A Personalized Approach to Web Privacy—Awareness, Attitudes and Actions

Craig E. Wills and Mihajlo Zeljkovic

Computer Science Department
Worcester Polytechnic Institute
Worcester, MA USA

Abstract

Purpose - Our research seeks to understand user attitudes and actions regarding Web privacy in the context of educating users on what information is obtained and inferred about them via behavioral tracking.

Design/methodology/approach - We take a novel approach to personalize user awareness by embedding JavaScript in a Web page to examine portions of a user's Web browser history to ascertain Web sites that the user has visited. We then personalize information reported to the user about what third-party sites are tracking the user's behavior along with demographic information these sites may be inferring from these visited sites and the user's geographic location. In this context users are then surveyed regarding their attitudes towards tracking and actions they take to control it.

Findings - We found that 63% of users agreed with a statement of concern for third parties monitoring activities, about half of our respondents agreed with a concern for knowledge about a user's location and a little more than half agreed to concern about inference of demographic information. We found that females are more concerned about these issues than males.

In terms of possible actions, a majority of users report using an ad blocker tool and even more delete cookies at least some amount of time. Using an opt-out mechanism or removing browser history are done by less than 20% of users. Despite expressing more concern for information known by third-parties, females are not significantly more likely to take actions that may limit what is leaked to these third parties. A contributor to this discrepancy is that females were much less likely to know their settings for many of the actions indicating less familiarity with them.

Social implications - Web privacy is an important social issue so helping users to better understand tracking of their actions along with what actions can be taken to limit tracking is important. The project Web site continues to exist and educate users.

Originality/value - The work takes an original approach to educating users on what third-parties learn about them and follows up that knowledge with a survey of user attitudes and actions on this important topic.

Keywords Web, Privacy, Behavioral Advertising, Tracking

Paper type Research

Security Tech EX2011
Home Depot v. Security Tech
IPR2024-01420

1 Introduction

As the Web has evolved, the use of "third-party" sites to serve advertisements on "first-party" sites that users choose to visit has continued to grow. The work of these third-party sites also has evolved as they move from *contextual* advertising where ads are served only on the basis of the first-party site that a user has visited to *behavioral* advertising where advertisers seek to track the behavior of users across first-party sites to build up a "profile" of user activity and interests. Tracking of a user's browsing behavior raises privacy concerns, particularly if such a profile can be linked with a user's identity. These concerns have become a public issue and are being examined in the media, by privacy rights groups and by governmental agencies concerned with consumer rights.

Recent studies have examined user attitudes towards behavioral advertising (McDonald and Cranor, 2009; TRUSTe, 2009; Turow et al., 2009). These studies have primarily collected user opinions via user surveys and interviews. A result from one of these surveys indicates that 68% of Americans would "use a browser feature that blocks ads, content and tracking code that doesn't originate from the site they're visiting" (TRUSTe, 2009). Tools have been created that show users what third-party sites are present on Web pages that they visit (Ghostery, 2009; Adblockplus, 2009). However these tools do not show users what these third-party sites do with this information nor are these tools widely available to users of all browser platforms. They also require installation by a user, which is an impediment to use for many users. Limitations such as these contribute to lack of understanding by many users on the implications of behavioral advertising for their own personal privacy.

A problem in Web privacy research is educating users on the privacy considerations of techniques such as behavioral advertising. If users do not understand what behavioral advertising means for them then any attitudes they express or actions they take are based on perception. This problem motivates our research to find ways to educate users on what behavioral advertising personally means for them while doing so in a manner to minimize impediments for participation by all users. Once users understand what behavioral advertising means for them then they are in a better position to express informed opinions on its use.

Our approach for tackling this problem is to design a tool that is accessible to a large segment of users while providing personalized information. The approach provides a Web page that any user can access via any browser from anywhere on the Internet. We are able to personalize the information provided to users by using JavaScript embedded in the Web page to examine portions of the user's Web browser history to ascertain Web sites that the user has visited. We then personalize information reported to the user about what third-party sites are tracking the user's behavior along with demographic information these third parties may be inferring from these visited sites and the user's geographic location. Our project is called "WhatTheyKnow" because it helps users understand what "they" (the third-parties) know about them.

Taking this approach, our research study seeks to answer the following questions:

1. Are we able to attract and educate a broad audience of users by providing personalized privacy *awareness* where users visit Web site without needing to install software or browser extensions? Our approach of providing benefits to participating users is in contrast to previous studies that have relied primarily on surveys.

2. What *attitudes* do educated users have on tracking and other information inferred about them when surfing the Web? Our approach allows us to survey users in the context of better understanding on what is known or inferred about them.

3. What, if any, privacy-related *actions* do users take to limit the capability of third-parties to track a user's activities? Previous work has found gaps between user intentions and behaviors when it comes to privacy (Norberg et al., 2007; Paine et al., 2007). This question allows us to understand how user concern for tracking of browsing behavior translates into actions for limiting its effectiveness.

The remainder of this paper begins with a discussion of related work that places our work in context in Section 2. We go on to discuss details of the approach for conducting our research study in Section 3 and the results that we obtain from it in Section 4. We conclude with a summary and future directions for our work in Section 5.

2 Related Work

There are many directions of work that provide context for our work. These include consumer concerns for privacy, differences in attitudes based on demographics, the growth of third-party advertising Web sites to track user behavior, the possible actions users can take to control tracking and the possible gaps between user concerns on privacy and the actions users actually take. We examine relevant work for each of these directions in this section.

Concern for monitoring of Web activity is not new, such as shown by a 2002 consumer research study (Graeff and Harmon, 2002). A more recent study looking at everyday tracking and recording technologies found that users expressed the most concern of any technology for recording of their Web activities (Nguyen et al., 2008). A number of recent studies have been published that examine user attitudes concerning tracking by third-party aggregators for behavioral advertising (McDonald and Cranor, 2009; TRUSTe, 2009; Turow et al., 2009). This work has primarily been done through surveys and interviews of user opinions based upon what users know about tracking of activities.

Studies have also examined differences in attitudes based on demographics such as age or gender. These studies include (Sheehan, 1999; Zukowski and Brown, 2007). A recent study examines evolution of online privacy values since 2002 (Anton et al., 2010). Our work contributes to the study of attitudes for different user demographics as well as studying differences between users based upon the specific Web sites that they visit.

A separate direction of work has examined the increasing presence of third-party sites used for advertising and analytics in a longitudinal study showing the penetration of the top-10 third-parties amongst popular first-party sites growing from 40% in 2005 to 70% in 2008 (Krishnamurthy and Wills, 2009b). Updated results show this penetration has increased to over 70% in September 2009 (Krishnamurthy and Wills, 2009c). A separate study shows that these third-parties are not only increasing their tracking of users, but the browsing behavior of users can be linked to personal information and identifiers via online social networking sites (Krishnamurthy and Wills, 2009a), which also employ these same third parties. These results indicate that tracking of user behavior is growing and that users need to better understand what it means for them.

