, Intelius, or WhitePages. Complaints about ZabaSearch were the most common within all three data sets. ZabaSearch made up 8% of the FTC GP5 complaints, 9% of the PRC complaints, and 18% of the COPP complaints. By comparison, in the FTC GP5 data set, the next three companies were Intelius (2.3%), US Search (1.6%), and Google (1.1%), followed by a long tail of companies that each made up less than 1% of the total.

An analysis of the complaints about ZabaSearch revealed two distinctive spikes in the numbers of complaints during the five-year period (see chart), one in mid 2005 and another in mid 2006. A conversation with the president and co-founder of ZabaSearch, Robert Zakari, revealed that the first spike coincided with a critical article in the San Francisco Chronicle, by David Lazarus [33]. The Privacy Rights Clearinghouse quickly picked up the story and discussed the company in its May 2005 newsletter [34]. A follow-up article [35] in August by the same author at the Chronicle explicitly mentioned the PRC, whose website states that complaints about ZabaSearch are among their most common complaints and which directs users to complain to the FTC.

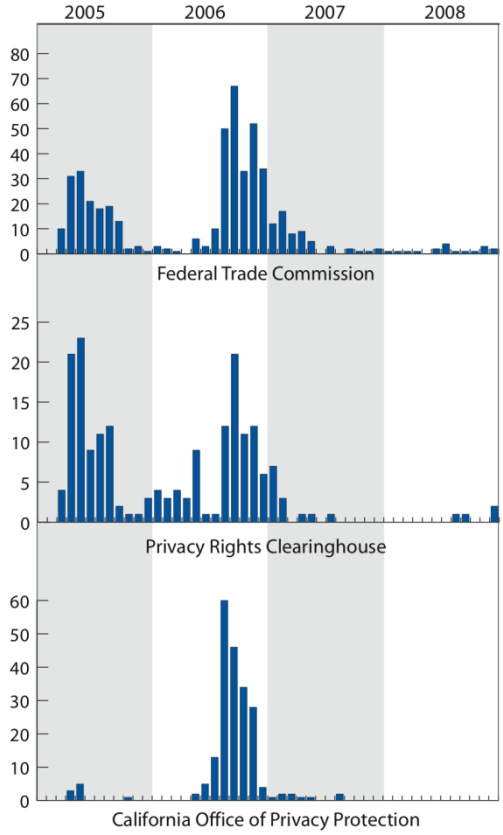


Figure 1 Complaints about ZabaSearch, 2005-2008

Zakari also pointed out that the 2006 spike coincided with ZabaSearch removing their opt-out policy from the website. Additionally, in July 2006, the PRC's monthly newsletter again featured ZabaSearch and specifically directed readers to complain to the FTC. We believe these spikes illustrate that when a specific instance of the public display of a consumer's personal information is made known to them, and they are provided with specific instructions regarding to whom to complain, consumers are concerned and will voice those concerns to advocacy organizations and regulators.

USERS CONCERNED WITH UNAUTHORIZED USE OF PERSONAL INFORMATION

The data from TRUSTe is different from the other three organizations. This is expected as TRUSTe serves a different purpose and only takes complaints about its member websites. Two of the three largest categories, in terms of volume of complaints over the past five years, were related to spam. However, the fastest growing complaint category was about the unauthorized creation of profiles with information about the user already filled in. Complaints in this category increased by 193% from 2007 to 2008.

	2004	2005	2006	2007	2008	Total
Total complaints	3864	7451	7645	6175	6537	**31672**
Privacy-related	1316	1177	970	1428	2150	**7041**
Unauth profile with my information		102	51	206	603	**962**
Email: sent without permission	326	29	119	302	464	**1240**
Shared personal info	251	100	70	277	423	**1121**
Email: unable to unsubscribe	382	408	396	298	288	**1772**
Unable to close account	256	486	261	256	282	**1541**
Unable to contact licensee	20	12	33	78	88	**231**

Table 3 TRUSTe Complaint Data, 2004-2008

5.1.2.2 QUALITATIVE DATA

USERS CONCERNED WITH LACK OF CONTROL AND PUBLIC DISPLAY OF DATA

In our analysis of the free text complaints from the FTC and PRC we found that by far the most common categories of concern involved the public display of personal information and the lack of user control (see the chart below). Fewer than 10% of the complaints had concerns about physical harms that could arise from the distribution of personal information, such as from stalkers. Nearly as many had concerns about the aggregation of data by companies the user had no relationship with, as well as marketing (spam), and security.

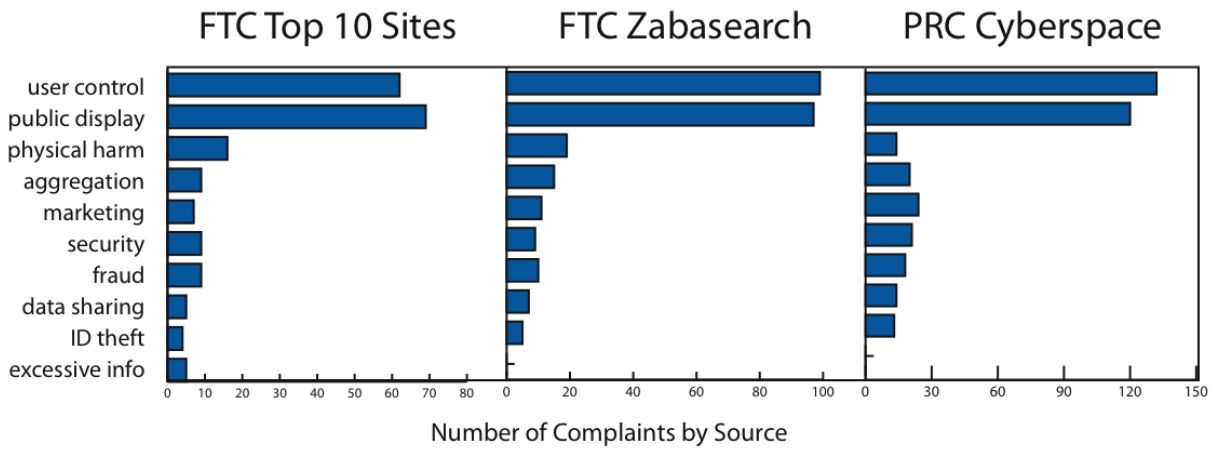


Figure 2 Free text complaints coding, 2004-2008, from three samples:
1) **Random sample of complaints made to FTC about websites in the top 10**
2) **Random sample of complaints made to FTC about data broker websites/portals**
3) **Random sample of complaints made to PRC about any website**

5.1.2.3 PROBLEMS WITH THE FTC DATA

The process of acquiring a data set of user complaints regarding Internet privacy from the FTC proved challenging.

CATEGORIES UNKNOWN

The FTC accepts two types of complaints submissions: those provided by phone, and those provided by web form. Using the web form submission interface we tried to estimate the category most closely encompassing complaints which might be related to web privacy and made a FOIA request for complaints categorized as "invasion of privacy."

After an initial review of this first data set, we found that the user interface categories may not directly map to corresponding database fields, and that a valid data request should be made using the database categories instead. From the datasets we received, we determined that each record contained a 'Product Service Code', 'Statue Code', and 'Violation Code'. To determine how each of these codes were related we submitted another FOIA request for a detailed list of the database categories. From this category index we were able to determine that the highest level category was 'Product Service Code', which contained 'Statute Codes', which in turn contained 'Violation Codes.' Product Service Codes describe the type of industry involved in the complaint, Statutes describe the general domain of the complaint, and Violations describe the specific complaint concern.

DATA IS NEITHER HIERARCHICAL NOR STABLE

However these categories are not structured hierarchically. Each category may be queried independently, thus while Violation Codes fall within Statute Codes, a specific Violation Code may be common to several Statute Codes, and likewise Statute Codes to Product Service Codes. For the datasets we received, for example, the complaints were marked with a Violation Code for 'other privacy violation (GP5)', a Statute Code for 'General Privacy (GP),' and one of a handful of Internet-related Product Service Codes (such as 'Internet Access Services' or 'Internet Information & Adult Services'). Each record may be coded in multiple categories, which makes cross category comparisons difficult.

We also discovered that the data fields change. Sometimes a category is dropped from the data structure and merged with another. Hence, there are no longer any GP1, GP2, or GP3 Violation Codes; they were rolled into the GLB1 and GLB2 Codes. This makes longitudinal studies difficult.

