Measuring the Effectiveness of Privacy Tools for Limiting Behavioral Advertising

Rebecca Balebako, Pedro G. Leon, Richard Shay, Blase Ur, Yang Wang, and Lorrie Faith Cranor
Carnegie Mellon University
{balebako,pedrogln,rshay,bur,yangwan1,lorrie}@cmu.edu

***Abstract*—Online Behavioral Advertising (OBA) is the practice of tailoring ads based on an individual's activities online. Users have expressed privacy concerns regarding this practice, and both the advertising industry and third parties offer tools for users to control the OBA they receive. We provide the first systematic method for evaluating the effectiveness of these tools in limiting OBA. We first present a methodology for measuring behavioral targeting based on web history, which we support with a case study showing that some text ads are currently being tailored based on browsing history. We then present a methodology for evaluating the effectiveness of tools, regardless of how they are implmented, for limiting OBA. Using this methodology, we show differences in the effectiveness of six tools at limiting text-based behavioral ads by Google. These tools include opt-out webpages, browser Do Not Track (DNT) headers, and tools that block blacklisted domains. Although both opt-out cookies and blocking tools were effective at limiting OBA in our limited case study, the DNT headers that are being used by millions of Firefox users were not effective. We detail our methodology and discuss how it can be extended to measure OBA beyond our case study.**

***Keywords*-behavioral advertising, tracking, privacy tools, do not track, third-party cookies**

I. Introduction

Online Behavioral Advertising (OBA) is the practice of tailoring internet advertising based on an individual's online history and behavior. This work is concerned with third-party behavioral advertising, in which a third-party ad company tracks an individual's web usage history across multiple sites in order to target advertisements. In the United States, third-party OBA is governed through advertising industry self-regulation, overseen by industry groups.

Prior work studying the privacy attitudes of Americans has identified substantial privacy concerns among consumers about tracking and OBA [1], [2]. In response to these concerns, industry groups have provided opt-out webpages on which consumers can state a preference for not receiving behaviorally targeted advertisements. Furthermore, several third-party privacy-enhancing tools exist, often in the form of web browser plug-ins or privacy features built into web browsers.

Although OBA arouses privacy concern in some users, advertising industry self-regulatory groups have argued that the availability of both industry-provided and third-party tools sufficiently enables privacy-concerned consumers to limit the behaviorally targeted advertising they see [3]. However, it has not been clear how to measure the effectiveness of these tools in reducing behavioral advertising. Simple visual inspection does not necessarily reveal behavioral targeting.

We take the first steps towards measuring the effectiveness of tools by presenting novel methods for measuring behavioral targeting in text advertisements based on a user's history of websites visited. We further propose a method for comparing the effectiveness of privacy-enhancing tools based on their ability to limit behavioral targeting in advertisements. As a case study, we measure behavioral targeting in text ads based on web browsing histories organized around different topics, detecting behavioral targeting in Google's text ads for most of these topics. As a further case study, we then test six representative privacy-enhancing tools, measuring behavioral targeting. We find that Firefox's Do Not Track feature does not limit behavioral advertising even though Do Not Track has been enabled by over 5% of Firefox users for whom this feature is available [4].

This work offers two main contributions. First, we introduce a method for measuring behavioral targeting in text ads based on browsing history. Second, we present a method to measure the effectiveness of privacy tools at limiting behavioral advertising. We present results of a case study that uses these methods.

We begin by explaining our motivation and discussing related work in Section II. In Sections III and IV, we present the data collection and analysis portions of our proposed method. In Section V, we describe our experimental method, including descriptions of the privacy-enhancing tools we test. In Section VI, we present our results, which we discuss in Section VII.

II. Background and Related Work

A. How Behavioral Advertising is Operationalized

Online advertising is a lucrative field, estimated at nearly \$15 billion in revenue for the first half of 2011 [5]. Through OBA, advertising agencies tailor ads to specific consumers who may have directly or indirectly indicated interest in specific products [6].

While we cannot gain complete knowledge of the algorithms companies use for data collection since agencies generally do not publish their exact methods, some companies have indicated that the websites a user visits influence the topics of advertisements shown to that user.[1] For example, a user that has visited websites about traveling in Europe may see more ads for travel or European vacations. This method involves

[1] Yahoo: http://info.yahoo.com/privacy/us/yahoo/adinfo.html

tracking the user across different sites. We also know that web users are profiled and put in demographic categories as they traverse websites.[2] Since the websites a consumer visits may lead to him or her being profiled as a member of a certain demographic, leading to seemingly unrelated advertisements targeted to that demographic, it is not always clear from visual inspection whether or not an ad is behavioral.

The proportion of advertising that is behavioral is unclear. Furthermore, behavioral advertising's importance to the economy [7] and the ethics of this practice [8] have both been debated. Advertising agencies say OBA improves market transactions [9], but some scholars argue that privacy concerns can create a "chilling effect" on Internet commerce [1].

Third-party cookies are often used for tracking [10]. Cookies are small pieces of text stored on the user's computer, while third-party cookies are those that are set by any domain other than the primary website visited. For instance, the companies that serve advertisements on a webpage are often a "third party" in the relationship between a website and the visitor since they are not the visitor's primary destination. However, tracking can also occur using Flash Locally Shared Objects (Flash LSOs) and the browser cache [10], [11]. In this paper, we assume that most tracking is done through cookies, based upon statements by the Network Advertising Initiative, an advertising industry trade group.[3]

B. Expectations for online privacy and tracking

Prior work on the privacy attitudes of Americans has found high levels of privacy concern about OBA. In 2009, Turow et al. found that 66% of Americans reject the idea of tailored advertising [2], while McDonald and Cranor found in 2010 that 64% found the idea invasive [1]. Consumers are concerned about having their actions tracked, and they don't necessarily comprehend the mechanisms used to track their actions online; while many web users have heard of "cookies," they don't always understand how they work or how they enable tracking [1]. Furthermore, much of the data collection and monitoring occurs without consumer knowledge [2], [9].

Users who do not wish to receive behavioral advertising do not always know how to protect their online privacy. Leon et al. conducted a 45-person usability study of nine privacy-enhancing tools related to OBA, finding significant usability problems in all tools tested [12]. These usability problems led some participants to believe that all OBA was blocked when the tools were not configured to block anything. In contrast, we examine whether tools that are properly configured are effective at limiting behavioral advertising.