Another direction of work has focused on developing tools that help users understand the extent of this tracking and provide means to control it. Making a single request to a third-party server may leak private information by encoding the information in the URL. Tools such as AdBlock Plus (Adblockplus, 2009) or Ghostery (Ghostery, 2009) allow syntactic string matching of the URL for monitoring and blocking requests from known aggregators. RequestPolicy (Samuel and Zhang, 2009) takes a white-listing approach. However these tools do not always detect hidden third-party servers that appear to be part of the first-party domain (Krishnamurthy and Wills, 2009b).

Information is also stored and passed to servers via cookies. Browsers do provide means for users to control the sending of first- and third-party cookies. However, the removal of cookies, particularly first-party, can lead to errors on some sites (Krishnamurthy et al., 2007).

A number of third-party sites now provide a mechanism for users to opt-out of targeted advertisements via third-party cookies. Some of these sites are part of the Network Advertising Initiative (NAI), a cooperative of online marketing and analytics companies (Network Advertising Initiative, 2009). Users can opt-out of receiving targets ads by any or all of the NAI members through the creation of an “opt-out” cookie. One weakness in this approach is that if a user removes all their cookies, the opt-out cookies are lost. The Firefox

TACO extension (Targeted Advertising Cookie Opt-Out, 2009) makes these opt-out cookies persistent.

It is possible for a Web site to determine if a browser has visited any given URL using a combination of CSS and JavaScript (Janc and Olejnik, 2010). This technique can produce a partial browsing history without the use of cookies, and may reveal private information, such as which bank and OSN sites a user visited. Users can prevent such attacks with the SafeHistory Firefox extension (Jackson et al., 2006).

Modern browsers have added features, such as InPrivate Browsing for Internet Explorer 8, incognito mode for Chrome, and Firefox Private Browsing. These features allow users to create a "private" browsing session, also known as "porn mode" for its suspected use (Keizer, 2008), where no history is recorded. In addition, Internet Explorer 8's InPrivate Filtering blocks requests to embedded objects that it encounters multiple times.

With many actions that are available for users to help control tracking of their activities, a key question in our research is to what degree these actions are employed by users. This question is particularly relevant when previous research has found gaps between user intentions and behaviors when it comes to privacy (Norberg et al., 2007; Paine et al., 2007). These studies have not specifically looked at actions taken in regard to tracking of behavior, but there is the potential that expressed user concern for tracking may not translate in actions to control it.

3 Approach

Our research seeks to understand user attitudes and actions regarding Web privacy in the context of educating users on what information is obtained and inferred about them via behavioral tracking. We take a novel approach for this research study by constructing a Web site that allows users to better understand what this tracking means for them and for them to share their opinions about it if they so choose. This approach was taken because we can personalize the experience for an individual user while minimizing what is needed from a user—simply visiting a Web site rather than requiring the user to install any tools.

To personalize the experience for each user we access the browser history of the user via JavaScript code to check if sites from a list of popular sites are included in the browser's history. Browsers maintain history information by default so that previously visited sites can be seen and their link color can be changed from the default when a page containing such a link is shown. It is not possible to list the contents of the history via such a script, but only to query whether or not specific URLs are contained within the history. For example, one of the sites in our list is `cnn.com`. If either the URL `http://cnn.com/` or `http://www.cnn.com/` is found in the browser history then the site `cnn.com` is marked as visited. The script works similarly for other sites in our list.

In the remainder of this section, we first provide an overview of a user's experience when visiting our Web site followed by further details of how the site works. The discussion references the architecture of our Web site shown in Figure 1.

3.1 Web Site Overview

Users visiting our site access its contents in two steps. In the first step, a user reads about the site and then clicks on a button, which invokes JavaScript code (step 1 of Figure 1) that probes the history of the browser for the presence of a list of popular Web sites. This list of visited sites is then sent to our server (step 2) where the third-parties used for each of the Web sites in the list is determined (step 3). A server-side script also uses demographic information about each site in the list to predict the gender and age range of the user. The JavaScript code is also used to determine a user's location via a service that maps Internet addresses to geographic locations. All of this information is then displayed to the user (step 4) as a personalized summary of what information is known and inferred about the user's behavior on the Web.

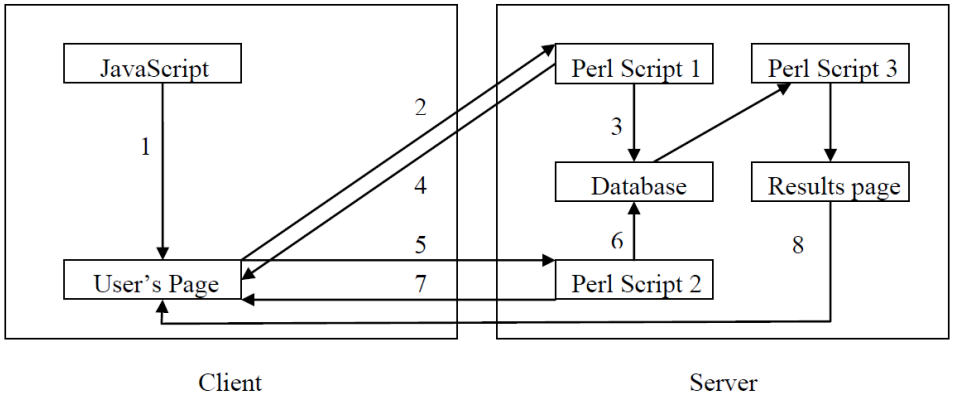

Figure 1: Site Architecture

Once the information is shown to a user, then the user has the choice to complete a second step. This step seeks feedback on the accuracy of age, gender and location information inferred about them as well as seeking feedback on the user's attitudes towards tracking and the user's use of actions that may prevent third-party sites from monitoring the user's browsing behavior. If a user provides this feedback then it is sent to our server, stored and acknowledged (steps 5-7). As an incentive for the user to complete this second step, the user can see the results for all users upon submission of results (step 8). More details on these steps are provided in the following.

3.2 Gathering User Data

We use modified JavaScript code (step 1 of Figure 1) from the work in (Raskin, 2009) with a different list of sites. We check each user's history for the top 1000 sites obtained from Quantcast (Quantcast, 2009) in July 2009 as well as additional lists of the most popular search engine, social networking and adult entertainment sites. The existence of such a scripting capability has been previously publicized for determining a user's social sites (Raskin, 2009) or a user's gender (Nolet, 2008). A more recent site checks a user's browser history against an extensive set of site lists from many categories of sites (Janc and Olejnik, 2010).

