Understanding Web Advertising Privacy Through Browser Instrumentation

Jovanni Hernandez
Drexel University
hj57@drexel.edu
TRUST- REU
Stanford Security Lab

Abstract

A growing number of websites serve tracking content from third party advertisers, advertising networks, advertising exchanges, advertising data providers, and more. Most consumers are unaware of what information is gathered and how it is used. We conducted web crawls with a new browser instrumentation tool to better understand the privacy-related business practices of the largely unregulated and unstudied online advertising ecosystem.

Keywords

Online Privacy, Browser Instrumentation, Fingerprinting, Cookies, Supercookies

1 Introduction

Within the last few years, online advertisement companies have switched their business models to include the emerging concept of third party tracking. In its most basic concept, third party tracking allows advertisers to follow users across multiple websites that are a part of their network or a partnering company's network. This allows advertisers to use techniques that map online activities into segments of ads that are likely more relevant to convert into a sale. Third party tracking is also used for advertisement analytics, frequency capping, and various other details.

There are many methods used by online advertisement companies to track users across websites. The most common method is a tracking cookie, or a small script that contains a unique identification tag or number that is stored on a user's computer by the website's server. In addition, some companies use web beacons [1] to generate log files on their servers and also to set cookies. Other possible tracking methods include browser fingerprinting, history sniffing, and Flash cookies.

Prior studies on the online advertising industry often fail to provide consistency or reliability because the measurement platforms used differ between research laboratories. Using a properly instrumented browser that is able to detect a myriad of dynamic web content solves this issue of inconsistency. Using the open sourced community platform, FourthParty [2], initially developed by Stanford researcher Jonathan Mayer solved those caveats. The platform was built using Mozilla Firefox's Addon-SDK, designed to work with production versions of Firefox 4.0 and up. The platform dumps dynamic content that is recorded into a SQLite Database to be later analyzed. For the studies below, the MozMil [3] automation extension for Firefox was used to crawl the desired Alexa Top Websites list [4].

2 Opt Out Behaviour and the DNT header

The first study [5] conducted using the FourthParty platform was designed to analyze the behavior exhibited by opt-out systems of members of the Network Advertising Initiative (NAI). The NAI is a cooperative of various online advertisers, marketers, and analytics companies that was designed to establish responsible business practices and increase consumer awareness. As a part of the NAI program, participants are required to provide a system that allows consumers to opt-out. Contrary to the typical understanding of an opt-out system which stops advertisers from tracking consumers, the NAI only requires its partners to stop showing consumers targeted ads based on the information gathered from their tracking.

Prior research has shown that that the NAI opt-out program confuses most consumers [6] and that only half of member companies go beyond the NAI commitment in their privacy policy and promise to stop tracking after a user opts out. [7] We conducted an experiment to measure how many NAI members leave tracking cookies in place after a user opts out. We also measured how many NAI members respond to the Do Not Track header [8], a new opt-out mechanism opposed by the online advertising industry.

2.1 Methods

To conduct this experiment, we initially navigated popular websites to locate tracking content from NAI member companies; we identified content from 64 of the 75 members. We carried out three tests on each company. First, we loaded the tracking content in a clean browser profile. Second, we opted out of the company on the NAI website, then loaded the tracking content. Third, we enabled the Do Not Track header and then loaded the tracking content. We manually labeled the cookies set during each test to identify tracking; if a cookie lasted longer than the current browser session and contained a seemingly unique value, we labeled it as a tracking cookie. Finally, we compared our results to a recent Carnegie Mellon study of opt-out commitments in NAI member privacy policies [7].

2.2 Results

The first and most noteworthy finding was that two companies, Media6Degrees and BlueKai, began to make headway in compliance with the Do Not Track header. In the case of Media6Degress, old tracking cookies are deleted and a new opt-out cookie is set upon receiving a Do Not Track request. In the case of BlueKai, old tracking cookies are kept, but new cookies are not set upon receiving a Do Not Track Request.

The second finding was that 32 of the 64 NAI companies studied kept tracking cookies in place after opting out. When comparing these results to Carnegie Mellon's privacy policy categorizations, 7 of the 32 companies that continued tracking after opting-out were seemingly in violation of their privacy policies. Digging deeper and manually retrieving these policies seems to confirm these findings, but may be interpreted otherwise [Appendix 2]. In response to these findings, various companies such as NetMining and Wall Street on Demand updated their privacy policies. In addition, other companies such as AudienceScience and Vibrant Media reached out to clarify their business practices [Appendix 3].

The last finding was that several companies went beyond their Carnegie Mellon categorization, removing their tracking cookies upon opting-out even though they promise to only stop behaviorally targeted of ads. These companies include: BluKai (retains city-level geolocation), Dapper, FetchBack, Google, Invite Media, Media6Degrees, Mediaplex, Quantcast, TidalTV, and YuMe.