C. OBA self-regulation in the United States

In the United States, OBA is self-regulated by the advertising industry. In response to a decade of scrutiny by the U.S. Federal Trade Commission (FTC) [9], [13], the advertising industry has formed self-regulatory coalitions, such as the Network Advertising Initiative (NAI) and Digital Advertising Alliance (DAA) [14]. These organizations offer websites on which consumers can opt out of tailored advertising. Based on the availability of these opt-out websites, as well as third-party privacy tools, industry leaders have argued that industry self-regulation is working effectively [3]. In contrast, Komanduri et al. investigated members of industry coalitions, finding many cases of non-compliance with self-regulatory principles [15]. Recently, the White House released a Privacy Bill of Rights concurrently with an FTC announcement of the ad industry's intention to support a Do Not Track button on browsers that would reduce OBA [16], [17].

While novel proposals for privacy preserving systems that allow OBA have been described in the literature, they have not been widely adopted. These proposed systems include a privacy-protective method for behavioral advertising that profiles users within the browser, rather than on a third-party server [18], as well as a system that allows users to manage the sharing of third-party cookies based on a trade-off between privacy costs and the benefits of ad relevance [19].

D. Measuring Behavioral Advertising

There has been limited prior work on measuring online behavioral advertising and ad targeting. Guha et al. developed a method for measuring OBA by examining differences in Google ads based on location and search history, as well as differences in Facebook ads shown to users with different profiles [20]. They found that location impacted the ads shown on Google, but the results for search history on Google had a less clear impact. We expand upon their work by measuring how a history of web pages visited leads to behavioral advertising, and to compare the effectiveness of privacy tools.

There has also been prior work on measuring tracking. For instance, work by Soltani et al. examined and counted Flash LSOs, which allow unique user tracking but can't be removed in the same way that typical cookies are deleted [11]. McDonald and Cranor did a follow-up study a year later and found that Flash LSOs were still being used [21]. A long-term study by [22] measured privacy diffusion through increasing aggregation of data by third-party agencies.

III. Data Collection Method

Collecting data to measure behavioral targeting is a complex process, on account of confouding factors such as IP address, browser fingerprints, and LSOs. It is also important to ensure that tests are run at the same time so that the influence of ad turnover is minimized.

In this section, we introduce our method for collecting data for the purpose of detecting behavioral targeting in text-based ads. We then present our methods for analyzing this data in Section IV. At a high level, we configure browser automation software to create a history of web browsing on a particular topic. The browser then visits a general interest site and captures the text advertisements on that page for analysis. This process is repeated a number of times, creating a set of advertisements for that topic. For each topic, we then compare

[2] Google Ad Preferences, Interest-based Advertising: How it Works, http://www.google.com/ads/preferences/html/about.html

[3] http://www.networkadvertising.org/managing/faqs.asp

these sets of advertisements with those advertisements presented in the general interest site with no previous browsing history. Similarly, we compare ads when different tools are in use with those obtained when no tool is in use.

A. Training and Testing

To induce one way behavior is tracked for OBA, we visit several websites centered on a specific topic. We dub this process "training" as it could train an advertiser's ad selection algorithm to indicate that the consumer is interested in this specific topic. We visit between five and ten unique domains for each topic, which we dub "training sites," usually visiting both the homepage and an article on that site. For each topic, we inspected the top two pages of Google search results in a search for that topic, choosing "training pages" from these results that have at least six third-party trackers. We iterated over several sets of training and test pages trying to find sets that were likely to reveal OBA.

After visiting each set of "training pages" for a topic, we visit a "test site." A "test site" is a general interest page, such as a news site, that displays text ads. These sites are chosen to appeal to a broad audience so that it is less likely that advertisements are contextually chosen.

Sites generally rotate through more ads than can be shown on a single visit to a test page. Guha et al. identified that over 80% of the unique ads in their experiment had been loaded by the 7th visit, but ad turnover takes over after around 10 ads [20]. We repeat the training process and visit each test site 7 times to grab a representative sample of the ads for that site at that moment.

We also store the cookies set by these sites. To store the cookies, we copy the cookie file from the active Firefox profile into a new location after each page load. Before downloading the ad and copying the cookie files, we wait seven seconds to allow the page to load completely.

B. Data Collection Algorithm

First, we collect data to measure the presence of OBA without any tool installed; then we collect data while the tools are in place to evaluate their effectiveness at limiting OBA. To measure behavioral advertising, all topics are run simultaneously on identical virtual machines. Therefore, the results from each topic can be compared individually to the results from running without any training topic (*no topic*).

The data-collection method for the privacy-enhancing tools differs from OBA data-colletion and testing in terms of what is run simultaneously (all tools for each topic) and what the control is (no tool). The tools are tested simultaneously on identical virtual machines running the algorithm described in Algorithm 1. We measure whether the ads have changed by comparing the set of ads run with each tool to the set run with *no tool* active, as we expect effective tools to yield different sets of ads than running without a tool.

All data collection is automated using Java and Selenium2.[4]

Algorithm 1: Algorithm Used to Test Tools. The algorithm used to find OBA is the same, except the initial **for** loop over topics is not run, as all topics are run simultaneously.

```
for each topic in (notraining, wedding, travel, camera,
bicycle) do
    for each testpage in (nytimes, chicagotribune,
    latimes, howstuffworks) do
        repeat
            Open browser
            Visit all training pages on topic unless topic
            is notraining
            Visit and save testpage
            Close browser, delete cookies
        until 7 visits;
    end
end
```

C. Controlling for Identical Environments

To ensure that we have a clean browser history between all tests, we delete the cookies, delete the cache of Flash LSOs, and close the browser after each visit to the test website. Additionally, each time Firefox starts, we use a clean copy of the user profile directory.

All simultaneous tests are run on identical virtual machines (VM) with clean installation of Windows 7, Firefox, and the necessary add-ons for the privacy-enhancing tools. The identical VMs prevent unique browser "fingerprints" from being created, which might enable tracking [23]. The use of separate VMs ensures that the different cookies, temporary files, or any other artifacts from browsing do not interfere with each other in separate tests. At the same time, it allows the tests to be run from identical environments.