To obtain a user's location we use a script from (Find-IP-Address, 2009) to embed location data in the user's page based upon the Internet address. Our JavaScript code extracts country, state and city information. If the user agrees to participate in our study, then as shown in step 2 of the figure, the list of visited sites and location are submitted to our Web server via a HTTP POST command. We also send the browser type and list of browser plug-ins to understand how well these values uniquely identify a browser. The use of browser type, plug-ins and similar information for fingerprinting a browser has been studied extensively in another project (Eckersley, 2010).

3.3 Processing User Data

Once the user data is received by our server it is processed by a Perl script to first determine the user's gender and age based upon demographic data obtained from Quantcast for each of the sites in our list. The gender data contains the probability of being a male visitor for each site. If $p_1, p_2, ...p_n$ represent the probability of a user being male for each of n sites visited then similar to (Nolet, 2008) we use the formula that the probability of the user being male is

$$P(male) = \frac{1}{1 + \prod_{i=1}^{n} \frac{1-p_i}{p_i}}$$

Similar age-based data is available for ten age groups: 3-12, 13-17, 18-20, 21-24, 25-34, 35-44, 45-54, 55-64 and 65+. To determine the probability for each age range we use a similar calculation as done for gender except we do not include sites for which data from more than one age group is missing.

The Perl script also uses the list of visited sites to approximate the set of third-party sites that know about each site a user has visited. The Perl script uses third-party data that was obtained separately with the Pagestats Firefox extension (DeDeo, 2006), which is used to visit the home page of each site in our list and record all third-party servers accessed for each page. Once the Perl script obtains the list of visited sites then it simply looks up the stored third-party information for each site. It is important to understand that the set of third-party sites determined for a list of visited sites is intended for awareness and is likely only an approximation for any user for the following reasons:

1. The script cannot examine a user's entire history, but only query for specific sites. Therefore not all Web sites visited by a user containing third parties are likely to be queried by the script.

2. The set of third-party sites present on a visited site may change over time. The third-parties shown in the results were determined in October 2009.

3. Third-party sites typically use cookies to track users. If a user removes cookie information or blocks advertisements then the determined list of third-parties will not be accurate for the user.

4. While this script uses browser history to determine some of the sites a user visits, third-party tracking sites do not (to our knowledge) use browsing history to track a user's behavior, although (Raskin, 2009) suggests using it to determine a user's social networks. If a user periodically removes history then the list of visited sites determined by our JavaScript code may be smaller than the list of tracked sites.

Despite these approximations, the calculated set of third-party sites does provide a user with information on which third parties are tracking the user's Web browsing behavior. As shown in step 3 of Figure 1, the list of visited sites, location, the calculated gender, age and third-party sites, as well as the browser type and list of plug-ins are stored in a MySQL database.

3.4 Presenting Information to Users and Obtaining Feedback

In step 4 of Figure 1 the list of visited sites along with the known third-party sites for each visited site is shown to the user. A sample of this output is shown in Figure 2 where a user has visited two search engines, a social networking site and three other popular sites.

In addition to the list of visited sites and associated third-parties, a user is also shown the inferred location, gender and age. At this point the user can choose to stop or provide additional information as a follow-up to the reported data. We seek feedback via a survey from the user on three related aspects. First, we ask the user if the inferred location, gender and age range are "correct", "close to correct" or "incorrect". Second, we survey users about their attitudes concerning the information presented to them. We ask them to indicate if they "agree", are "not sure" or "disagree" with specific statements about concern for third-party monitoring of activities, concern that a user's location is known and concern that third-parties can infer demographic information.

Here is the list of **2 search engines** you have visited along with count and list of third-party tracking sites that know about you (list of site abbreviations).

Visited site	Count	Third-party sites
google.com	1	google
yahoo.com	3	yahoo, yimg, mplex

Here is the list of **1 social network** you have visited along with count and list of third-party tracking sites that know about you (list of site abbreviations).

Visited site	Count	Third-party sites
facebook.com	3	dblclick, atdmt, adv

Here is the list of **3 other popular sites** you have visited along with count and list of third-party tracking sites that know about you (list of site abbreviations).

Visited site	Count	Third-party sites
mail.yahoo.com	1	yimg
maps.google.com	0	
cnn.com	8	dblclick, omniture, scresearch, g-syndication, revsci, aol, dl-rms, vfive

Figure 2: Sample List of Visited Sites Along with Associated Third-Party Sites Reported to WhatTheyKnow Users

Finally, we ask users what, if any, actions they take to prevent or limit tracking of their browsing behavior. These six questions ask about use of ad blocker tools, cookie blocking, cookie deletion, opt-out mechanisms provided by third-party sites, browser history removal and private browsing session features.

Once a user has responded to each question (or chosen not to respond) and supplied "any additional comments about what you learned" in a text box then the user submits their responses to another Perl script on our site as shown in step 5 of Figure 1. This script locates the user's prior database record from step 3 by matching a recent record containing the same location, visited sites, browser type and list of plug-ins to update the record in step 6. The user is thanked with a confirmation in step 7 and as an enticement to participation the user is given a link to see the results of all participants in step 8 where the results page is generated by a Perl script from information stored in the database.

4 Results

The Web site at `http://whattheyknow.cs.wpi.edu/` went public in January 2010. We did not try to control the set of users with access to the site, but rather sought to encourage broad participation through a set of initial "seed" email messages to various groups with the expectation that word-of-mouth would extend use of the site to a much larger audience. This is a similar approach as used in (Anton et al., 2010). Broad participation was a key objective for the education aspect of the work. We expected that such a "seed-and-spread" approach would not necessarily lead to a representative sample group, but a large enough number of users would allow us to focus our results analysis on identifiable sub-groups of users rather than on the aggregate set of users. We are also able to profile the set of Web sites visited by our users relative to the population at large to better understand potential biases by our users.

We used a number of seed email messages beginning with an initial announcement of the site to faculty, students (primarily undergraduates) and staff at our institution. WPI is a private university with approximately 4000 full-time students and 300 faculty primarily in fields of science and engineering with support from 500 staff. Students and faculty are roughly 70% male with staff gender approximately balanced. Subsequently, we followed up this announcement with a message to a mailing list of over 2500 computer science educators from around the world, to a mailing list of over 100 people in an author's local neighborhood and

to approximately 50 people doing Internet measurement and Web privacy work. We have evidence that the existence of the site spread to many others that were never directly contacted by us.

3749 users had visited the site by the end of February 2010 and completed the first step of running the JavaScript program to see their personalized results. 1853 (49%) of these users then followed up these results by completing the survey step with most questions answered by between 1750 and 1800 users.