Based on our research, we were able to create a dynamic treemap of the FTC's category system using the ManyEyes software. You can view the interactive map here: http://manyeyes.alphaworks.ibm.com/manyeyes/visualizations/ftc2. Take note that the sizes of the boxes in this visualization do not represent the quantity of complaints filed. Each color coded box represents a statute code. The medium sized boxes inside the Statute Codes represent the various Violation Codes that are paired with that statute. Inside each Violation Code box are numerous Product Service Codes (the small boxes in white outlines) that are found together with that violation. Thus the sizes of the Statute Code boxes are determined by the number of product service codes related to it.

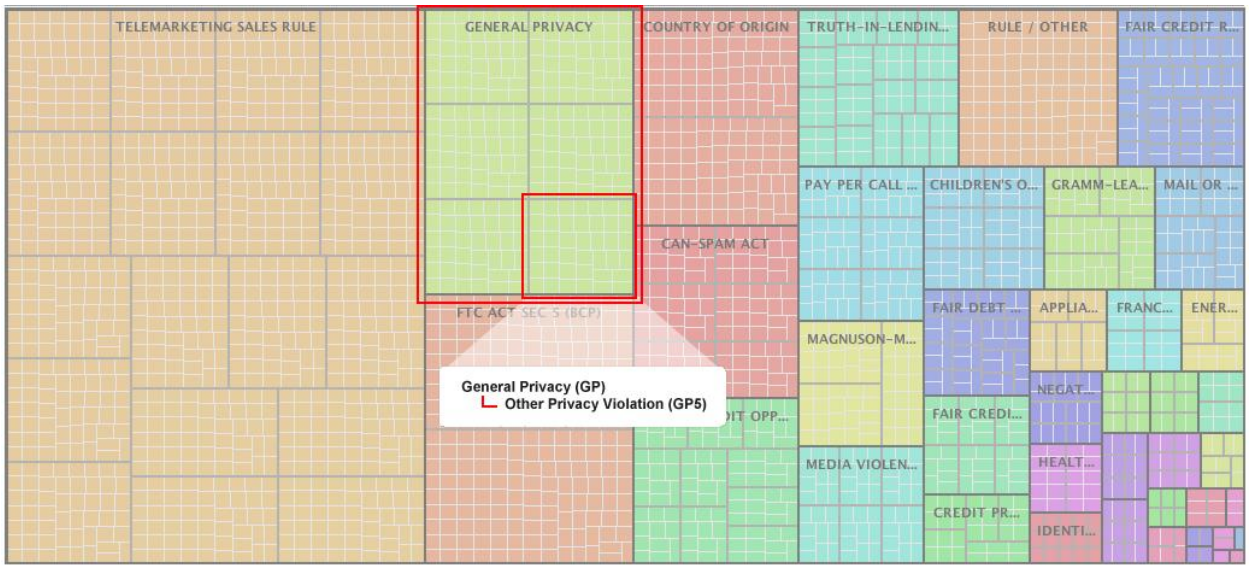


Figure 3 FTC categories. Statute Code->Violation Code->Product Service Code

USER INTERFACE COULD INTRODUCE BIAS

In order to get to the "other privacy" category which we looked at, the user must click through several drop-down menus. An uninformed user could follow an incorrect path through the menu system and file a complaint in the wrong category. For instance, the first menu asks the user if the complaint is concerned with identity theft. In our analysis of the free text complaints we found that many users who are concerned about services like ZabaSearch are often worried they could be victims of identity theft. If they went to the FTC's complaint form they may just click

on the first prompt for identity theft, even though that may not be what the actual situation is about. For screenshots of the path towards complaints about Internet privacy, see Appendix D. From 2004-2008, the FTC received 1,315,179 complaints in the category for ID Theft, as opposed to 6,713 for General Privacy and 42,765 for CAN-SPAM.[*] While much of this could be due to the proliferation of these types of concerns as well as the FTC's clear ownership of the Identity Theft domain, there could also be bias introduced simply on account of the method in which these complaints are submitted. We hope to conduct future research into the free text complaints in the ID Theft category to determine if those complaints are correctly categorized.

5.1.3 News Stories

Our analysis of news stories for the past two years found a fairly even distribution of coverage for most of the topics covered in this report. Slightly more coverage was given to topics related to behavioral profiling.

The one category our textual analysis found significantly lacking was the issue of websites sharing data with affiliates. To verify this deficiency, we ran another specific query for this topic. For the same three news sources for the five year span from 2004 to 2008, we searched for all articles containing 'internet' and 'privacy' in the same paragraph that also contained the words 'subsidiar*' or 'affiliate*' anywhere in the article. Of the same 1,778 'internet privacy' stories, we found 84 hits, though most of these were false positives as well. After we culled all the irrelevant uses of the words, we found only 9 valid hits that mentioned websites sharing data with affiliates. Thus, we found that although these three newspapers mention data sharing, they miss an important element of which users may not be aware.

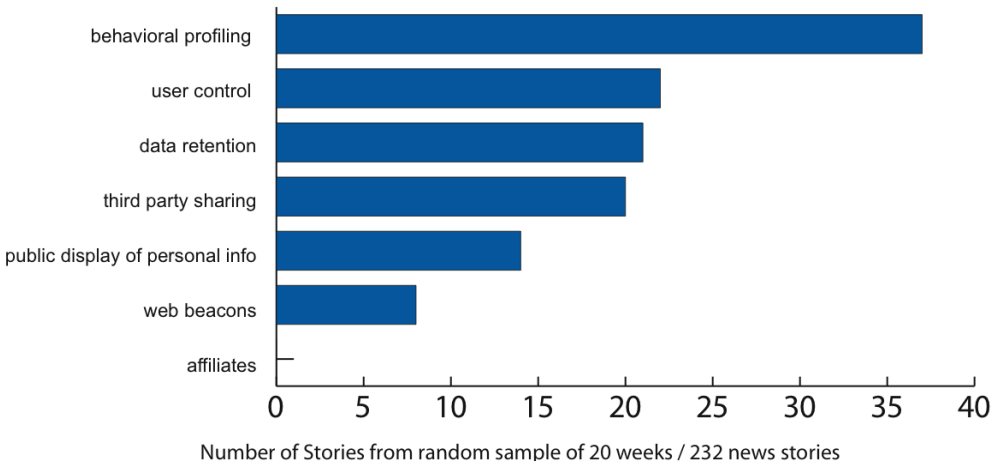


Figure 4 News story coding, 2007-2008, from random sample of 20 stories per year, from three major newspapers: New York Times, Washington Post, San Jose Mercury News

The codes we used for categorizing news stories bundled all both first- and third-party tracking technologies together. To get a better understanding of how much exposure third-party tracking was receiving we did a specific query among the same three newspapers for the date range of January 1, 2004 through December 31, 2008, for all articles that contained 'internet' and 'privacy' in the same paragraph and also contained the terms 'bug' or 'beacon' somewhere in the

[*] These numbers were tallied from FTC Public Affairs announcements, not from our requested data sets. See the following URLs: http://www.ftc.gov/opa/2009/02/2008cmpts.shtm; http://www.ftc.gov/opa/2008/02/fraud.shtm; http://www2.ftc.gov/opa/2007/02/topcomplaints.shtm; http://www.ftc.gov/opa/2006/01/topten.shtm; http://www.ftc.gov/opa/2005/02/top102005.shtm

article. This query returned a few dozen hits, though most were false positives for topics such as software glitches or Facebook's Beacon initiative. Once these were culled from the list, we found only six valid hits among 1,778 total 'internet privacy' stories. Thus, while issues of tracking and profiling were mentioned in these three papers, little was said about the actual technology that enables it.

5.2 Website Practices

5.2.1 Policy Analysis

We analyzed the privacy policies of the top 50 most visited websites (according to Quantcast as of March 1, 2009).

5.2.1.1 Data Types

From the website privacy policies we see that the top 50 websites collect a significant amount of information about users. All 50 collect computer information such as IP address or type of operating system. This is expected, as this type of information is automatically collected by most server logs and useful in investigating security breaches or attacks. However, 49 of the top 50 also collect some form of contact information, such as name, address, or phone number (the only exception is Wikipedia, for whom contact information is optional for site registration – users can make edits pseudo-anonymously using only an IP address as an identifier). The majority of the top 50 websites also collect demographic, financial and interactive (click stream) data. Only a few affirmatively stated that they collect content information (such as communications, media files, etc.), though none of them affirmatively stated that they did not. Most of the policies were unclear about it, or simply did not mention it.