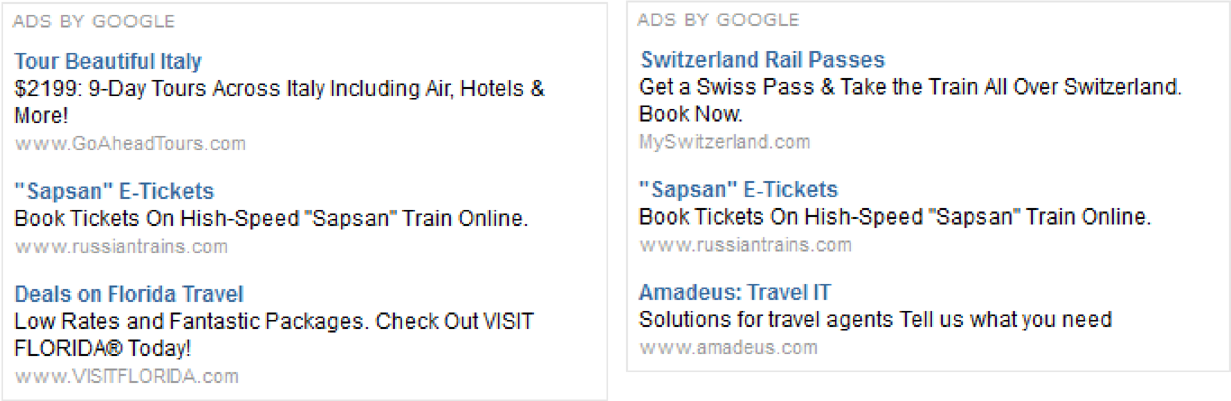
The virtual machines all use a proxy to access the Internet and appear to come from the same IP address to outside networks since Guha et al. found that different IP addresses yield different ads [20]. While IP addresses could be used by advertising agencies to track users, we do not observe this behavior in our results. We do not see leakage of specific words from the initial topics appearing in topics that were run later, which may happen if tracking tactics besides cookies were used.

Some hits to websites may take longer than others to load. To ensure that all tests are synchronized, we use a server to control the start of each test across all VMs. In our case study, most sets finished within 60 seconds of each other, although we did observe up to three minutes of difference in the worst case.

IV. Analysis Methods

We first describe the anatomy of an advertisement, highlighting the information we used to detect behavioral advertising. We then describe how we use cosine similarity to compare sets of ads using either the URLs displayed on those ads or by considering all words present in each set of ads.

[4] SeleniumHQ: Web Application Testing System, http://seleniumhq.org/



ADS BY GOOGLE

Tour Beautiful Italy
$2199: 9-Day Tours Across Italy Including Air, Hotels & More!
www.GoAheadTours.com

"Sapsan" E-Tickets
Book Tickets On Hish-Speed "Sapsan" Train Online.
www.russiantrains.com

Deals on Florida Travel
Low Rates and Fantastic Packages. Check Out VISIT FLORIDA® Today!
www.VISITFLORIDA.com

ADS BY GOOGLE

Switzerland Rail Passes
Get a Swiss Pass & Take the Train All Over Switzerland. Book Now.
MySwitzerland.com

"Sapsan" E-Tickets
Book Tickets On Hish-Speed "Sapsan" Train Online.
www.russiantrains.com

Amadeus: Travel IT
Solutions for travel agents Tell us what you need
www.amadeus.com

Fig. 1. Ads on two different visits to the Chicago Tribune after visiting web sites about European travel.

TABLE I
ANATOMY OF AN EXAMPLE ADVERTISEMENT FROM FIGURE 1

Display URL:	www.GoAheadTours.com
Title:	Tour Beautiful Italy
Content:	$2199 9-Day Tours Across Italy Including Air Hotels More
Source URL:	http://i.goaheadtours.com/gc/italy%3Fmkwid %3DcrCtiOElZ%26pcrid%3D6438802697%26 utm%3Dgoogle%26utm_campaign%3DGAT_Italy-Content%26utm_term%3Ditaly%2520travel%2520tour utm_medium%3Dcpc

A. Anatomy of an ad

In Figure 1, we show examples of the text advertisements from loading www.chicagotribune.com on two occasions after visiting several websites related to European Travel. Each ad box contains up to three different ads. Each ad includes a *title* (such as "Tour Beautiful Italy"), descriptive text that we term the *content* of the ad, and what we term a *display URL* (e.g. www.GoAheadTours.com), which is the URL displayed on the ad itself. In the source code, additional information is available; we define the *source URL* as the URL in the web page's source code to which the advertisement redirects. This URL generally differs from the *display URL* by including an exact destination and many parameters, whereas the *display URL* is often the domain of the destination. Table I provides an example of all information we considered from each advertisement.

B. Comparison using display URL

In our first of two comparisons, we use the display URL of ads to compare "sets" of advertisements. This same method has previously been used by Guha et al. in determining whether advertisements were customized based on location or search history [20]. A "set of ads" comprises the ads collected while visiting all test pages after training on a specific topic while using a particular privacy tool (or no tool).

C. Comparison using all words

In our second comparison, we combine the titles and full content of all ads in each set, resulting in a set of all words in the ads. To combine words with the same root, such as "cycling" and "cyclist," we use an implementation[5] of the Porter stemming algorithm [24]. In our case study, stemming reduced the number of unique words by about 12%.

D. Comparison using Cosine Similarity

To perform both the comparison between sets of advertisements' display URLs and the comparison of all words in a set of ads, we use cosine similarity. The cosine similarity of two vectors measures the similarity between these two vectors, yielding a value from 0 (completely different) to 1 (exactly the same). Thus, two sets of ads with cosine similarity near 1 contained roughly the same advertisements.

In the first comparison using display URLs, each vector contained the frequencies of the display URLs observed in a particular set of ads. In the second comparison, each vector contained the frequencies of each word appearing in either the *title* or *content* of ads.

Cosine similarity is defined as:

$$\frac{\bar{\mathrm{A}}\cdot\bar{\mathrm{B}}}{||\bar{\mathrm{A}}||\;||\bar{\mathrm{B}}||}, \qquad \bar{\mathrm{A}} = [w_{A,e}] \qquad (1)$$

A and B are the frequency vectors of elements in the union of both sets. Each element e is, respectively in our two comparisons, either a display URL or a word found in the ads. $[w_{A,e}]$ is the weight of element e in vector A, which we calculate as the number of times that element appears in the set.

V. CASE STUDY METHODOLOGY

In our case study of text ads, we use five different topics for *training*, test on five different *test pages* for the presence of behavioral targeting, and then test the effectiveness of six privacy-enhancing tools. All tests were run on virtual machines with the Windows 7 operating system and Firefox 7.0.1 browser. All data were collected during November 24-26, 2011.