In the remainder of this section we provide results from our analysis for the roughly 1800 users taking the survey unless otherwise noted. We first provide a profile on respondents followed by their attitudes on various aspects of tracking. We then examine users actions taken to control tracking and finish with additional results.

4.1 Profile of Respondents

In our survey about Web privacy, we choose not to explicitly ask for user demographic information, but by correlating user responses on the correctness of location, gender and age inferences we can determine information about user demographics. We also characterize our set of users based upon the set of sites that they visit. We note that other potential factors affecting responses such as users' culture, abilities, personality and motivation are not explicitly examined in our work.

4.1.1 Location

Users were asked about the correctness of their location inferred from their Internet address. 64% responded that their location was correct, 23% indicated it was "close to correct" and 13% responded that it was incorrect. Although we choose the particular IP location service because it provided a script that integrated with our JavaScript code, we note that it provided reasonable performance given that 87% of users responded with correct or close to correct location.

Focusing on these users responding with correct or close to correct location, we find that 10% of our respondents are located in Worcester, the city in which WPI is located, 26% were in the state of Massachusetts, the state in which WPI is located, and 87% were in the United States. These results indicate a representation of users in and around WPI, but with the majority of users beyond WPI.

4.1.2 Gender

As part of the results, we displayed the gender predicted based upon the set of sites found in the browser history along with a confidence on the correctness of the prediction. For example, such a result might be "Based upon the sites you visit there is a 62% chance that you are male." In the case that there was equal chance of male and female, such as when no sites could be found in the browser history, then we arbitrarily reported "Based upon the sites you visit there is a 50% chance that you are female." By indicating a gender, even if uncertain, allows us to ask the user if it is correct.

Overall, we reported that 52% of users were more likely to be male, 33% of users were more likely to be female and 15% were of equal likelihood. Based on user feedback, our predicted gender was correct for 64% of the users. As expected, we observe better correctness values if more sites are detected in a browser's history and if there is greater confidence in the predicted gender. Combining the results on predicted gender and user feedback on the correctness, we determine that 72% of our respondents are male while 28% are female.

4.1.3 Age

We made a similar prediction on the age range of a user based on the set of sites visited where a result such as "Based upon the sites you visit there is a 45% chance that your age range is 35-44" was shown to the user.

Unfortunately the predicted age range of users was skewed to only four (out of the possible ten) predicted ranges: 42% for 25-34, 39% for 35-44, 4% for 13-17 and 15% for 3-12 where this latter value was the default (with 10% confidence) when no visited sites could be found in the browser history. We conjecture this skewness occurs for a couple of reasons. First, there are more categories and less prediction confidence in each when compared to gender prediction. Second, because the age ranges in the demographic data were not uniform in size, the two 10-year ranges consistently had the highest predicted confidence values. In hindsight, it would have been better to combine some of the ten age ranges to have more uniform sizes for each.

Overall, the age range was predicted correctly for 19% of users while being "close to correct" for 23% and incorrect for 58%—again these results improve when more sites are detected in the browser history. These results do not provide a means to obtain the age-range breakdown of all users in our data set as was possible for gender. However in subsequent analysis, we do examine responses from users in the 25-34 and 35-44 age ranges who reported that their age range is correct or close to correct.

4.1.4 Visited Sites

Another means of characterization that our data set affords is to examine the set of sites visited by our set of users. The mean number of visited sites for all 3749 users was 15 while the median was 8. Table 1 shows the top-10 most visited sites where `google.com` was found in the browser history for 59% of the users. In 13% of the cases, no sites from our list were found in the browser history. If we only consider cases where at least one site was found in the history then `google.com` was found 68% of the time.

Table 1: Top-10 Visited Sites

Rank	Visited Site	% of All Users	% with History
1	google.com	59	68
2	facebook.com	51	58
3	youtube.com	42	48
4	maps.google.com	31	36
5	cnn.com	24	28
6	weather.com	24	28
7	yahoo.com	24	27
8	ebay.com	23	27
9	twitter.com	20	23
10	apple.com	19	22

While this set of visited sites is interesting, a more important question is how this set compares with the set for a broader range of users. Such a comparison would allow us to understand how the characteristics of our users compare with this broader range. We make such a comparison in Figure 3 where we compare the proportion of WhatTheyKnow users who visit popular sites with the same percentages as reported by Quantcast for these same sites in January 2010. The sites are listed in order of popularity according to Quantcast. The list of sites includes the top-20 Quantcast sites that were in the list of sites we checked along with any remaining sites in the top-20 sites for our set of users.

The list of visited sites is shown for four sets of users in Figure 3. The WhatTheyKnow results are for all 3749 users. The Survey results are for only the set of users that submitted the survey. The closeness in results between these two sets of users is meaningful because most of our analysis is done for this smaller set. The results show that sites such as `yahoo.com`, `msn.com`, `live.com` have much less of a presence