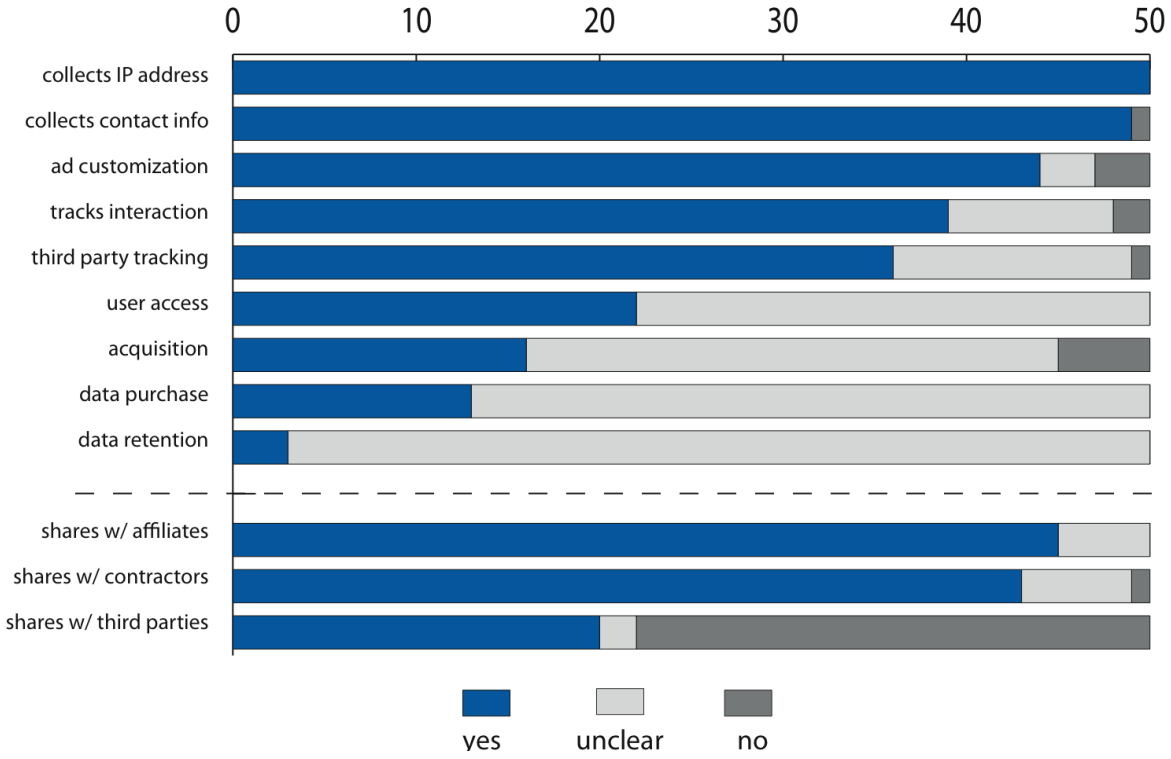


Figure 5 Privacy Policy Coding, for top 50 most visited websites

5.2.1.2 Sharing

Websites make distinctions between sharing with affiliates, contractors, and third parties. Of the top 50 sites, 29 stated that they do NOT share user data with unrelated third parties. However, 45 affirmatively state that they share data with affiliates, and 36 affirmatively state that they allow third-party tracking. The average consumer might assume an affiliate or tracker to be a third party, but given the actual usage of these terms in privacy policies, that assumption would be mistaken.

Of the top 50 sites, 43 state affirmatively that they share data with third-party contractors, including all 29 of the sites who state that they do not share with unrelated parties. Although consumers may consider these entities to be third parties as well, this form of sharing may not be as problematic. Most of these contractors are necessary (for instance, e-commerce sites must share consumer contact information with shipping agencies), and they are often contractually limited in their use of consumer data.

5.2.1.3 Purpose

We coded the purposes for which the policies stated the data collection was to be used. After review of our data we reduced our set of purpose codes down to three essential purposes: public display, ad customization, and third-party sharing. Ad customization includes both contextual, one-time customization (such as when a search query is used to dynamically generate ads) as well as longitudinal behavioral profiling, where your activity across multiple visits is used to build a behavioral profile about you.

The websites are almost evenly split on the publication of user data for public display. Many of the top 50 sites incorporate some kind of social networking functionality, so this number is not surprising. However, an overwhelming majority (44 of top 50) stated that information collected about users will be used for purposes of customizing advertisements.

Thirty-six of the sites stated that they allow trackers, while fourteen of them were unclear. Some of them contained what some users could perceive to be contradictory statements. For instance, Microsoft's policy stated that "*Microsoft may also employ Web beacons from third parties in order to help us compile aggregated statistics regarding the effectiveness of our promotional campaigns or other operations of our sites. We prohibit Web beacons on our sites from being used by third parties to collect or access your personal information*" [36]. The policy suggests that Microsoft employs third-party trackers, but prohibits them from collecting personally identifiable information. However, in their current form, these two sentences may sound contradictory to an average user. Furthermore, the statement is problematic because a web beacon automatically collects information about a user, such as IP address, which can be used to determine other information, such as geographic location. Beacons also enable trackers to identify the content a user chooses to read and view, which is arguably very personal information. Microsoft may not deliver registration information to the tracker that would allow it to personally identify the user, but if the beacon comes from a tracker that already has a cookie installed on the user's machine, then the user may already be identified.

5.2.1.3 Other Findings

Although access/participation is a core FTC FIP, more than half of the policies were unclear as to whether users can access, edit, or delete their personal information.

Many web companies are started by entrepreneurs who ultimately wish to sell their business. However, most privacy policies were unclear about the fate of user data in the event of acquisition or bankruptcy.

While many websites emphasize the idea that trust drives consumers to reveal personal information to websites, voluntary sharing is only one way sites obtain personal information. Twelve of the 50 sites affirmatively state that they also purchase data from third parties to supplement or enhance their data.

How long a company keeps personal information about its users is a topic of increasing public importance. However, of the top 50 sites, 47 have a retention policy which is unstated or unclear.

5.2.1.4 Responses From Companies

We received responses from seven companies, representing 12 of our top 50 websites. Most of them stated that our interpretation of their policy was generally correct (Adobe gave us complete approval), though they also pointed out that some of our findings were dependent on context. Many were concerned with our use of the "unclear" tag. For instance, Microsoft's Director of Privacy Strategy wrote to us, stating:

"Privacy policies are usually more nuanced than such categorized analysis allows for. For example, it is indicated that we do not provide data to third parties. This is most often the case, but there is a case where, with the opt-in consent of the customer, we do provide data such as an email address to third parties for marketing purposes. I can think of a number of other examples of where the yes/no analysis results in both a conditional 'yes' and a conditional 'no.' Therefore, I worry that the conclusions, if published as they are, will be misleading."

This response raises difficult problems for the notice and choice regime favored by businesses and the FTC. This regime is predicated on user choice, informed by privacy policies. If there are nuanced situations that create conditional yes or no answers to these basic questions about a site's data collection and sharing practices, then it is unclear how an average user could ever understand these practices if the nuances are not explained in the privacy policy. Choice, therefore, cannot be informed.

In this context, for purposes of this report, yes/no dichotomies are still tenable, because we are exploring whether information is shared with third parties without opt-in consent.

5.2.2 Web Bugs Data

The data from Ghostery identified 117 unique web bug servers on 393,829 unique domains visited during the month of March by approximately 30,000-45,000 users.

USERS ARE TRACKED BY DOZENS OF COMPANIES

Many websites featured multiple web bugs; some had several dozen. The two sites with the most web bugs were both blogging sites: Blogspot [*] had 100 and Typepad had 75 (Blogger came in fourth with 31). This does not mean that these sites had this many web bugs on them at once. Rather, it means that during the month of March a number of unique web bugs were reported by various users during different visits to a particular website (a website delivers different web bugs at

Domain	Web Bugs
blogspot.com	100
typepad.com	75
google.com	44
blogger.com	31
msn.com	29
aol.com	28
yahoo.com	27
huffingtonpost.com	27
photobucket.com	25
tripod.com	25

Table 4 Sites with most web bugs, March 2009

[*] Note that Blogspot is a part of Blogger, a subsidiary of Google. Individual blogs are still served from the blogspot domain, but traffic to the blogspot.com main page is directed to blogger.com.

different times and to different users). The prevalence of web bugs on blogging sites are likely the result of individual bloggers' use of third-party trackers on their blogs, rather than the actions of the site operators. However, this number seems larger than what a typical user might expect upon visits to a website. See Appendix E for a large chart of the top 100 websites' web bugs.

TRACKING COMPANIES HAVE EXTENSIVE COVERAGE

In addition to the abundance of web bugs on individual sites, the data shows that tracking companies have the potential to cover vast swaths of the Internet. The biggest players showed up on hundreds of thousands of unique domains.