[5] Porter Stemmer http://tartarus.org/~martin/PorterStemmer/

A. Training Topics

In order to examine whether a history of browsing on these topics led to behaviorally targeted ads, we used a clean install of a virtual machine to visit a handful of webpages focused on a particular topic. For the purposes of our case study, we chose topics which were likely candidates for OBA, as they involve expensive items or services that might have a specific audience. These topics were chosen from a list of a dozen topics collaboratively developed by the researchers and are not representative of all advertising topics. For each training topic, we chose up to ten sites from the first two pages of Google results for that topic as *training pages*.

Based on the number and variety of third-party advertisers we observed in pilot studies on training pages for our original list of topics, we chose to run the case study with the topics "wedding," "European travel," "digital camera," "bicycling," and "pregnancy." The final topic, "pregnancy," was chosen particularly since health topics tend to be more privacy-sensitive.

B. Test Sites

To test whether the *training* had led to behaviorally targeted advertisements, we examined the ads on *test sites* that we chose. For the case study, we chose sites with a large regional or national audience that contained several text advertisements. These test pages were: http://www.nytimes.com, http://www.cnn.com, http://www.chicagotribune.com/news/local/breaking/, http://www.latimes.com, and http://entertainment.howstuffworks.com/.

C. Privacy-Enhancing Tools

In our case study, we tested six privacy-enhancing tools. These tools can be grouped into three broad categories: opt-out cookies, "blocking" tools, and Do Not Track.

D. Opt-out Cookies

Opt-out cookies allow users to specify their desire to "opt out" of behavioral advertising, storing this request in a cookie on their computer. Opt-out cookies can be set and read by each individual ad agency. If a particular ad agency's opt-out cookie exists on a user's computer, the ad agency is notified that the user does not wish to receive behavioral advertising. Although the opt-out process is a core part of the self-regulation of OBA, it has not been widely adopted by users [25].

The Digital Advertising Alliance (DAA) provides a web site allowing consumers to "opt-out" of advertising by the agencies in their alliance.[6] The Network Advertising Industry (NAI) has a similar page that allows users to opt-out of the individual agency's behavioral advertising.[7] We tested both of these opt-out mechanisms by opting out of all agencies listed on each page at the time of our data collection.

E. Blocking Tools

"Blocking" tools prevent tracking and third-party advertising by refusing content (such as cookies or scripts) from specific domains on a blacklist. We tested three "blocking" tools; one of these is built into browsers, while the other two are browser add-ons.

Setting a web browser to block all third-party cookies may prevent tracking since advertising agencies commonly use third-party cookies to store information about the user's web history. Estimates of the percentage of Internet users blocking third-party cookies range from 5-18%.[8] In this work, we test the option built into Firefox to block all third-party cookies.

The second and third "blocking" tools we tested were browser plugins that protect against tracking and behavioral advertising by blocking blacklisted third-party content. Typically, these plugins' blacklists include the domains of known advertising agencies and trackers. In this work, we tested TACO 4.40, developed by Abine, with opting out of targeted ad networks and blocking web trackers enabled.[9] We also tested Ghostery 2.6.2 with blocking and cookie protection enabled.[10] Ghostery is developed by Evidon and has over 300,000 Firefox users.[11]

1) Do Not Track: One proposal for allowing users to control tracking is the "Do Not Track" (DNT) header. A user's web browser sends an HTTP header notifying websites that the user does not want to be tracked. The W3C has released a draft of a technical standard for these headers [26], [27]. While the definition of "Do Not Track" is not universally agreed upon, Mozilla, Microsoft, and Apple have already implemented DNT headers in their browsers [28]. Around 6 million Firefox users have enabled DNT on their browsers.[12] An informal survey done by privacychoice.org of users with DNT enabled found that three-quarters believed that some or all ad agencies would honor their request [29]. In a survey of web users, McDonald et al. found that 79% expected a Do Not Track button would limit data collection, and 26% thought it would limit cookies [30]. At the time of our study, only two out of hundreds of existing advertising agencies had agreed to respect Do Not Track headers [31]. We tested the Do Not Track option built into Firefox 7.0.

F. Case Study Limitations

In the first part of the case study, in which we measured behavioral advertising, not all topics and test pages had measurable behavioral advertising. We only test the privacy-enhancing tools on those topics and test pages on which we found behavioral advertising. Furthermore, all text ads on these test pages for which there was measurable behavioral advertising were served by Google. Therefore, the results from testing the tools is limited to Google ads.

[6] http://www.aboutads.info/choices

[7] http://www.networkadvertising.org/managing/opt_out.asp

[8] Fighting Intenet Surveillence http://www.grc.com/cookies/cookies.htm

[9] http://abine.com/preview/taco.php

[10] http://www.ghostery.com

[11] https://addons.mozilla.org/en-US/firefox/addon/ghostery/

[12] http://blog.mozilla.com/privacy/author/afowlermozilla-com/

Our topics were limited to a very small subset of areas. We do not claim they are representative; instead they are used to test our methodology and provide initial results. Furthermore, our results represent a snapshot in time. Although pilot studies indicate that our results will hold up, future work could replicate the results over longer time periods, such as every few days for several weeks.

VI. Results of the Case Study

To demonstrate the method we have introduced, we present the results of a case study of Google text ads on a limited number of training topics. Throughout our results, we report cosine similarities for both the comparison of display URLs and the comparison of all words contained in the ad. The results for display URLs and words are similar.

A. Baseline Measurements for Similarity

Agencies may have a large set of ads available for display, and a particular test will capture only a subset of these ads. Furthermore, due to churn, the set of ads available changes over time. To account for both artifacts of seeing only a subset of the ads available at a particular moment, as well as ad churn, we measured the cosine similarity of sets of ads created under identical conditions as a baseline measurement.

To compute this baseline, we compared the results of identical tests without any privacy-enhancing tools. We controlled the tests to ensure the time of visit, operating system, browser, and sites visited were identical. The difference between these two no-training, no-tool controls provides a baseline measurement for similarity at a particular moment. The cosine similarity between the two sets across all test pages was .97 for word frequency and .97 for URL frequency. Running the same test again the next day, we found .97 for words and .95 for URL frequency.

Since these sets that were designed to be the same had a cosine similarity of at least .95, the remaining .05 can be attributed to subset selection and churn. To be conservative, we broaden the margin for our baseline measurement to be .10. Therefore, we assume that a cosine similarity between two sets of .90 or above indicates that the sets are from the same distribution of ads.

B. Measuring Behavioral Advertising

In order to verify that the tools stop behavioral advertising, we need first to confirm that our set produced behavioral ads based on the training topics. We compared the ads that resulted from each training topic to the ads from an untrained machine (no-training, no-tool).