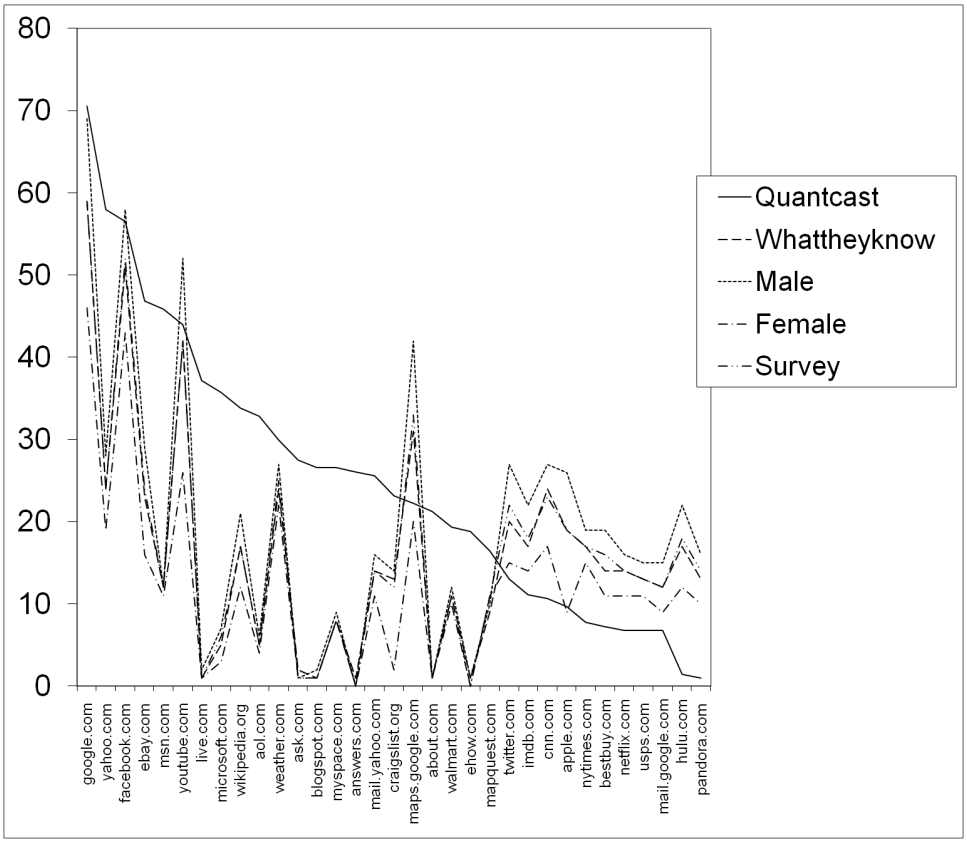

Figure 3: WhatTheyKnow User Profiles Relative to Published Profile

amongst our set of users. One reason for this discrepancy could be that our set of users simply do not have the same profile as those measured by Quantcast. Another reason might have to do with our methodology where we only test the presence of a site's home page in the browser history. If a user does not visit the home page of a site then our JavaScript code does not detect it. Despite this limitation, Figure 3 shows some sites detected at higher rates than observed by Quantcast. These sites include `maps.google.com`, `cnn.com` and `hulu.com`.

The figure also shows results for male and female users where males consistently have larger representation than females for the list of sites. The largest differences between the two genders are for the sites `craigslist.org`, `maps.google.com` and `apple.com`.

4.1.5 Third-Party Sites

We also determined the third-party sites that were most prevalent for our set of users. These results are shown in Table 2 where `doubleclick.net` was present on an average of 7.8 visited sites per user. As a comparison, in March 2010 we made use of the same methodology as described in (Krishnamurthy and Wills, 2009b) to obtain the list of third-parties used by a set of over 1000 popular sites. The rank ordering based on these separate results is shown for each site in Table 2.

As expected, the table shows much similarity between the top third-party sites for our WhatTheyKnow users and those sites found for a large list of popular sites. The lists contain seven common third-parties with the three other third-party sites falling just out of the top-10 in the other ranking.

Table 2: Top-10 Observed Third-Party Sites

Rank	Third-Party	Ave. # of Sites Per User	Rank in March'10 for Methodology of (Krishnamurthy and Wills, 2009b)
1	doubleclick.net	7.8	1
2	atdmt.com	3.9	6
3	google-analytics.com	3.9	2
4	omniture.com	3.8	7
5	quantserve.com	3.4	4
6	scorecardresearch.com	2.6	5
7	advertising.com	2.6	14
8	yieldmanager.com	2.1	9
9	revsci.net	1.7	11
10	yimg.com	1.6	13

4.2 User Attitudes on Tracking

Users were shown their list of detected visited sites and associated third-parties such as illustrated in Figure 2. We then asked users a number of questions, including three questions regarding user attitudes on various aspects of tracking. The results for all WhatTheyKnow users completing the survey and answering these three questions are:

1. I am concerned that third-party tracking sites have this level of monitoring of my activities.

   63% Agree
   25% Not sure
   12% Disagree

2. I am concerned that Web sites have this level of information about my location.

   48% Agree
   27% Not sure
   26% Disagree

3. I am concerned that third-party tracking sites can infer information about demographic information such as age and gender based on the sites I visit.

   54% Agree
   25% Not sure
   21% Disagree

With a sampling error of $\pm 2\%$ (95% confidence interval), these results show statistical significant differences between the three responses. Users are most concerned about monitoring of activities with 63% agreeing to a statement of concern on this issue and only 12% disagreeing. This result is similar to a survey by Turow et al. where 66% of Americans do not want behavioral advertising (Turow et al., 2009).

There is less concern about revealing a user's location, which is determined based upon the Internet address of the user's machine, although only 26% indicate no concern for this information. Concern for inference of demographic information lies between the other two pieces of information.

While the overall results represent a large number of respondents we have no control over the nature of these users—for example as reported in Section 4.1.2 we know that 72% of our respondents are male. In the next step of our work we took these three questions and analyzed their results based upon four independent

characteristics about a user: gender, age range (limited to a subset of users as described in Section 4.1.3), location and concern about monitoring of activities. This last characteristic is based on a user's response to the first question and is intended to understand how a user's level of concern correlates to other concerns and actions. The results of this analysis are shown in Table 3.

Table 3: Attitudes Based on Four User Characteristics

		Gender		Age		Location			Concerned w/ Monitoring	
Attitude	All	M	F	25-34	35-44	Worc	US-Worc	non US	Agree	Disagree
Concerned that third-party tracking sites monitor activities.										
Agree	63	*61*	*67*	63	**69**	**54**	65	65	100	0
Disagree	12	*14*	*7*	15	10	17	11	13	0	100
Concerned that Web sites have information about location.										
Agree	48	**45**	**55**	47	44	47	49	52	**69**	**5**
Disagree	26	**29**	**17**	30	27	30	24	23	**11**	**88**
Concerned that tracking sites can infer demographic info.										
Agree	54	*52*	*58*	53	53	48	54	57	**76**	**10**
Disagree	21	**24**	**15**	26	21	25	21	23	**7**	**81**

The table drops the "not sure" responses for each question and focuses on respondents that agree or disagree with each statement. The second column in Table 3 simply repeats the result of all users for easier comparison. Responses that are statistically significant at the 95% level in comparison with responses from users without the given characteristic are shown in **bold** font. Statistically significant results at the 90% level are shown in *italics* font.

The results show that for all three questions females have more concern about the information than males. In all cases this difference is statistically significant at the 90% or 95% level. Recent studies (TRUSTe, 2009; Turow et al., 2009) did not specifically report differences in attitudes between male and female users, although previous studies have generally found females to have more privacy concerns (Graeff and Harmon, 2002; Sheehan, 1999).

There are not strong distinctions in results based on the 21% of users we could identify as in or close to the 25-34 age range or the 19% of users in or close to the 35-44 age range. We did observe a statistically significant difference for the 35-44 age-range users that showed a concern for monitoring of activities by third parties.

In analyzing the results based on location we separated the users for whom the location was correct or close to correct into three groups: those located in Worcester, MA (Worc); those located outside of Worcester, but in the United States (US-Worc); and those located outside of the United States (non US). We expect the Worcester results to exclusively represent students, faculty and staff at WPI and thus wanted to explicitly separate out this population for study. As shown, the only statistically significant result is that Worcester users are less concerned with monitoring of activities than all other respondents. Although we do not know the age composition of WPI respondents we do know that 70% of the email messages sent to initially publicize the site went to undergraduate WPI students and results in (Turow et al., 2009) found this age group the most accepting of tailored ads. This finding likely explains the result we obtained.