In the tables below, it is apparent that Google is the dominant player in the tracking market; it operates the top three trackers and four of the top 10. We found five trackers overall operated by Google, including Analytics, DoubleClick, AdSense, FriendConnect, and Widgets. Among the top 100 websites this project focused on, Google Analytics appeared on 81 of them. When combined with the other trackers it operates, Google can track 47 of the top 50 websites, and 92 of the top 100 websites. Further, a Google-operated tracker appeared on 348,059 of 393,829 distinct domains tracked by Ghostery in March 2009, i.e. over 88% of the domains tracked by Ghostery that month. This trend appears to be consistent with more recent data as well. Preliminary analysis suggests that the Google trackers cover more than 80% of approximately 766,000 unique domains reported through April 2009. Please note that the data set does not include websites that have no web bugs at all. Thus, our data does not claim that Google can track across 88.4% of the Internet. Rather, it means that of the domains that use at least one form of third party tracking we found that Google is used by 88.4% of them.

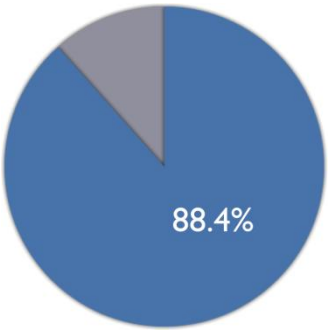


Figure 6 Combined coverage of Google trackers, March 2009. 348,059 out of 393,829 unique domains reported to use some form of third party tracking.

We are not claiming that Google aggregates information from each of these trackers into a central database, though it does possess the capability to do so. It appears that they strive to keep data in silos. For instance, their Analytics FAQ indicates that they give website operators control over how Google may use the data. Operators can keep data gathered by Google Analytics from being used by Google for other services [37]. However, Google creates incentives for site operators to share by offering premium services only to those websites that share data.

Tracker	Number of unique domains found on	Percent of all unique domains
Google Analytics	329,330	84%
Google Adsense	162,584	41%
DoubleClick	122,483	31%
Statcounter	26,806	7%
AddThis	24,126	6%
Quantcast	24,113	6%
Google Custom Search Engine	20,601	5%
OpenAds	17,608	4%
Omniture	13,126	3%
Wordpress Stats	11,475	3%

Tracker	Percent of Top 100 found on
Google Analytics	81%
DoubleClick	70%
Microsoft Atlas	60%
Omniture	57%
Quantcast	57%
PointRoll	54%
Google Adsense	52%
Dynamic Logic	48%
Insight Express	41%
ValueClick Mediaplex	41%

Tables 5 & 6 Percentage of Domains each Web bug was found on, March 2009

NO ACCOUNTABILITY FOR THIRD-PARTY TRACKING

In our analysis of privacy policies, 36 of the websites affirmatively acknowledged the presence of third-party tracking. However, each of these policies also stated that the data collection

practices of these third parties were outside the coverage of the privacy policy. This appears to be a critical loophole in privacy protection.

LIMITATIONS

We would like to point out a few constraints that limit the conclusions that can be made from the GhostRank data. One constraint is the grouping of all web pages under a single domain. Thus if a single web page in a given domain has a web bug, then the entire domain is marked as using a web bug, even though in some cases every other page within the domain does not employ web bugs. Furthermore, the prevalence of web bugs on the blogging sites illustrates the different sources that web bugs may have. Some web bugs could be placed by users rather than the site operators, though from a user's perspective, the presence of a web bugs is of greater importance than the responsible party. Both of these constraints could have the effect of overestimating the use of web bugs by site operators.

However, other constraints in the data could result in underestimating the use of web bugs. For instance, the GhostRank data set does not include third-party tracking performed through first party cookies and JavaScript or other methods such as DNS aliasing, which has been discussed in other research [26]. Additionally, it does not include tracking that employs other technologies, such as Flash or Silverlight.

5.2.3 Affiliate Investigation

In our analysis of the privacy policies we found that 46 of the top 50 companies affirmatively state that they share data with affiliates, and the four remaining were unclear. We sent each company a request via email or an online web form for a list of each affiliate they may share data with. We received 14 replies, but none included the lists we asked for. Most stated that they do not disclose corporate information. Some companies did offer a little information. For instance, eBay mentioned that some of their more notable subsidiaries include PayPal, Half.com and Rent.com. Based on our experience, it appears that users have no practical way of knowing with whom their data will be shared.

Our search for corporate families in the Mergent Online database turned up some surprising information. Many of these websites are owned by parent companies that have hundreds of subsidiaries. MySpace, one of the most popular social networking sites (especially among younger users), is owned by NewsCorp, which has over 1500 subsidiaries. Bank of America has over 2300. It should be noted that these numbers include several foreign affiliates. For instance, Google has 137 subsidiaries, including Google Canada, Google Belgium, Google Israel, and other foreign offices. However, the numbers we compiled do not include subsidiaries of subsidiaries, so there may be more that are missing. The numbers at least give us an idea of the vast corporate families to which many of these websites belong. Information pulled from these websites could potentially find its way to all of these affiliated companies.

Privacy law has typically treated third-party information sharing differently than affiliate sharing. Third-party information sharing is often subject to more restrictions, including opt-in or opt-out consent requirements. These restrictions are based upon the heightened risk associated with sharing information with unrelated entities, which may have different incentives than the company that collected the information. The law on affiliate sharing generally is more permissive. Incentives for security and fair treatment of data are assumed to exist among affiliates. However, given the large size of affiliate networks, the fact that many affiliates are essentially unrelated entities with different business models in entirely different fields, and the

practical challenge of identifying their size and scope, the more liberal treatment of affiliate sharing should be reexamined.

6. Discussion

6.1 User Concerns, Complaints and Knowledge

Our review of survey data suggests that users are very concerned about privacy and do not want websites to collect and share their personal information without permission. Yet, the number of complaints made to the various organizations is low relative to the number of Internet users. The FTC had only 6,713 for five years (in the General Privacy category), the PRC had 2,202 for the same period and the COPP had 1,152. TRUSTe had 7,041 that it categorized as privacy related.

Website operators and direct marketing agencies might point to this low volume and claim that users don't care about data collection. However, that would be a misinterpretation of the data. It is apparent from our research that users do care. The low number of complaints simply conforms to our hypothesis that users file complaints only when two conditions are met: 1) they perceive an invasion of their privacy, and 2) they know where to file a complaint.

The largest numbers of complaints at all four of the institutions we received data from were concerned with public displays of personal information. In the case of the FTC, PRC, and COPP these complaints were about online data brokers and portals like ZabaSearch. For TRUSTe, which does not represent those companies, the fastest growing complaint category is for the unauthorized creation of a profile with personal information. These are special cases where users were able to see their data being collected and distributed without their permission and felt that their privacy was violated.

However, most users are unaware of the majority of the data collection and sharing that goes on. Consumers may have heard that websites can track their behavior, but the tracking is done passively, and is therefore not salient in the minds of the users. Furthermore, several of our data sources show that users may not actually be aware of how data collection works and do not fully understand the policies that govern it. The analysis of news coverage showed a dearth of coverage about web bugs and affiliate sharing, and the surveys indicate that users do not read the privacy policies and misunderstand what they say. Our analysis of the actual policies shows that they are often vague or misleading.

In the case of ZabaSearch, some users were made aware of the FTC's complaint form via the Privacy Rights Clearinghouse website, which was referenced by the media. Thus, they had knowledge of where to file a complaint. Members of TRUSTe are required to link to TRUSTe from their privacy policy pages. This may be why TRUSTe receives more complaints than the other organizations, despite only serving a tiny fraction of the entire Internet.

In all likelihood there is no particular agency to which users are most likely to express privacy concerns. More often than not users probably direct their complaints to the specific entity with which they have a concern, particularly when the user has a direct relationship with that entity. That users care about their privacy and often complain directly to the website involved is supported by evidence from the incident involving Facebook's Beacon initiative in November 2007. Many users were upset with a commonplace practice, one that was made salient and obvious by the Facebook Beacon system. That system enabled e-commerce sites to share data about their transactions with Facebook, which in turn posted the data on its users' public news feeds. In this case, users were made aware of a use of private information that they had not authorized. Furthermore, users could voice their objections to the practice by joining a

protest group on the site itself. Over 50,000 users joined the group in ten days (over 80,000 in one month) strongly suggesting that users do care about privacy [38].

The similarities between the practices of online data broker portals like ZabaSearch and online behavioral advertising are interesting. Both are conducted by entities that a person has no relationship with. Both involve the collection of information about a person from various sources, which can be bought and analyzed by other entities that do have relationships with the person. Both practices are essentially invisible. The striking difference between the two is that ZabaSearch displays the end result publicly. When people were made aware of ZabaSearch's practices (and a proper forum for complaints), they complained. This raises the question of whether or not users would complain about behavioral profiling if they could see the end results and knew where to file a complaint. One might argue that it is not the collection of information by ZabaSearch that is of concern to users, but rather it is the public display of that information. This argument would deem behavioral advertising of no concern as the information is never distributed publicly. However, the data from multiple previous surveys that we looked at all point to user concerns over websites collecting information about them and using it to deliver targeted ads. Public display is only part of the reason people complained about ZabaSearch. What was of ultimate concern was control. Users want control over who can collect, share, and use information about them.