1) Behavioral Advertising by Topic: We compared the results both by topic and by test page to see which areas showed the most behavioral advertising. Figure 2 presents the results of comparing each topic to the no-tool, no-training set across all test pages. Most training topics yielded different ads, both in terms of unique URLs and the words found. Digital camera, European travel, bicycling, and wedding all have cosine similarities between .27 and .50 for display URLs. This indicates that the ads were different than the ads from no-training. The cosine similarity between these topics and no-training ranged from .35 to .60 for word frequency.

The results from training on pregnancy did not show the same level of behavioral targeting. The cosine similarity of display URL frequency between pregnancy and no-training is .90, and it is .93 for the set of words. Based on our baseline measurements, this level of cosine similarity does not seem to indicate behavioral targeting. Therefore, either our training on pregnancy was inadequate, there were no ads for pregnancy on the days we ran the test, or Google does not use visits to pregnancy sites to build an advertising profile.

Further examination of the words used in the ads shows evidence of behavioral advertising. Some of the most frequently seen words after training are semantically related to the training topic. Table II shows the most frequent words for each topic; many of these words appeared only for that topic. Table III shows the most frequent words that appear only for that topic. For example, words with the stem "wed" and "favor" were found frequently after having visited wedding pages, but at no other times.

TABLE II
FIVE MOST FREQUENT WORDS, BY TOPIC, IN ORDER OF FREQUENCY

topic	frequent words
travel	on, eurail, pass, sapson, to
wedding	free, for, wed, label, your
camera	camera, free, sale, ship, for
bicycle	bike, mountain, and, you, for
pregnancy	depress, for, symptom, free, have
no training	depress, for, symptom, a, now
no training 2	depress, for, symptom, now, new

TABLE III
WORDS THAT APPEARED MORE THAN 15 TIMES FOR A TOPIC, BUT ONLY FOR THAT TOPIC. PREGNANCY AND NO TRAINING HAD NO SUCH WORDS. STEMMED WORDS ARE APPENDED TO SHOW THE ORIGINAL WORD.

topic	unique words
travel	eurail, pass, sapsan, ticket, train, europ[e], rail, e-ticket, acp, american
wedding	wed, label, candy, favor
camera	camera, digital, leica, olympus, canon
bicycle	bike, mountain, cycl[e], appalachian, near

On the other hand, pregnancy training yielded ads with the same health-related words as no-training, such as "depress" and "symptom," although these words did appear more frequently. This gives further credence to our observation that we either did not train sufficiently on the topic, or the behavioral advertising is too subtle to measure. We do not include pregnancy in further analysis.

2) Behavioral Advertising by Test Page: Figure 3 shows the amount of behavioral advertising by test page and topic. Cosine similarities that are close to .9 indicate that the training made very little difference. The text ads on cnn.com do not show measurable behavioral advertising. We therefore do not include cnn.com in further analysis. All remaining text ads were from Google.

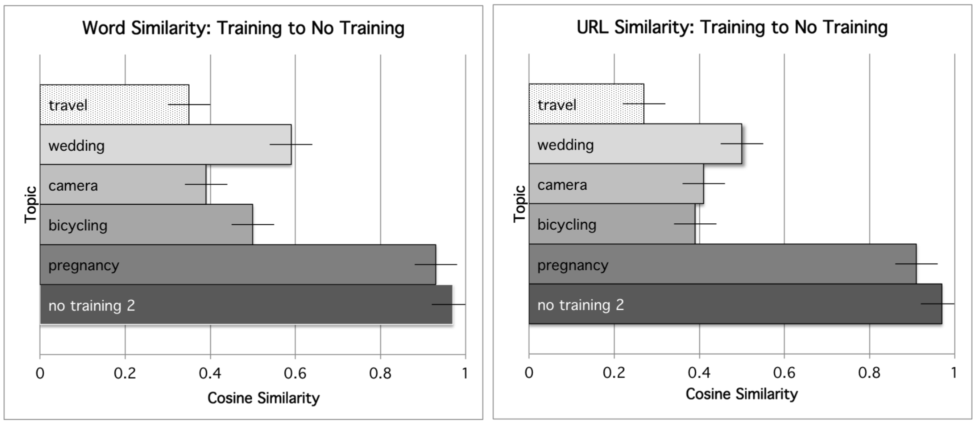


Fig. 2. Results from training on each topic compared to "no training." Results from all test pages are combined. Topics with cosine similarity less than .9 show evidence of behavioral advertising. Smaller similarity indicates more OBA. "No training 2" is a second set with no training, used to determine our baseline measurements. Error bars represent the .10 difference that can be attributed to ad selection and churn.

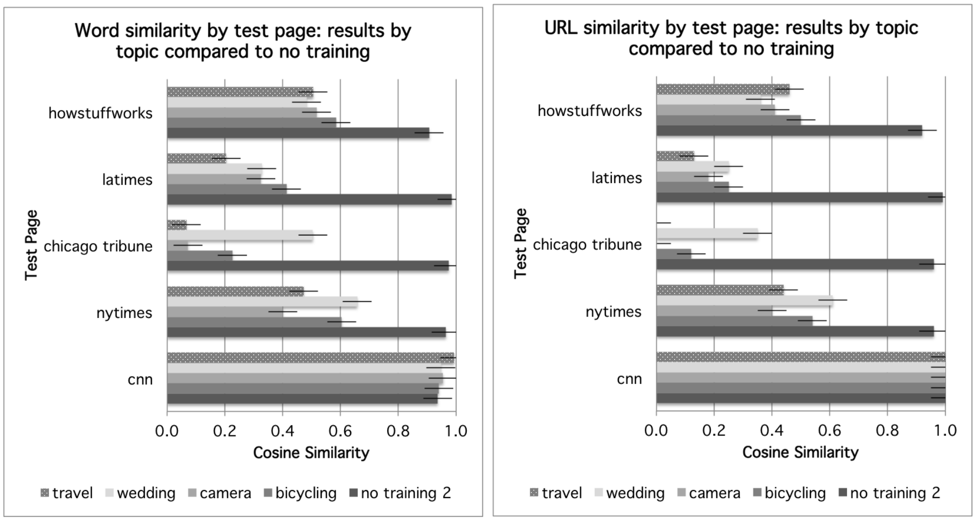


Fig. 3. Ads from all test pages and training topics compared to ads from no-training. We consider similarities below .9 to be evidence of behavioral advertising. No-training 2 is an identical test to no-training, so cosine similarity is expected to by high.