Finally, we characterized users by their concern for monitoring of activities to see if this concern correlated to other concerns. By definition, the first result shows 100% values, but as expected the responses for the other two questions correlate to the answer for the first question with 69% of users who expressed concern for monitoring activities also expressing concern for Web sites having location information. Similarly, 76% of users expressing concern for monitoring also expressed concern about inferring demographic

information. Correlation between statements of disagreement were even stronger at 88% and 81%.

We next used our data set to understand the variation in user response according to the sites they visit. For this portion of the analysis we examined the characteristics for the set of users confirmed to have visited the most popular sites of our set of users in Figure 3. Table 4 shows selected results for the top few sites as well as other sites exhibiting significant differences for these responses or in the set of actions taken in the following section.

Table 4: Attitudes Based on a Sampling of Visited Sites

	Visited Site								
Action	Google	Facebook	CNN	Weather	Yahoo	Twitter	Hulu	Pandora	USPS
Concerned that third-party tracking sites monitor activities.									
Agree	63	63	**72**	62	65	62	61	**56**	60
Disagree	12	13	9	12	12	14	14	14	14
Concerned that Web sites have information about location.									
Agree	47	46	48	43	46	**42**	42	**38**	**38**
Disagree	28	29	25	28	25	**34**	*34*	**36**	30
Concerned that tracking sites can infer demographic info.									
Agree	52	50	55	49	54	**48**	**45**	**43**	47
Disagree	23	24	20	25	23	26	28	26	24

These results show few significant results for concern on the monitoring of activities with only CNN users at 72% (when compared with non-CNN users at 60%) more likely to agree with concern and Pandora users at 56% (when compared with non-Pandora users at 66%) less likely to agree with concern. Users of a number of sites were significantly less concerned that Web sites have information about their location when compared with non-users of these sites. These sites include Twitter, Hulu, Pandora and USPS (United States Postal Service). Users of three of these sites—Twitter, Hulu and Pandora—were also significantly less likely to be concerned about inference of demographic information.

4.3 User Actions on Tracking

In the next portion of our analysis, we examined how these attitudes towards privacy tracking translated into actions taken by users to control it. Preventing the disclosure of location based upon Internet address is difficult because a user's Internet address is contained in each Web request. One prevention method is to use a Web proxy that services requests from users in a number of locations and consequently anonymizes the location of users making use of it. Similarly preventing tracking sites from inferring demographic information once they know the set of sites a user visits is even more difficult. Other than preventing tracking altogether, users could try to "pollute" their profile by intentionally visiting sites that would cause their inferred profile to be inaccurate. While possible, we did not survey users if they took any of these actions

Rather we focused on six questions regarding actions discussed in Section 2 that users can take regarding blocking ads, blocking and deleting cookies, opting out of targeted ads, removing history and using private browsing. The specific questions and survey results for possible answers are as follows:

1. Use of ad blocker tools will prevent some tracking by third-parties identified for your set of visited sites. Do you use any ad blocker tools to prevent the display of advertisements in Web pages?

55% Yes
33% No
11% Don't know

2. Use of cookie blocking by your browser will prevent some tracking by third-parties identified for your set of visited sites. Do you:

   37% Allow all cookies (Internet Explorer and Firefox default)
   43% Allow cookies for only sites I visit (block cookies to third-party sites)
   3% Allow no cookies (block cookies for all sites)
   17% Don't know

3. How often do you delete cookies?

   21% Often
   52% Sometimes
   21% Never
   6% Don't know

4. Some third-party sites provide an "opt-out" mechanism to avoid receiving targeted ads. The Network Advertising Initiative (NAI) is a cooperative of such sites. Do you use opt-out cookies for third-party sites?

   16% Yes
   55% No
   29% Don't know

5. Periodic removal of browser history prevents scripts like this one from detecting what sites you visit, but will not prevent third-party sites from tracking your behavior (they typically use cookies). What browser history settings do you use:

   68% Use browser default for managing history
   15% Clear history when browser closes
   4% Set browser to not remember history
   12% Don't know

6. Newer versions of browsers have features to create a "private" browsing session where history is not recorded. These include InPrivate Browsing for Internet Explorer, incognito mode for Chrome and Firefox Private Browsing. Do you use any of these features?

   33% Yes
   57% No
   10% Don't know

The results show that a majority of users report using an ad blocker tool, although the result is less than the 68% of respondents of a previous survey (TRUSTe, 2009) indicating they would use such a browser feature. Nearly three-quarters of users delete cookies at least some amount of time. In contrast, less than 20% of users report using an opt-out mechanism or removing browser history.

For easier comparison across the set of possible actions, we determine the percentage of users that have changed from the default for each. For some actions, we compute the percentage based on one possible answer while for other actions we sum the percentages of two possible answers. In all cases, we treat an answer of "don't know" as using the default for the action. Using this approach, the six possible actions by a user and answers contributing to the percentage for each are: use ad blocker ("yes"), block cookies ("block cookies to third-party sites" + "block to all sites"), delete cookies ("often" + "sometimes"), use opt-out

("yes"), remove history ("clear history on close" + "not remember history"), and use private ("yes"). The second column in Table 5 summarizes the results for all users for these six actions. In addition, the table shows the results based on the same four user characteristics used for analysis in the previous section plus one additional characteristic.

Table 5: Actions Taken Based on Five User Characteristics

		Gender		Age		Location			Concerned w/ Monitoring		Use Ad
Action	All	M	F	25-34	35-44	Worc	US-Worc	non US	Agree	Disagree	Block
Use Ad Blocker	56	55	56	52	**48**	61	55	53	56	53	100
Block Cookies	46	46	44	**39**	43	42	46	48	47	40	**54**
Delete Cookies	73	74	71	69	71	67	75	70	74	70	**79**
Use Opt-Out	16	15	20	12	11	16	17	13	16	15	*19*
Remove History	19	20	19	14	**9**	13	19	19	21	17	**24**
Use Private	33	**40**	**13**	*39*	32	39	32	38	32	**42**	35

In examining each of the characteristics there are no significant differences between male and female users except for the use of browsers' private browsing modes. Males are substantially more likely to use this feature than females, which may be explained by the "porn mode" nickname for this feature (Keizer, 2008). The lack of statistical difference for the other five privacy-related actions is notable because females were much more likely to express concern for leakage of information in the results shown in Table 3. One possible explanation for this discrepancy is that for all six actions, females were significantly (at 95%-level for five, at 90%-level for one) more likely to answer "don't know" as to their use of an action. These significant differences (not shown in the table) suggest less familiarity with these actions by females, which may limit their use in translating concern into actions. They may also indicate a lack of motivation in following up concern with actions.

Looking at age ranges, we see that in all cases except one, users in the two age ranges are less likely than all of our users to use the available actions and in three of these cases the differences are significant. However, the results show that users ages 25-34 are more likely to use the private browsing feature.