6.2 Control

The FTC has placed a particular importance on privacy "harms." However, complaints to that agency we analyzed showed an overwhelming concern about lack of control. Users want the ability to edit and delete information about them as well as to determine who can have access to certain types of information. Consumer complaints demonstrated great discomfort with the ability of data broker portals to sell data to anyone, meaning that in reality, no one is in control of the data. While the FTC has framed online privacy issues in terms of "harm," consumers' complaints focus on lack of control over personal information.

Our review of the policies showed that only 23 of the top 50 affirmatively stated that users could have access to some portion of the information the website had collected about them. The remaining 27 policies lacked mention of access or their statements about access were unclear.

However, none of the policies specified that a user could access *all* the data that had been gathered. Instead, most of them allow users to edit information the user had offered through registration forms, communications, file uploads, etc. None of them explicitly offered users the ability to view or delete click stream data. Therefore, claims of user access are only partially true. Furthermore, users have no ability to discover which data were shared with affiliates.

Self-regulation is based on the premise that if users do not like a website's practices, they can simply avoid the website. Giving users access to data about them and enabling participation in the data collection process are methods by which site operators can make their practices more appealing and prevent users from going elsewhere. However, third-party trackers are not governed by a website's privacy policy. Therefore, they have no incentive to allow users to view or delete information collected about them. In addition to this lack of participation, users have no ability to avoid third-party tracking. There is no opt-out, let alone opt-in.

The Network Advertising Initiative (NAI), a "cooperative of online marketing and analytics companies" [39], currently has an opt-out mechanism that requires users to download a cookie, which will let direct advertisers know not to install any third-party tracking cookies on the user's computer. This method of opt-out is unacceptable. First, it only governs members of the NAI; tracking companies that are not members will still be able to use cookies and web bugs to collect

data about users. Second, users that decide to delete cookies on their machine may delete the NAI cookie inadvertently and open up their machine to third-party tracking again.

Users cannot avoid trackers by avoiding websites that use them; our data shows that trackers are ubiquitous on the web. Many browsers give the user the option to block third-party cookies, but this does not block JavaScript web bugs. Browser technology could create a system by which users could block content coming from a server other than the one serving the web page. However, that would also block a lot of desired content, such as embedded videos, or framed websites that result from a Google image search, and would totally disrupt web advertising norms. This is a case of market failure, as users have no options to protect their privacy.

Furthermore, the argument that users should simply avoid certain websites is unrealistic. More and more of our social and political discourse is taking place on these popular websites. Colleges have begun communicating to students via Web 2.0 sites like Facebook [40]. The Obama administration has also begun engaging the public via social networks as well as media websites like YouTube and Flickr [41].

6.3 Deceptive Practices

Our analysis of privacy policies found that most companies do not share personal information with third parties, where "third parties" is defined to exclude contractors and affiliates. Many may share data in an aggregated form, but do not divulge identifiable information. This seems to conform to users' concerns over the unauthorized sharing of personal information. Whether this practice is due to the concerns of users remains to be seen. It may just be good business sense to withhold the valuable information and act as an intermediary between the users and direct advertisers. However, companies are not as protective of private information as users would like them to be. Data is still flowing to other entities through affiliate networks or via third-party tacking bugs.

Most users do not know the corporate families to which these websites belong. How many users know the Internet Movie Database is owned by Amazon? How many know that MySpace is connected to Fox News via its parent company NewsCorp? For some users, this may be common knowledge; other users may be completely oblivious to these facts. Furthermore, we have found that it is difficult for a user to discover exactly who these affiliates are, even if they took the time to ask.

Website operators should reevaluate a common practice we discovered: claiming that they do not share information with third parties, but allow third-party trackers. We think that these statements are inherently contradictory. A practice is deceptive for purposes of the Federal Trade Commission Act if it involves a "material representation, omission or practice that is likely to mislead a consumer acting reasonably in the circumstances, to the consumer's detriment" [42]. The conflicting statements in the privacy policies would most likely confuse or mislead a reasonable consumer. The confusion would also likely be to their detriment, as surveys indicate that users do not want companies to collect data about them without permission. Deception is a legal term, and we do not claim that these practices necessarily meet the standard. However, to the extent that website operators wish to avoid stricter regulations, they should pay more attention to practices that may even appear to be deceptive.

7. Conclusions/Recommendations

Based on our findings we offer the following recommendations:

7.1 Access, Control, and Salience

The biggest concern among the complaints we coded was the lack of control. Users do not want websites to collect or share data without permission, and they want the ability to access, edit, and delete records about themselves. In 2003, Joseph Turow found that 94% of his sample of 1,200 American adults agreed or agreed strongly with the statement, "I should have a legal right to know everything that a website knows about me" [43].

We recommend regulation by which both websites and third-party trackers must allow users to see *all* the data that has been collected about them, not just user-provided information. Additionally, users should also be allowed to see with whom their data has been shared. The imposition posed upon companies by such a requirement could be greatly mitigated by merely requiring that websites provide users with the information they have about the user in a form no less convenient than the form in which it is available to the company.

We recommend that companies request permission from users before sharing data about them with any outside party, regardless of affiliation.

The presence and purpose of third-party tracking should also be made more salient in the minds of users. We recommend that all browser developers provide a Ghostery-like function in their browsers that alerts users to the presence of third-party trackers.

7.2 Authority & Metrics

Our analysis of user complaints brings to the fore a larger problem with data collection policy in the United States: no one knows who is in charge of protecting privacy. The fairly low number of complaints to the various organizations we contacted reveals that users do not know to whom they should complain. Furthermore, the FTC's new principles for behavioral tracking make no mention of any enforcement or accountability principles.

According to the FTC's Privacy Initiative web page, it safeguards consumer privacy by enforcing the Gramm-Leach-Bliley Act, the Fair Credit Reporting Act, and the Children's Online Privacy Protection Act. It also states that the FTC strives to educate "consumers and businesses about the importance of personal information privacy" [5]. We recommend that the FTC become more aggressive in protecting privacy on the Internet.

The first step for the FTC is to improve the integrity of its current system for taking user complaints. We recommend an overhaul of both the user interface as well as the database architecture. The current system may introduce bias in its presentation of complaint categories, especially for users who may lack the technical understanding to accurately describe their exact concerns, thereby affecting the data collected. If the FTC is going to protect privacy it must be able to gauge public sentiment and measure the efficacy of its policies in an accurate manner.

It should also strive to get a larger picture of user concerns. Therefore, we recommend that the FTC make more users aware of the complaint assistance system. One possible way to achieve this is to require websites that collect personal information about users (other than the automated IP logs) to include a link on their privacy policies to the FTC's website. This would direct users to the FTC and help it gain insight into user concerns.

7.3 Better Notice

Notice is the FTC's primary Fair Information Principle. Users must be made aware of data collection practices if they are to make informed decisions. In the Introduction we discussed

several reasons why privacy policies are an ineffective means of notifying users of practices. However, to the extent that they remain the primary method of notice, we have some suggestions for improvement.

First, the policies should be readable for average users. Despite years of research showing problems with the language of privacy policies, they are still difficult to read. We conducted a Flesch-Kinkaid readability test on the 50 policies we analyzed and found that the average grade level was 13.83 (the lowest was Chase with 8.66, and the highest was Adobe with 17.29, standard deviation was 1.89).

Beyond the problems with language, the policies are often vague about actual practices, and contain statements that are contradictory or misleading. Many state that data is not shared with third parties even though the data may be shared with affiliates with whom the user has no relationship. Allowing third-party tracking while claiming that data is not shared with third parties is also misleading. By sharing space on a web page for tracking companies to collect information, website operators are in effect sharing user information with third parties.

We recommend that users be given clear and proper notice as to whom the data will be passed, regardless of affiliation or method of sharing. The policies should not contain conflicting statements that third-party sharing is not allowed but third-party tracking and affiliate sharing are. Therefore, we recommend the FTC adopt strict definitions for the terms "affiliate" and "third party." In addition, users should be informed as to whether or not the flow of data will stop with the affiliate or if the affiliate may share data with another company.

We also recommend that the practice of third-party tracking be made more transparent. It currently operates in a policy loophole, by which neither the website nor the tracker are clearly accountable for the data collected. We recommend that websites define the policies of the third-party trackers it allows on its site or, at a minimum, link to the appropriate policies on the tracking companies' websites and specify which practices fall under each policy.