C. *Comparing Tools*

After confirming that we could measure behavioral advertising, we used a similar method to test the tools across the four topics and four test pages. Our measure of effectiveness for tools is the similarity between ads from the tool and training to ads from no-tool and the same training. We expect the tools to have an impact, reducing the number of behavioral ads, and therefore to be different than the no-tool control set. We find the cosine similarity between the tool when trained on a topic, and no-tool trained on the same topic. For these comparisons, the smaller the cosine similarity, the more effective the tool is for this subset of topics and advertisers.

1) Number of Ads and Words: We collected the ads from visiting the test pages seven times for each topic. There were 215 unique ads (based on display URL) across all tools, test pages, and topics. Only one of these URLs was also a training page.

For all tools and topics and test pages, the average number

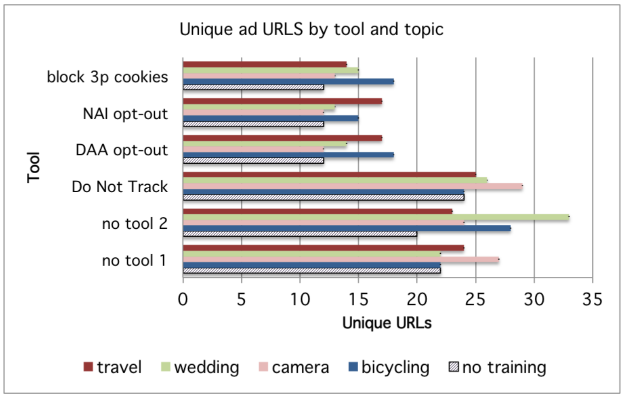


Fig. 4. The number of display URLS (used to measure unique ads) for the tools across all test pages. Abine Taco and Ghostery eliminated Google text ads and are thus not shown.

of total ads was 19.45 ($\sigma = 4$) for the seven hits to that page. The mode number of total ads was 21 as each page typically had 3 ads. Not including Ghostery and Abine Taco, there was very little variance ($\sigma^2 = 0.10$) between the tools for the average number of ads shown in a set.

Ghostery and Abine Taco both completely eliminated all Google text ads that we examined in this study, although they did not remove all advertising. The following graphs and discussions thus exclude Ghostery and Abine Taco.

As seen in Figure 4, the DAA, NAI, and Firefox third-party cookie blocking tools all reduced the number of unique ads (as measured by number of unique display URLS). This occurred whether or not there was training on a topic. This implies that the tools reduce the set of ads shown, even without training and no history for OBA.

2) Words by Topic: To see if the tools reduced ads with topical words, we examined the two most frequent words that were found only in each training topic on no-tool, no-training. These words are indicative of behavioral advertising. As seen in Table IV, the opt-out cookies and Firefox third-party cookie blocking tools did not have ads with these words.

3) Cosine Similarity: Tool to No-Tool: Figure 5 shows the results from this comparison across all test pages, using both word frequency and URL frequency. Firefox third-party cookie blocking, NAI, and DAA are different than no-tool for each of the training topics. Do Not Track is less effective. While DNT is not as similar as the two no-tool controls are to each other, the similarities are much higher than the other tools.

D. Examination of Cookies

In order to improve understanding of how the tools operate, we examined their impact on cookies. First, we examined the number of cookies set by the opt-out tools and when they were set. Second, we looked at the cookies from new domains that were set during training. This observation cannot determine which cookies are used for tracking, or how much tracking is being done by each ad agency.

The tools DAA, NAI, and Abine Taco set opt-out cookies, which indicate that a user does not want OBA. DAA set 136 cookies from 96 unique domains when we opted-out, while NAI set 120 cookies from 91 unique domains. For both DAA and NAI, 66% of domains had a cookie that appeared to be for opt-out purposes, which meant they included the word "privacy," set a unique ID to zeros, or matched the case-insensive regular expression opt.*out. Abine Taco sets a larger list of opt-out cookies: 191 cookies from 150 unique domains, 69% of which appear to be opt-outs.

Figure 6 shows how many new domains set cookies after training on each topic. For the opt-out tools, this includes only new domains that set cookies after the opt-out cookies were already set. The top line shows how many unique domains were visited in the test set, and thus the maximum number of first-party cookies that could be expected.

DNT seems to have little impact on the number of cookies, which aligns with the industry's low adoption of DNT. In contrast, the blocking tools block the most cookies. Opt-out reduces the number of new cookie domains. By opting-out, a large number of cookies will already be set on the machine, reducing the number of new domains that can set cookies. Many of the domains that were able to set cookies are not members of DAA or NAI. This points to a limitation in the opt-out method; an individual can only opt-out from companies that choose to participate and respect the self-regulation standards. Furthermore, many of these companies only promise to stop delivering targeted ads and make no statements about reducing tracking.

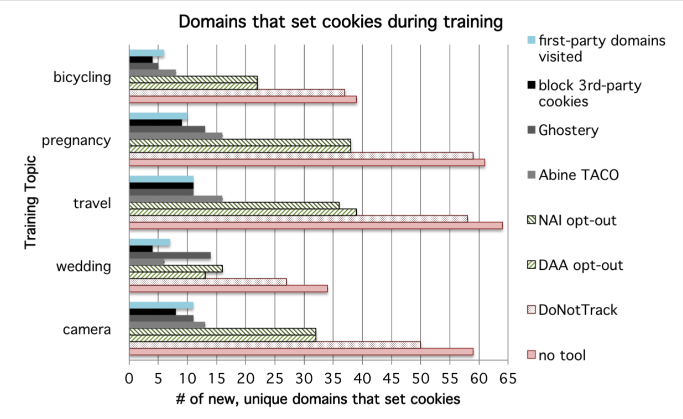


Fig. 6. Number of unique domains that set cookies during training, organized by tool and topic. For the opt-out tools, this includes only new domains that were set after the user had opted-out. The first-party domains indicate the number of unique domains visited during training, hence the maximum expected number of first-party cookies.

VII. Discussion

A. Online Tracking and OBA

We provide a method of determining whether advertising is behaviorally targeted. However, tracking tends to be the cause for privacy concern, and seeing behavioral advertising is not equivalent to "seeing" tracking. While behaviorally targeted advertisements imply that tracking has occurred, a lack of targeted advertisements does not necessarily imply that a consumer is not being tracked. Although five of the six tools we tested in our case study were effective at limiting OBA, we cannot evaluate the level of protection they offer against online tracking. In fact, the DAA and NAI trade groups do not guarantee that their members' opt-out cookies will limit data collection or tracking, only that they will prevent the consumer from seeing behaviorally targeted ads from member agencies.