There were no significant differences found due to location. However there is generally a weak, but not statistically significant, correlation between user concern for monitoring and actions taken. For all actions, users agreeing with concern for monitoring report using actions at the same or slightly higher levels than average for all users. Similarly, users disagreeing with this concern use actions at a lower level than average except for the private browsing mechanism where its use is at a significantly *higher* level than average. These overall results indicate that the use of the private browsing feature is not well correlated with other possible actions available to users.

The last column in Table 5 includes all users that report using an ad blocker. This characterization is introduced to understand whether the use of one action correlates with the use of other possible actions. The results show there is such a correlation for all other actions except for private browsing. The correlation for most actions confirms what we would expect to be the case.

We next analyzed differences in preventive actions taken compared with the same set of visited sites used in Table 4. These results are shown for five of the possible actions in Table 6. The sixth action, removal of history, is not shown because its results for all sites is relatively low due to our approach of using browser history to determine if users visit a site. It means that if we detect a site *is* visited then it is less likely that a user has removed history.

The most popular sites, such as Google and Facebook, are less likely to show distinguishing characteristics since they encompass a majority of WhatTheyKnow users. The remaining sites tend to have some

Table 6: Actions Taken Based on a Sampling of Visited Sites

	Visited Site								
Action	Google	Facebook	CNN	Weather	Yahoo	Twitter	Hulu	Pandora	USPS
Use Ad Blocker	54	55	53	53	56	54	58	53	56
Block Cookies	42	44	46	**38**	**38**	44	44	**37**	41
Delete Cookies	71	71	71	**65**	69	71	**65**	**66**	**65**
Use Opt-Out	14	14	16	13	15	**12**	13	14	16
Use Private	35	36	33	31	32	**42**	**42**	36	35

actions that show significantly less use than the average for all sites, although again the private browsing feature is reported to be used significantly more for users of Twitter and Hulu.

4.4 Additional Results

We examined information regarding accuracy of the determined location, potential fingerprinting of browsers based upon browser type and plug-ins and user written comments. Results for each of these additional examinations are provided in the following.

4.4.1 Location Correctness

We asked respondents in our survey about the type of their location and found 56% at work/school (combining these because work is the expected daytime location for adults and school the expected daytime location for students), 40% at home and 4% at a public location. We also looked at the correctness of the location information based on the type of their location. Focusing on incorrect location, we previously reported that 13% of respondents indicated their location was incorrect. However in correlating this figure with location type we found that 12% of work/school and 12% of home users reported an incorrect location while 33% of public location users reported an incorrect location. This significant difference indicates that at least the IP location service we employed yields much poorer performance for public locations. We also compared the correctness of locations known to be in the U.S. and outside of the U.S. finding that 11% of each type were incorrect for no net difference.

4.4.2 Browser Fingerprinting

We also examined the uniqueness of browser configurations by computing a fingerprint of the browser type and list of installed plug-ins. The values not only contain names, but also version numbers, which help to make these strings distinct for each browser. In computing this fingerprint for our 3749 users we found 3166 (84%) unique browser fingerprints. These results indicate that such a fingerprint is not unique, but it is a possible means to track users, particularly when combined with other information such as IP address, even if users employ preventive measures. These results are similar in tone with a related project (Eckersley, 2010).

4.4.3 Written Responses

135 (7%) of the respondents who took the survey added a written response when submitting their answers. A noticeable number of comments were about the age or gender being incorrect, which is not surprising as

these demographics were intended more to get people's attention than to necessarily make accurate predictions with what was limited data in many cases. The most common comment was an appreciation for the site and the awareness it brought to users. The following are a sample of comments submitted by users.

> "I think the privacy issue is overrated. The age distribution about me was a joke."
>
> "Curious where your demographics data come from? Also, would like to have more choices on the multiple choice. I just picked "not sure" because either answer is not what I want to say. That is, I am concerned at a broad level about the way that we categorize people into statistical buckets, but I also recognize that my use of free services on the internet - gmail, wowwiki, google, etc... - demands that those companies make some money off of me - and so I have come to the decision that I will to some extent pay for these services via my data, so I am not sortof angry/grumpy/concerned/surprised. Also wish the private browsing mode was more nuanced–I have used before, but very rarely. I answered no."
>
> "I used to limit cookies from non-third party sites, but found too many sites broke if I did that. I often use NoScript as a privacy measure (but have found it causes problems since many legit sites forward traffic elsewhere). I would use these features more if the controls were more accurate."
>
> "Thanks for the info. I'm changing some of my settings!"
>
> "Made me realize that I ought to think more about my privacy settings and cookie settings."
>
> "Nice tool"
>
> "Interesting (and timely) idea. I look forward to hearing more about this project."
>
> "this is a great program and I hope it gets lots of use by many people and raises awareness of the issues!!"
>
> "Your questions opened my eyes to new features and protection mechanisms available. Thank you!"
>
> "Very interesting and yet disturbing!"

5 Conclusions

In this work we undertook a research study based upon a novel approach that educates users on what information about their browsing behavior is sent to third-party aggregators and the information that is inferred based upon their behavior. Our approach can be used by any user from any browser simply by visiting our Web site where JavaScript code probes a browser's history to determine popular Web sites that have been visited. We then map the known third-parties of these visited sites, determine age and gender demographics based upon the list of visited sites and use an IP location service to look up a user's location. This information is shown to users with a follow-up survey used to gauge user attitudes towards the availability of this information to third-parties as well as find out what actions users take to protect their privacy.

We sought answers to three research questions in our study. Our first question was about the viability of this approach for attracting a broad representation of users. We had nearly 4000 participants use the site in the first couple of months it was available to learn what is known about them with nearly half of these participants responding to a follow-up survey. These numbers are significant representing a number of identifiable sub-groups. Written feedback from users yielded a number of comments from users who appreciated what they learned from the site.

Our second research question examined user attitudes in the context of this knowledge. We found that 63% of users agreed with concern for third parties monitoring activities while 12% disagreed and the

remainder were not sure. This level of concern is comparable to other studies that have been done. About half of our respondents agreed with a concern for knowledge about a user's location with a little more than half agreeing to concern about inference of demographic information. In examining these responses based on specific user characteristics we found that females are more concerned about these issues than males. We also found strong correlation between the responses of a user for each of these three issues. We found that users of sites such as Twitter and Pandora showed less concern for Web sites having location information or being able to infer demographics.

Our third research question focused on what actions users take to control tracking of their actions. A majority of users report using an ad blocker tool and even more report deleting cookies at least some amount of time. Using an opt-out mechanism or removing browser history is done by less than 20% of users. Users expressing concern for monitoring were only slightly more likely to take preventive actions. Users who blocked ads were more likely to take other actions.