We also recommend that the FTC create an opt-in standard for enhancement, the practice of buying information about users from outside sources. The FTC's self-regulatory regime is premised on the idea that consumers will selectively disclose personal information to websites they trust. Enhancement circumvents this process, and allows websites to obtain this same information without user participation. A user who decides to reveal a small amount of personal information to a website that she does not fully trust loses all defenses when that site can simply bump up the submitted data with extrinsic, enhanced data.

8. Acknowledgements

We would like to thank several people for their guidance and support with this project. Foremost, we are greatly indebted to our faculty advisor, Brian Carver, for offering a large portion of his time to our project, continuously providing us with sound advice, and always engaging us with enthusiastic support. We are also much obliged to Chris Hoofnagle for providing us with his insights from years of experience working in this field of research, as well as helping us with numerous FOIA requests to the FTC. Many thanks are also due to Eric Kansa for procuring funds to help us hire our undergraduate assistant.

We would also like to acknowledge the help we received from the project's assistant members. Our undergraduate assistant, Sona Makker, did an excellent job helping us with our research, and Mark McCans provided us with helpful legal advice.

This project would not have been possible without access to data from several different organizations. We owe our sincerest thanks to Beth Givens at the Privacy Rights Clearinghouse,

David Cancel from Ghostery, the staff at the Federal Trade Commission, Joan McNabb at the California Office of Privacy Protection, and Simona Nass at TRUSTe.

We have received a wealth of advice from various privacy and legal scholars throughout our work on this project. Many thanks go to the following people for providing us with their insights: Deirdre Mulligan, Ryan Calo, Robert Gellman, Aleecia McDonald, Lori Cranor, Jim Dempsey, Jason Schultz, and Alessandro Acquisti.

Thank you to the administrators and judges of the Bears Breaking Boundaries contest, specifically the Science, Technology & Engineering Policy Group.

Thank you to the I-School IT staff for their support and hosting of our project website.

Thank you to Reid Oda for helping us retrieve data.

Thank you to our fellow I-School student, Nick Rabinowitz, for use of his textual analysis software.

Thank you to Robert Zakari of ZabaSearch for his time and forthright insights.

9. References

[1] Leibowitz, Jon, "Concurring Statement," *FTC Staff Report: Self-Regulatory Principles for Online Behavioral Advertising*, 2009.

[2] FTC (Federal Trade Commission), "Self-Regulatory Principles For Online Behavioral Advertising" 2009. http://www.ftc.gov/os/2009/02/P085400behavadreport.pdf

[3] Schatz, Amy, "Lawmakers Examine Privacy Practices at Cable, Web Firms," *Wall Street Journal*, April 23, 2009. http://online.wsj.com/article/SB124050539070948681.html

[4] Boucher, Rick, "Communications Subcommittee Holds Hearing on Network-Based Technologies and Privacy," Online Office of Congressman Rick Boucher, April 23, 2009. http://www.boucher.house.gov/index.php?option=com_content&task=view&id=1654&Itemid

[5] FTC (Federal Trade Commission), "Fair Information Practice Principles." http://www.ftc.gov/reports/privacy3/fairinfo.shtm

[6] Gellman, Robert, "Fair Information Practices: A Basic History," 2008. http://bobgellman.com/rg-docs/rg-FIPshistory.pdf.

[7] Walenta, Toasz, "Do Consumers Understand the Role of Privacy Seals in E-Commerce?" *Communications of the ACM.*, vol. 48, no.3, 2005. http://portal.acm.org/citation.cfm?id=1047674

[8] Edelman, Ben, "Certifications and Site Trustworthiness," 2006. http://www.benedelman.org/news/092506-1.html

[9] Anderson, Ross; Moore, Tyler, "Information Security Economics and Beyond," Information Security Summit, 2008.

[10] California State Legislature, *California Business and Professions Code*, vols. 22575-22579. http://media.gibsondunn.com/fstore/documents/pubs/BP22575-22579.pdf

[11] TRUSTe, "Consumers Have False Sense of Security About Online Privacy – Actions Inconsistent With Attitudes," 2006. http://www.truste.org/about/press_release/12_06_06.php

[12] Anton, Annie, "The Lack of Clarity in Financial Privacy Policies and the Need for Standardization," *IEEE Security & Privacy*, vol. 2, no. 2, 2004. http://www.truststc.org/wise/articles2009/article4.pdf

[13] Hoofnagle, Chris; King, Jennifer, "What Californians Understand About Privacy Online," Samuelson Law, Technology & Public Policy Clinic, 2008. http://www.law.berkeley.edu/clinics/samuelsonclinic/files/online_report_final.pdf

[14] McDonald, Aleecia; Cranor, Lorrie Faith, "The Cost of Reading Privacy Policies," CyLab, Carnegie Mellon University, 2008.

[15] Cranor, Lorrie Faith, et al., "2006 Privacy Policy Trends Report," CyLab Privacy Interest Group, 2007.

[16] Acquisti, Alessandro, "Privacy in Electronic Commerce and the Economics of Immediate Gratification," 2004. http://www.heinz.cmu.edu/~acquisti/papers/privacy-gratification.pdf

[17] Acquisti, Alessandro; Grossklags, Jens, "Privacy and Rationality," *Privacy and Technologies of Identity*, 2006. http://www.dtc.umn.edu/weis2004/acquisti.pdf

[18] Acquisti, Alessandro; Grossklags, Jens, "What Can Behavioral Economics Teach Us About Privacy," *Digital Privacy: Theory, Technologies and Practices*, 2007. http://www.heinz.cmu.edu/~acquisti/papers/Acquisti-Grossklags-Chapter-Etrics.pdf

[19] Nehf, James, "Shopping for Privacy Online," *Journal of Consumer Affairs*, vol. 41, 2007. http://ssrn.com/abstract=1002398

[20] World Wide Web Consortium, "P3P: The Platform for Privacy Preference," November 20, 2007. http://www.w3.org/P3P/

[21] EPIC (Electroni Privacy Information Center), "Pretty Poor Privacy: An Assessment of P3P and Internet Privacy," 2000. http://epic.org/reports/prettypoorprivacy.html

[22] Quantcast, "Description of Methodology: Delivering An Actionable Audience Service," 2008. http://www.quantcast.com/docs/display/info/Methodology

[23] Cranor, Lorrie Faith; Byers, Simon; Kormann, David, "Automated Analysis of P3P-Enabled Web Sites," *Proceedings of the 5th International Conference on Electronic Commerce*, 2003. http://lorrie.cranor.org/pubs/icec03-final.pdf

[24] Reeder, Robert; Kelley, Patrick Gage, McDonald, Aleecia; Cranor, Lorrie Faith, "A User Study of the Expandable Grid Applied to P3P Privacy Policy Visualization," *Workshop on Privacy in the Electronic Society*, 2008.

[25] Jensen, Carlos; Sarkar, Chandan; Jensen, Christian; Potts, Colin, "Tracking Website Data-Collection and Privacy Practices with the iWatch Web Crawler," *Symposium On*

Usable Privacy and Security, 2007. http://cups.cs.cmu.edu/soups/2007/proceedings/p29_jensen.pdf

[26] Krishnamurthy, Balachander; Wills, Craig E., "Privacy Diffusion on the Web: A Longitudinal Perspective," *WWW 2009*, 2009. http://www2009.eprints.org/55/1/p541.pdf

[27] Consumers Union, "Consumer Reports Poll: Americans Extremely Concerned About Internet Privacy," 2008. http://www.consumersunion.org/pub/core_telecom_and_utilities/006189.html

[28] Harris Interactive, "Majority Uncomfortable with Websites Customizing Content Based Visitors Personal Profiles," 2008. http://www.harrisinteractive.com/harris_poll/index.asp?PID=894.