B. Sensitive topics

We were unable to measure behavioral ads when we trained on the topic of pregnancy, which was the most privacy-sensitive topic we tested. Google notes that it does not use health information in behavioral advertising [32], but this result might not generalize to other ad agencies. For example, the New York Times reported on Target's efforts to identify

TABLE IV

THE FREQUENCY OF SELECT WORDS THAT INDICATE BEHAVIORAL ADVERTISING, ACROSS TOOLS AND TOPICS. THE TOOLS DAA, NAI, AND FIREFOX THIRD-PARTY COOKIE BLOCKING HAD NO ADS WITH THESE WORDS. *THE WORDS SELECTED FOR NO-TRAINING ARE NOT INDICATIVE OF BEHAVIORAL ADVERTISING, AND ARE JUST THE MOST FREQUENT WORDS.

training topic	word	no tool 1	no tool 2	DNT	DAA	NAI	block 3p cookies
travel europe	eurail	28	24	26	0	0	0
	pass	22	22	34	0	0	0
wedding	wed	54	43	48	0	0	0
	label	41	36	34	0	0	0
digital camera	camera	58	60	55	0	0	0
	digit	18	17	15	0	0	0
bicycling	bike	21	26	35	0	0	0
	mountain	8	17	21	0	0	0
no-training*	depress	54	54	59	56	56	56
	symptom	30	25	30	35	35	35

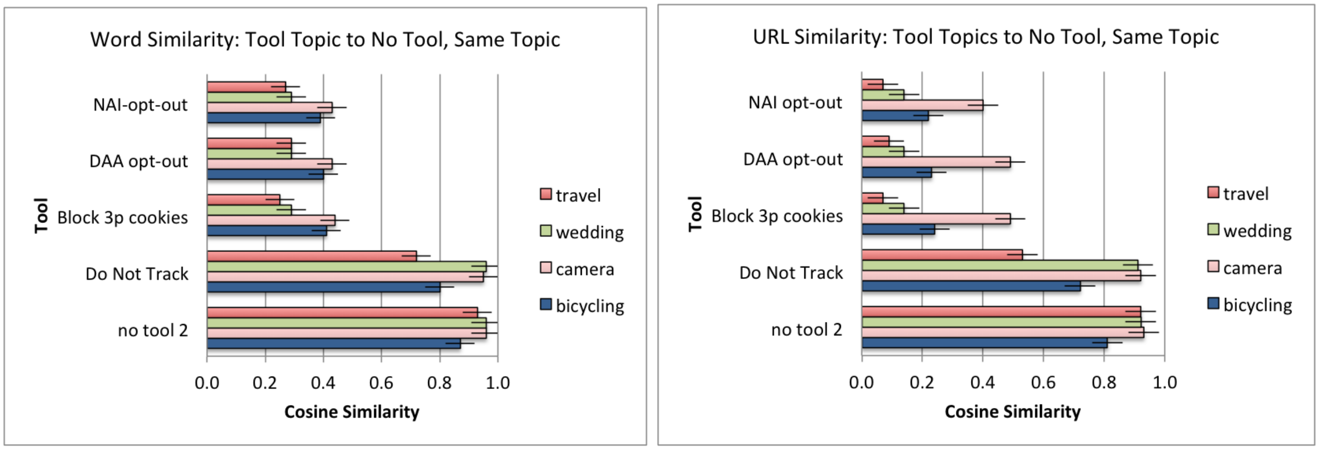


Fig. 5. Results from each tool compared to ads from no-tool on the same topic. Effective tools have low cosine similarities when compared with no-tool as they reduce the number of behavioral ads. Abine Taco and Ghostery eliminated the Google text ads, so they are not shown.

pregnant customers based on their purchases [33]. Furthermore, individuals have different ideas of what is private, and ad agencies are likely not responsive to individual preferences.

Furthermore, all of our tests trained only on a single topic, leading to ads tailored to that same topic. Some profiling may be more subtle by inferring demographics, such as determining that a particular history of web usage fits the profile of a 65 year-old woman. Then, advertisements thought to appeal to that demographic could be shown. Further work could investigate whether particular topics, or combination of topics, lead to these profiles and thus ads targeted towards a demographic rather than an interest.

C. Do Not Track

In our case study, Do Not Track headers did not seem to limit behavioral targeting of ads. Unfortunately, there are currently millions of Firefox users with DNT enabled who might expect it to have some impact. Although a recent announcement indicates that companies including Google will begin supporting Do Not Track later in 2012, [17] the preferences of DNT users are currently being disregarded.

D. Limitations

A major limitation of this work is that Google text ads were the only ads we tested. For a more representative view of OBA moving beyond a case study, our techniques should be extended to ads that are not text-based and that represent a spectrum of advertising agencies. Furthermore, our results represent a snapshot in time. Although pilot studies suggest that our results will hold up, future work could replicate the results over longer time periods.

The data collection process is easily reproducible; we imagine access to multiple virtual machines that can be run simultanreously would be the limiting factor in scaling up.

E. Future work

A major step in understanding the full spectrum of OBA is to expand our measurement techniques from text ads to multi-media ads. Analysis of the text ads was also automated; it remains to be determined whether image and multi-media ads can also be analysed in an automated manner.

Our work looked at both the text content of ads and the display URLs. We found that ad URLs and text yielded similar results for measuring OBA and the tools. However, in image or video ads, the text might not be available. Our work implies that comparison results are not sensitive to the field of the ad used to do the comparison, and that available fields (such as URLs alone) can be used to compare ads.

VIII. Conclusion

In this work, we have made two contributions to the study of privacy and behavioral advertising. First, we have presented a method for measuring behavioral targeting in text ads based on web history. Second, we have developed a method for testing the effectiveness of privacy tools that claim to limit behavioral advertising. Finally, we have demonstrated our method on text ads from Google for four training topics.

The methods developed here lay the groundwork for future testing of tools across different advertising agencies since detecting behavioral advertising based on web history is not a trivial task. We automated the collection of ads in a way that controlled for the time of visit, IP address, machine setup, and deletion of Flash LSOs.