Despite expressing more concern for information known by third-parties, females are not significantly more likely to take actions that may limit what is leaked to these third parties. A contributor to this gap between expressed concern and action is that females were much less likely to know their settings for many of the actions indicating less familiarity with them. Males were much more likely to use the so-called "porn mode" private browsing feature of browsers.

Moving forward, the WhatTheyKnow Web site continues to live and educate users. There are also a number of directions for future work. First, we could improve our ability for analysis by directly asking users for demographic information instead of trying to infer it. We could also improve our calculation of the age range and allow users to provide sites instead of only relying on browser history.

The tool could be further used in a teaching role to help educate students on privacy and security. One of the authors has already used the site in a social implications of computing course. Use of before and after testing would help assess any behavioral changes from use of the tool. While we did receive some user comments indicating what they learned, our own survey would have been improved if we had specifically asked users how their own attitudes and actions were changed based upon the results instead of simply asking about their attitudes and actions.

A related direction is to use ideas from this tool to explore how users make privacy-related decisions. Such a tool could help researchers better understand why users are concerned with monitoring of activities on the Web, yet do not necessarily take actions to control such monitoring.

A final direction is to not just ask users about what protective actions they take, but to develop a tool that can directly determine what actions are used and provide users personalized recommendations on what changes in protections they could make. Quantifying the potential benefits of privacy protection actions for individual users could lead to more effective use of these actions.

References

Adblockplus (2009), "Adblock plus: Save your time and traffic", available at: `http://adblockplus.org/`.

Anton, A. I., Earp, J. B., Young, J. D. (2010), "How Internet users' privacy concerns have evolved since 2002", *IEEE Security and Privacy*, Vol. 8, No. 1, pp. 21–27.

DeDeo, S., (2006), "Pagestats", available at: `http://www.cs.wpi.edu/~cew/pagestats/`.

Eckersley, P. (2010), "How unique is your web browser?", Electronic Frontier Foundation, available at: `https://panopticlick.eff.org/browser-uniqueness.pdf`.

Find-IP-Address (2009), available at: `http://www.find-ip-address.org/`.

Ghostery (2009), "Ghostery: Find out how Web sites are watching you", available at: http://www.ghostery.com/.

Graeff, T. R., Harmon, S. (2002), "Collecting and using personal data: Consumers' awareness and concerns", *Journal of Consumer Marketing*, Vol. 19, No. 4, pp. 302–318.

Jackson, C., Bortz, A., Boneh, D., Mitchell, J. C. (2006), "Protecting browser state from web privacy attacks", in *Proceedings of the International World Wide Web Conference*, Edinburge, Scotland, pp. 737–744.

Janc, A., Olejnik, L. (2010), "Feasibility and real-world implications of Web browser history detection", in *Proceeedings of the Web 2.0 Security and Privacy Workshop*, Oakland, CA USA, pp. 26–33.

Keizer, G. (2008), "Microsoft adds privacy tools to IE8", available at: http://www.computerworld.com/s/article/9113419/Microsoft_adds_privacy_tools_to_IE8.

Krishnamurthy, B., Malandrino, D., Wills, C. E. (2007), "Measuring privacy loss and the impact of privacy protection in web browsing", in *Proceedings of the Symposium on Usable Privacy and Security*, Pittsburgh, PA USA, pp. 52–63.

Krishnamurthy, B., Wills, C. E. (2009a), "On the leakage of personally identifiable information via online social networks", in *Proceedings of the Workshop on Online Social Networks in conjunction with ACM SIGCOMM Conference*, ACM, Barcelona, Spain, pp. 7–12.

Krishnamurthy, B., Wills, C. E. (2009b), "Privacy diffusion on the web: A longitudinal perspective", in *Proceedings of the World Wide Web Conference*, Madrid, Spain, pp. 541-550.

Krishnamurthy, B., Wills, C. E. (2009c), "Privacy diffusion on the web: A longitudinal perspective (updated graphs)", Submitted as public comment to Federal Trade Commission Exploring Privacy Roundtable Series, available at: http://www.ftc.gov/os/comments/privacyroundtable/544506-00009.pdf.

McDonald, A. M., Cranor, L. F. (2009), "An empirical study of how people perceive online behavioral advertising", available at: http://www.cylab.cmu.edu/files/pdfs/tech_reports/CMUCyLab09015.pdf.

Network Advertising Initiative (2009), "Opt out of NAI member ad networks", available at: http://networkadvertising.org/managing/opt_out.asp.

Nguyen, D. H., Kobsa, A., Hayes, G. R. (2008), "An empirical investigation of concerns of everyday tracking and recording technologies", in *UbiComp '08: Proceedings of the 10th international conference on Ubiquitous computing*, ACM, New York, NY, USA, pp. 182–191.

Nolet, M. (2008), "Using your browser URL history to estimate gender", available at: http://www.mikeonads.com/2008/07/13/using-your-browser-url-history-estimate-gender/.

Norberg, P. A., Horne, D. R., Horne, D. A. (2007), "The privacy paradox: Personal information disclosure intentions versus behaviors", *Journal of Consumer Affairs*, Vol. 41, No. 1, pp. 100–126.

Paine, C., Reips, U.-D., Stieger, S., Joinson, A., Buchanan, T. (2007), "Internet users' perceptions of 'privacy concerns' and 'privacy actions"', *International Journal of Human-Computer Studies* Vol. 65, No. 6, pp. 526–536.

Quantcast (2009), available at: http://www.quantcast.com/.

Raskin, A. (2009), "Vote! how to detect the social sites your visitors use", available at: http://www.azarask.in/blog/post/socialhistoryjs/.

Samuel, J., Zhang, B. (2009), "RequestPolicy: Increasing Web browsing privacy through control of cross-site requests", in *Proceedings of the 9th Privacy Enhancing Technologies Symposium*.

Sheehan, K. B. (1999), "An investigation of gender differences in on-line privacy concerns and resultant behaviors", *Journal of Interactive Marketing*, Vol. 13, No. 4, pp. 24–38.

Targeted Advertising Cookie Opt-Out (2009), available at: http://taco.dubfire.net/.

TRUSTe (2009), "2009 study: Consumer attitudes about behavioral targeting", available at: http://www.truste.com/pdf/TRUSTe_TNS_2009_BT_Study_Summary.pdf.

Turow, J., King, J., Hoofnagle, C. J., Bleakley, A., Hennessey, M. (2009), "Americans reject tailored advertising and three activities to enable it", available at: http://ssrn.com/abstract=1478214.

Zukowski, T., Brown, I. (2007), "Examining the influence of demographic factors on Internet users' information privacy concerns", in *SAICSIT '07: Proceedings of the 2007 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries*, ACM, New York, NY, USA, pp. 197–204.