[29] TRUSTe, "2009 Study: Consumer Attitudes About Behavioral Targeting," 2009. http://www.truste.com/about/bt_overview.php

[30] Turow, Joseph, et al., "The FTC and Consumer Privacy in the Coming Decade," 2006. http://works.bepress.com/cgi/viewcontent.cgi?article=1011&context=joseph_turow

[31] Pew, "Project Poll Database," *Pew Internet and American Life Project*. May 2000. http://webapps.ropercenter.uconn.edu/cfide/psearch_v11/webroot/question_view.cfm?qid=429137&pid=53&ccid=53

[32] Pew, "Project Poll Database," *Pew Internet and American Life Project*, November 2006. http://webapps.ropercenter.uconn.edu/cfide/psearch_v11/webroot/question_view.cfm?qid=1728783&pid=53&ccid=53

[33] Lazarus, Daivd, "It's Impressive, Scary to See What a Zaba Search Can Do," *San Francisco Chronicle*, April 15, 2005. http://www.sfgate.com/cgi-bin/article.cgi?f=/c/a/2005/04/15/BUG3JC8U341.DTL

[34] PRC (Privacy Rights Clearinghouse), *PRC's Privacy Update*, no. 3:3, May 18, 2005. http://www.privacyrights.org/newsletter/050518.htm

[35] Lazarus, Daivd, "Search Site to Add Free Blogs," *San Francisco Chronicle*, August 26, 2005. http://www.sfgate.com/cgi-bin/article.cgi?f=/c/a/2005/08/26/LAZ.TMP

[36] Microsoft, "Microsoft Online Privacy Statement," May 2008. http://privacy.microsoft.com/en-us/fullnotice.mspx

[37] Google, "Frequently asked questions for the Google Analytics data sharing options," 2009. http://www.google.com/support/googleanalytics/bin/answer.py?hl=en&answer=87515

[38] Story, Louise; Stone, Brad, "Facebook Retreats on Online Tracking," *New York Times*, November 30, 2007. http://www.nytimes.com/2007/11/30/technology/30face.html

[39] NAI (Network Advertising Initiative), "Opt Out of Behavioral Advertising," 2009. http://www.networkadvertising.org/managing/opt_out.asp

[40] Guess, Andy, "Taking Facebook Back to Campus," *Inside Higher Ed*, 2008. http://www.insidehighered.com/news/2008/10/24/socialweb

[41] Smith, Steve, "Obama signs on with Facebook, Twitter and MySpace," *The Tech Herald*, May 4, 2009. http://www.thetechherald.com/article.php/200919/3593/Obama-signs-on-with-Facebook-Twitter-and-MySpace

[42] FTC (Federal Trade Commission), "FTC Policy Statement on Deception," 1983. http://www.ftc.gov/bcp/policystmt/ad-decept.htm

[43] Turow, Joseph, "Americans & Online Privacy: The System is Broken," Annenberg Public Policy Center, 2003. http://www.asc.upenn.edu/usr/jturow/internet-privacy-report/36-page-turow-version-9.pdf

Appendix A – FTC Statute Codes

The FTC categorizes the complaints it receives from consumers with Codes for various Statutes that it enforces. Below is a list of all the Statute Codes.

Alternative Fuel/Fueled Vehicles Rule	Gramm-Leach-Bliley
Appliance Labeling Act	Health Violations
CAN-SPAM Act	Hobby Protection Act
Care Labeling Rule	Holder-in-Due-Course Rule
Children's Online Privacy Protection Act	Home Repair Deceptions
Consumer Leasing Act	Identity Theft Act
Contact Lens Rule	Internet Access Related Services Violations
Country of Origin	Jewelry Guides
Credit Practices Rule	Leather Goods Guide
Door-to-Door Sales Rule	Magnuson-Moss Warranty Act
Electronic Fund Transfer Act	Mail or Telephone Merchandise Order Rule
Energy Savings Violations	Media Violence
Equal Credit Opportunity Act	Negative Option Rule
Fair Credit Billing Act	Pay Per Call Rule
Fair Credit Reporting Act	Prescription Release Rule Violation
Fair Debt Collection Practices Act	Rule / Other
Fair Packaging & Labeling Act	R-Value Rule
Feather/Down Guides	Telemarketing Sales Rule
Franchise Rule	Textile Act
FTC Act Sec 5 (BCP)	Truth-In-Lending Act
Fuel Rating Rule	Unordered Merchandise
Funeral Rule	Used Car Rule
Fur Act	Watch Guides
General Privacy	Wool Act

Appendix B – Free Text Complaint Coding Facets

In our qualitative analysis of user complaints, we categorized the complaints using the sets of facets. One set for the type of concern the user had, one for the type of website involved and one for the type of data involved.

Complaint Concerns

Each free-text user complaint was assigned codes based on the nature of the concern. Codes are not mutually exclusively as concepts are often related and complaints may contain multiple concerns. The codes were distilled from the most common concerns found in the pilot evaluation.

Aggregation	User concerned that company is aggregating data or building a profile about them, and user has NO relationship with the aggregator
Excessive information	User thinks too much information is required for the purpose of completing a given transaction
Fraud	User has received potentially fraudulent communications
ID theft	User is concerned with or has been victim of identity theft
Marketing	User concerned with receipt of unsolicited marketing / spam
Public display	User concerned with the public display of personal or private information
Security	User concerned with security, breach, or information system integrity issues
Sharing	User concerned that company with whom user has relationship is sharing user data
Threat	User perceives potential for physical harm / stalking / personal threats
Control	User concerned with lack of ability to access, edit, delete, or remove from public view private or personal information collected by a website

Service Type and Data Type Codes

Service Type
Broker (data broker or portal)
Search (search engine)
ISP
Email
Software
Socinet (social network)
Ecom (e-commerce)
Gov (government)
Other (other)

Data Type
Contact (name, address, phone, email, ssn
Demo (demographic)
Computer (IP address, browser info, OS)
Interactive (browsing behavior, search history)
Financial (credit info, purchase history, account numbers)
Content (communications, files)

Appendix C – Privacy Policy Coding Facets

We conducted an analysis of the top 50 websites privacy policies using the facets below. Each policy received an evaluative code of **YES**, **NO**, or **UNCLEAR** for each category. **YES** and **NO** codes were only assigned if the distinction could clearly be made based on the wording of the site's privacy policy. **UNCLEAR** was assigned if the given information was not specified or was to too nuanced or vague to be determined.

Types of User Data Collected. In order to be coded **YES** or **NO** the policy must explicitly state whether the site governed by the policy collects the given information type.

Contact	personal contact information, including name, mailing address, email, or phone number
Demographic	demographic data, including gender, age, race, or income
Computer	IP address, browser type, or operating system
Interactive	browsing behavior or search history
Financial	account status or activity, credit information, or purchase history
Content	contents of personal email, textual communications, stored documents or media files (includes services which offer online content storage or hosting)

General Practices. In order to be coded **YES** or **NO** the policy must explicitly state whether the site governed by the policy allows the given behavior.

Ad Customization	User data may be used for the purpose of customizing advertising for users
Public Display of Personal Information	User data may become publicly viewable or publicly available as part of the service offered (includes information voluntarily made public in the case of services which offer public facing user profiles)

Practices Regarding Stored Data. In order to be coded **YES** or **NO** the policy must explicitly state the site's practice regarding the given facet.

User Access	Users may access and correct personal data collected (user is allowed access to at least *some* personal data beyond just contact information)
Data Retention	Explicitly stated duration of retention for personal data collected (must state the specific amount of time, even if that duration is indefinite)
Event of Acquisition	User will be notified and given the chance to delete personal data in the event of bankruptcy or acquisition/merger
Data Purchase	Site purchases data from third parties to supplement or enhance their aggregate user data

Data Sharing. In order to be coded **YES** or **NO** the policy must explicitly state the whether the site will share user data with the given entity.

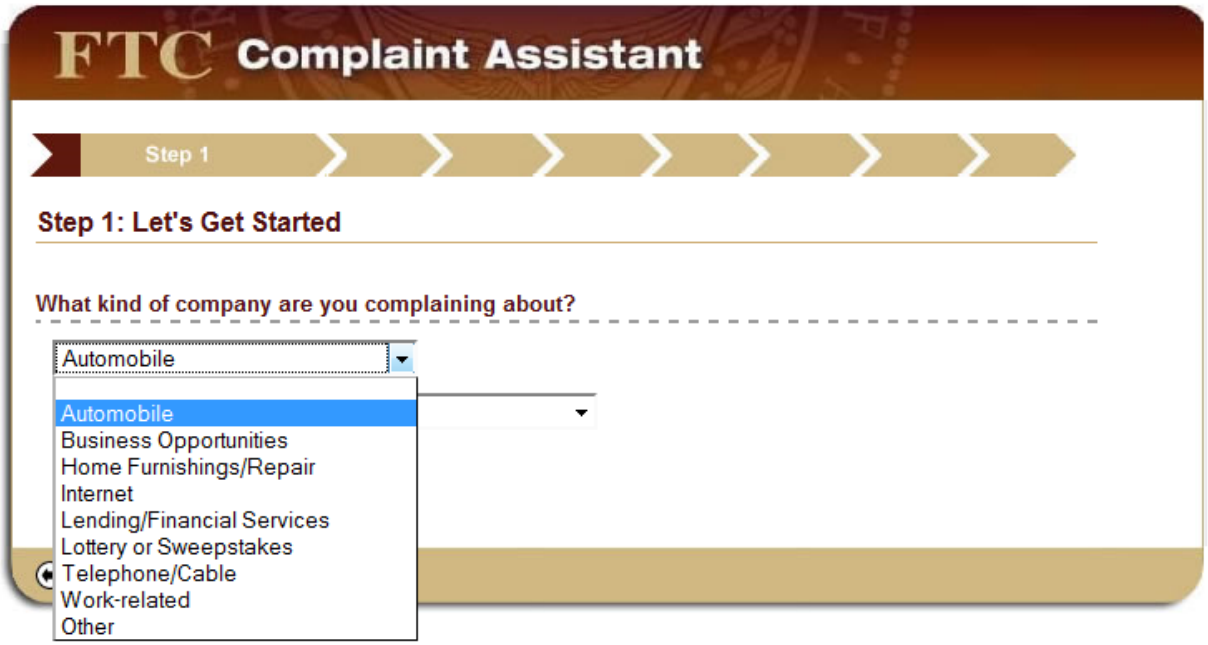
Sharing with Affiliates	User data may be shared with affiliates and subsidiaries of the primary entity who are bound by the same privacy practices
Sharing with Contractors	User data may be shared with third-party contractors (entities employed to assist with site administration, data analysis, or transaction processing, and who are bound by the same privacy practices)
Sharing with Third Parties	User data may be shared with third parties not subject to the same privacy practices (includes sharing of aggregate data which may not contain personally identifiable information)