In the case study comparing the effectiveness of tools on Google text ads, we find that the add-ons Ghostery and Abine Taco are effective in preventing behavioral advertising since they remove these ads from the website. The cookie-based tools we tested – blocking third-party cookies, NAI opt-out cookies, and DAA opt-out cookies – are also effective in reducing behavioral advertising. Using these tools, ads are still displayed, but they are similar to the ads shown without training. In contrast, Do Not Track headers were ineffective in reducing behavioral advertising. These differences suggest that further work is needed applying the techniques we propose across a wider range of advertising agencies and across all privacy-enhancing tools available to better capture the extent to which protectively configured tools are effective in limiting behavioral advertising.

Acknowledgment

This research was funded in part by a grant from The Privacy Projects and by NSF grants DGE0903659, CNS1012763, and CNS0831428. The authors would like to thank Michael Strroucken and the CMU Parallel Data Lab.

References

[1] A. McDonald and L. Cranor, "Americans' attitudes about internet behavioral advertising practices," in *Proceedings of the 9th Annual ACM Workshop on Privacy in the Electronic Society (WPES 2010)*, pp. 63–72.

[2] J. Turow, J. King, C. Hoofnagle, A. Bleakley, and M. Hennessy, "Americans reject tailored advertising and three activities that enable it," *Departmental Papers (ASC)*, p. 137, 2009.

[3] Interactive Advertising Bureau, "IAB tells congress privacy bills may harm business and consumers," http://www.iab.net/public_policy/1296039, July 2011.

[4] A. Phadke, "Understanding DNT adoption within firefox," http://blog.mozilla.com/metrics/2011/09/08/understanding-dnt-adoption-within-firefox/, September 2011.

[5] I. A. Bureau, "Press release," http://www.iab.net/about_the_iab/recent_press_releases/press_release_archive/press_release/pr-092811, September 2011.

[6] D. Evans, "The online advertising industry: Economics, evolution, and privacy," *Journal of Economic Perspectives*, 2009.

[7] J. Mayer, "Do not track is no threat to ad-supported businesses," Stanford Law School http://cyberlaw.stanford.edu/node/6592, Jan. 2011.

[8] A. Massey and A. Antón, "Behavioral advertising ethics," *Information Assurance and Security Ethics in Complex Systems: Interdisciplinary Perspectives*, 2010.

[9] Federal Trade Commission, "Online profiling: A report to congress, part 2 recommendations," *Federal Trade Commission, Washington, DC*, 2000.

[10] M. Ayenson, D. J. Wambach, A. Soltani, N. Good, and C. J. Hoofnagle, "Flash cookies and privacy II," *SSRN eLibrary*, 2011. [Online]. Available: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1898390

[11] A. Soltani, S. Canty, Q. Mayo, L. Thomas, and C. J. Hoofnagle, "Flash cookies and privacy," *SSRN eLibrary*, 2009. [Online]. Available: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1446862

[12] P. Leon, B. Ur, R. Balebako, L. Cranor, R. Shay, and Y. Wang, "Why Johnny Can't Opt Out: A usability evaluation of tools to limit online behavioral advertising," *CHI 2012, forthcoming*.

[13] Federal Trade Commission, "FTC staff report: Self-regulatory principles for online behavioral advertising, 2009," *Federal Trade Commission, Washington, DC*, 2009.

[14] D. Hirsch, "The law and policy of online privacy: Regulation, self-regulation or co-regulation?" 2010.

[15] S. Komanduri, R. Shay, G. Norcie, B. Ur, and L. Cranor, "AdChoices? Compliance with online behavioral advertising notice and choice requirements," *I/S: A Journal of Law and Policy for the Information Society.*, 2012, forthcoming.

[16] White House, "Consumer data privacy in a networked world: A framework for protecting privacy and promoting innovation in the global digital economy," *White House, Washington, DC*, February 2012.

[17] J. Angwin, "Web firms adopt 'no track' button," http://online.wsj.com/article/SB10001424052970203960804577239774264364692.html, Feb 2012.

[18] V. Toubiana, A. Narayanan, D. Boneh, H. Nissenbaum, and S. Barocas, "Adnostic: Privacy preserving targeted advertising," in *17th Network and Distributed System Security Symposium*, 2010.

[19] J. Freudiger, N. Vratonjic, and J. Hubaux, "Towards privacy-friendly online advertising," *Proceedings of W2SP*, 2009.

[20] S. Guha, B. Cheng, and P. Francis, "Challenges in measuring online advertising systems," in *Proceedings of the 10th annual Conference on Internet Measurement (IMC)*, 2010, pp. 81–87.

[21] A. McDonald and L. Cranor, "A survey of the use of adobe flash local shared objects to respawn http cookies," http://www.casos.cs.cmu.edu/publications/papers/CMUCyLab11001.pdf, CyLab, Carnegie Mellon University, Tech. Rep., 2011.

[22] B. Krishnamurthy and C. Wills, "Privacy diffusion on the web: A longitudinal perspective," in *Proceedings of the 18th international conference on the World Wide Web (WWW)*, 2009, pp. 541–550.

[23] P. Eckersley, "How unique is your web browser?" panopticlick.eff.org/browser-uniqueness.pdf, Electronic Frontier Foudation, EFF Report, 2009.

[24] M. Porter *et al.*, "An algorithm for suffix stripping," 1980.

[25] M. Creamer, "Despite digital privacy uproar, consumers are not opting out," http://adage.com/article/digital/digital-privacy-uproar-consumers-opting/227828/, may 2011.

[26] R. T. Fielding, "Tracking preference expression (dnt): W3c working draft 14 november 2011," http://www.w3.org/TR/2011/WD-tracking-dnt-20111114/, Nov. 2011.

[27] J. Brookman, S. Harvey, E. Newland, and H. West, "Tracking compliance and scope: W3c working draft 14 November 2011," http://www.w3.org/TR/2011/WD-tracking-compliance-20111114/, Nov. 2011.

[28] N. Wingfield, "Apple adds do-not-track tool to new browser," http://online.wsj.com/article/SB10001424052748703551304576261272308358858.html, April 2011.

[29] J. Brock, "Why not track? results from an informal survey," http://blog.privacychoice.org/2011/09/14/why-not-track-results-from-an-informal-survey/, Sep. 2011.

[30] A. McDonald and J. Peha, "Track gap: Policy implications of user expectations for the'do not track'internet privacy feature," *Information Privacy Law eJournal*, vol. 5, 2012.

[31] R. Singel, "Google holds out against do not track flag," http://www.wired.com/epicenter/2011/04/chrome-do-not-track/all/1, Apr. 2011.

[32] Google, "Advertising and privacy," http://www.google.com/privacy/ads/, Apr. 2011.

[33] C. Dugihh, "How companies learn your secrets," http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?src=me&ref=magazine, Feb 2012.