2.3 Conclusion

This area of online advertising has not been studied before in such depth. Although our methodology was fairly naïve, it gave an initial view of what most companies view as an opt-out and how that opt-out actually affects consumers. Ultimately, these types of studies have begun to lead companies to create greater transparency in their business practices

3 Effectiveness of Tracking Protection Lists

The second study conducted using the FourthParty platform was designed to test the effectiveness of common tracking protection lists. Tracking protection lists are community or organization maintained lists designed to protect consumer privacy by blocking third party content using whitelists or blacklists. Whitelists contain known good content serving domains and allows them to communicate with the browser, blocking domains that are unknown. Blacklists work in the opposite way, blocking known bad domains from communicating with the browser and allowing all that are unknown.

As of the release of Microsoft's Internet Explorer 9, the built in ability to add TPLs has been touted as a main feature. Among the recommended lists for use with Internet Explorer 9 are Abine [9], Easy Privacy [10], two variations of Privacy Choice [11], and TRUSTe [12].

3.1 Caveats

Directly comparing these tracking prevention lists is not the most effective method of measurement. Because these lists are community or organizationally driven, the methods used in ensuring privacy vary greatly. Below are the descriptions for appropriate usage as explained by the creators of the lists:

According to Abine, their list is designed to:

> *"…blocks many online advertising and marketing technologies that can track and profile you as you browse the Web."*

According to the EasyList group, their EasyPrivacy list is designed to:

> *"… completely removes all forms of tracking from the internet, including web bugs, tracking scripts and information collectors, thereby protecting your personal data. "*

According to Privacy Choice, their first list is designed to:

> *"[Blocks] companies not subject to oversight by the Network Advertising Initiative (currently 508 domains)."*

While the second list is designed to:

> *"[Blocks] all tracking company domains in the PrivacyChoice database (currently 672 domains). "*

According to Trust, their list is designed to:

> *"…activate TRUSTe's Tracking Protection List (link below) to block companies that offer poor privacy protection, while ensuring that trustworthy companies who protect their privacy can continue to provide them with a richer, more personalized browsing experience. "*

This separates the lists into two categories, lists that block all known third party tracking content, and lists that block *select* known third party tracking content. In order to properly gauge their effectiveness, it is only possible to compare lists of the same type. Therefore, only Easyprivacy, PrivacyChoice list 2, and Abine can be compared.

3.2 Methodology

Keeping in mind the previously mentioned caveats, only the Easyprivacy, PrivacyChoice list 2, and Abine lists can be compared in terms of their effectiveness. In addition to these TPLs, two vanilla crawls with no TPL were run to serve as controls along with a test of Easylist alone and Easylist combined with Easyprivacy. This experiment was initiated by loading the TPLs into Firefox by using a plug-in developed by Abine called Do Not Track Plus [13]. Using the Mozmil automation extension, the Alexa Worldwide Top 1,000 websites were visited three times. For each page load, private browsing was turned on, and turned off before loading the next page, starting the cycle over again. This was done to ensure that no cookies from viewing the previous page influenced the cookies that were set from visiting the new page.

When the crawls for all five tests were completed, they were first filtered to only show cookies set by known tracking domains using the list compiled by PrivacyChoice. Next, they were manually filtered to remove any cookies that appeared to not contain uniquely identifying strings, session cookies, and interest segments. These results were then compiled into a list showing the frequency of cookies set and their unique domains.

3.3 Results

The data collected proved that there is a strong correlation in blocking advertisements and the amount of tracking cookies that are set. When combing EasyList and EasyPrivacy, the amount of trackers that were set dropped down to 246 from 25 unique tracking domains. Using EasyList alone, 2,686 tracking cookie were set from 73 unique tracking domains. EasyPrivacy alone allowed 4,967 tracking cookies to be set from 125 unique tracking domains.

On the less effective side, PrivacyChoice list 2 allowed 11,573 tracking cookies to be set from 199 unique tracking domains. Abine, which also performed poorly, allowed 13,089 from 197 unique tracking domains.

Putting these numbers into perspective the first vanilla crawl resulted in 13,943 tracking cookies being set from 201 unique tracking domains. The second vanilla crawl resulted in 13,647 tracking cookies being set from 203 unique tracking domains. Combined, these average to 13, 795 tracking cookies being set from 202 unique tracking domains.

3.4 Conclusion

These results are still very early findings, and there is still much work to be done in order to accurately test the effectiveness of tracking protection lists. Although it is proven that directly comparing certain tracking prevention lists is inadequate, there are other methods that could be used. In the case of PrivacyChoice list 1, the post crawl results can be filtered to show domains that still set tracking cookies and are not a part of the NAI. Unfortunately, testing TRUSTe's TPL is currently not possible because their interpretation of, "*companies that offer poor privacy protection,"* has not been defined and could possibly mean anything.

4 Limitations of Browser Instrumentation

While a properly instrumented browser that records dynamic web content is very useful for online research, it does have its limitations. When dealing with some third party content and the methods in which that content is served, incomplete data may result. This was the case in experimenting with the platforms ability to detect the presence of the AdChoices program on various websites and advertisements.

4.1 Background

The AdChoices program is a collaborative effort of the nation's largest self-regulatory groups. It was designed to create transparency in their business models and give consumers better understanding and control over interest-based advertisements. The program revolves around a simple concept: a triangular icon with a centered "i" that is placed on advertisements or on websites where data is collected.