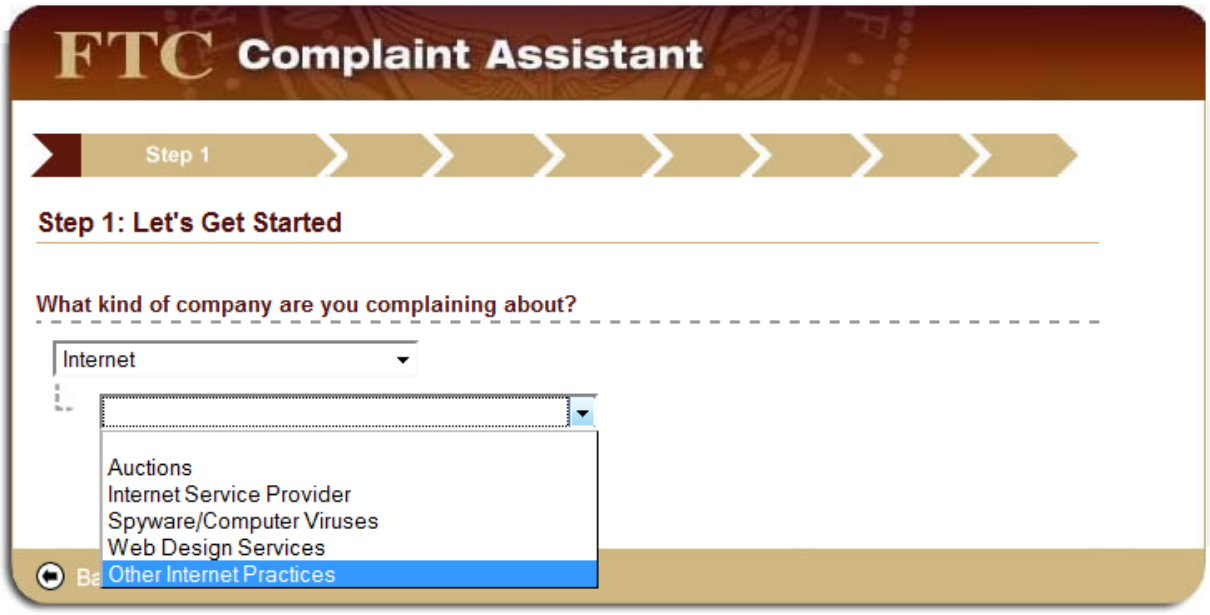
Appendix D – Screenshots of FTC Complaint form interface

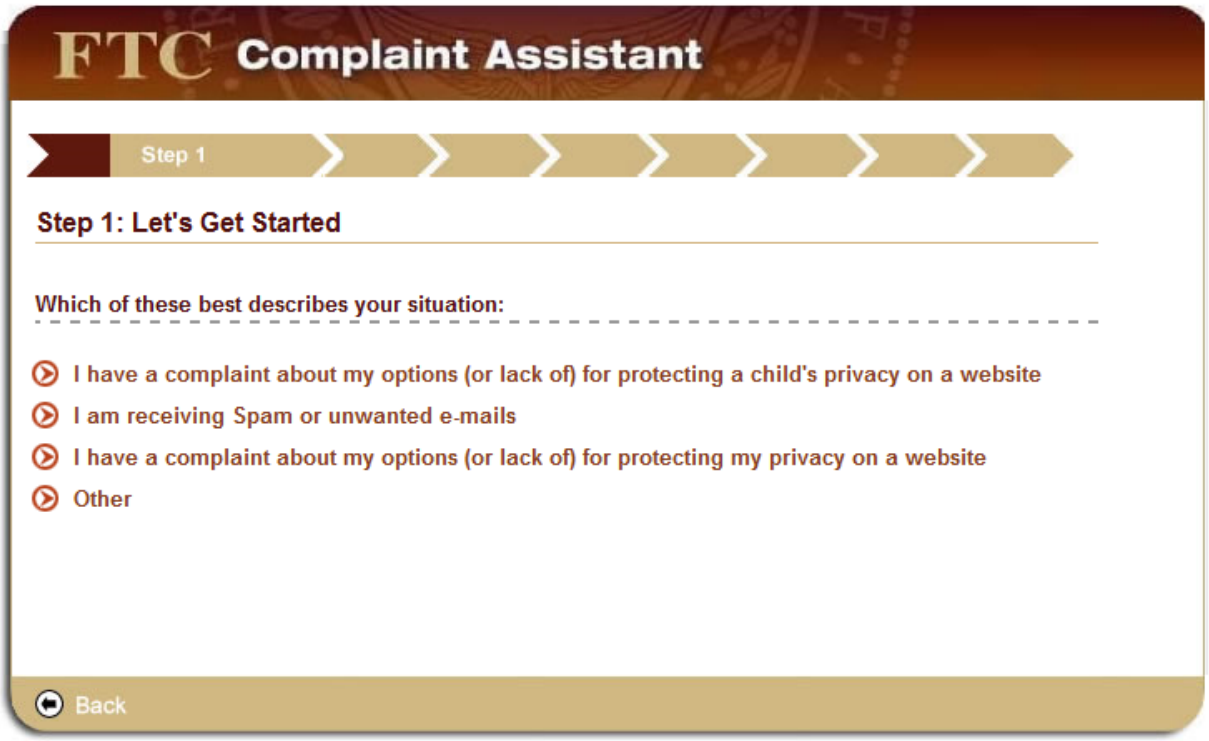
From 2004-2008, the FTC received 1,315,179 complaints in the category for ID Theft, as opposed to 6,713 for General Privacy and 42,765 for CAN-SPAM. While much of this could be due to the proliferation of these types of concerns as well as the FTC's clear ownership of the Identity Theft domain, there could also be bias introduced simply on account of the method in which these complaints are submitted.

The screenshots below show just the first few steps required to navigate to the section in which a consumer can make a complaint about general privacy problems on the Internet. Notice that the first question asks if the complaint is about identity theft. For an average user that discovers a website using their personal information in a manner they perceive as an invasion of privacy, such as a data broker selling their profile information, the term identity theft may seem appropriate, regardless of whether or not they are an actual victim.


FTC Complaint Assistant
Step 1
Step 1: Let's Get Started
Is your complaint concerning identity theft?
Yes
No
Is your complaint most closely related to:
Debt collectors or debt collection practices
Credit Reports
Dissatisfaction with other business practices


FTC Complaint Assistant
Step 1
Step 1: Let's Get Started
What kind of company are you complaining about?
Automobile
Automobile
Business Opportunities
Home Furnishings/Repair
Internet
Lending/Financial Services
Lottery or Sweepstakes
Telephone/Cable
Work-related
Other


FTC Complaint Assistant
Step 1
Step 1: Let's Get Started
What kind of company are you complaining about?
Internet
Auctions
Internet Service Provider
Spyware/Computer Viruses
Web Design Services
Other Internet Practices


FTC Complaint Assistant
Step 1
Step 1: Let's Get Started
Which of these best describes your situation:
I have a complaint about my options (or lack of) for protecting a child's privacy on a website
I am receiving Spam or unwanted e-mails
I have a complaint about my options (or lack of) for protecting my privacy on a website
Other
Back

Appendix E – Websites with Most Web Bugs

The figure below depicts the websites from the top 50 with the greatest number of bugs reported in March 2009. For another graph that shows both this quantity of bugs per site as well as the number of sites each bug was found on, follow this link: http://www.knowprivacy.org/newsite/web_bugs_analysis.html.