In a June press release the Digital Advertising Alliance, the organization behind the AdChoices program, touted that 100 companies are currently participating [14]. They claimed that 9 of the nation's 10 largest advertising companies are fully engaged in the program. In addition, the press release claimed that 90 more companies had registered for the program and would be deploying in two months. Lastly, the press release claimed that over two trillion online ads have been displayed with the AdChoices icon since the programs launch. We conducted a study to measure how frequently the icon appears to previous experiments ran, it was assumed that the actual ratio is much lower than the paper seemed to imply.

4.2 Methodology

To test this theory of ratio, a large single visit crawl of the Alexa Wordwide Top 10,000 Websites was recorded. When the crawl had completed, the recorded results were filtered to show HTTP requests and HTTP referrers from the visited websites. The filtered results were then queried for requests with a value containing "adchoice" or "ad choice." It was apparent by the low amount of results that this method was ineffective. Manually inspection of ads on several popular websites revealed that there is no standardized URL schema for ads or icons.

5 Stepping away from Browser Instrumentation

Because of this caveat, a new method had to be used to gather the data. With no other automated solutions, manual inspection was required, which in turn led to a smaller test and sample size [15]. First, the Alexa top 500 websites list in the United States was gathered. Next, each website's homepage was manually inspected taking note of the number of ads served by third party advertisement companies and the number of those ads that actually contained the AdChoices icon on or near the advertisement. In addition, the placement of the icons relative to the advertisements and the page in which they were linked to were recorded. Using Chrome's "inspect element" the different sizes of icons were determined. Lastly, everything was documented by using Webpage Screenshot [16], a Chrome extension, to capture the entire website.

5.1 Results

Although very time consuming, manual inspection did reveal a great deal of information concerning the AdChoices program. Out of the 627 advertisements encountered, 62 (9.9%) contained the AdChoices icon. Restricting our dataset to the 449 non-explicit, domestic websites in the Alexa U.S. top 500 list, we spotted 512 third-party ads. 58 (11.3%) of those ads had AdChoices symbol. The typical size of the icon is 13x13 pixels and non-descriptive when compared to the smallest standard display ad size of 88x31 pixels.

In several cases, the icon was actually not in the advertisement but in the footer of the webpage serving the ads, which was observed 13 (2.6%) times. Often times, The accompanying text is in small font and reads, ambiguously, "AdChoices." Now we learn that the icon rarely shows up and, half the time, doesn't even include any text.

A full spreadsheet of results is available in Excel Format from the initial coverage [15]

5.2 Conclusion

Our study suggests that the actual ratio of ads displaying the AdChoices icon is miniscule in comparison to the amount of ads served daily. We also prove that the current system lacks consistency in the ways the icon is displayed in terms of positioning, size, and overall appearance.

6 Final Notes

The uses of proper browser instrumentation on a readily available and popular production browser gives researchers the opportunity to measure different types of online dynamic content without the need to spend time modifying non-production browsers. This allows for consistency in results and a better focus on the study of data in place of the actual collection of data. The case studies provided show various examples on how data collected using this method can be used.

Using an instrumented browser gave insight into the online advertising industry in a way that has not been previously studied. In this paper, opt-out behavior was observed along with the different patterns exhibited by these mechanisms and how they affect consumers. In addition, the effectiveness of several tracking prevention lists were measured. Lastly, various caveats were explored and possible workarounds were provided. This area of research still remains widely unknown and has many available opportunities for further research.

7 Acknowledgments

I thank my fellow research assistant Akshay Jagadeesh for his collaborative efforts on the various experiments that we conducted. In addition, I thank Jonathan Mayer for introducing me to this area of research, guiding our team on these experiments, developing the initial release of the measurement platform used, and for reviewing all of the collected data. Lastly, I thank John C. Mitchell for helpful feedback on research and suggestions for improvement.

This work was supported in part by TRUST (Team for Research in Ubiquitous Secure Technology), which receives support from the National Science Foundation (NSF award number CCF-0424422) and the following organizations: AFOSR (#FA9550-06-1-0244), BT, Cisco, DoCoMo USA Labs, EADS, ESCHER, HP, IBM, iCAST, Intel, Microsoft, ORNL, Pirelli, Qualcomm, Sun, Symantec, TCS, Telecom Italia, and United Technologies.

Appendix I – Opt-Out Behaviour and the DNT header Results

Company	Tracking cookies remain after opt out?	Tracking cookies remain after DNT?	CMU Privacy Policy Categorization
[x+1]	Yes	Yes	
24/7 Real Media	Yes	Yes	Collect no data
33Across	No	Yes	Don't target ads
Adara Media	Yes	Yes	Don't target ads
AdBrite	Yes	Yes	Collect less data
AdChemy	No	Yes	Collect no data
Adconion	Yes	Yes	Stop tracking
AddThis	No	Yes	
AdMeld	No	Yes	Collect no data
AggregateKnowledge	No	Yes	
Akamai (aCerno)	Yes	Yes	Don't target ads
AlmondNet	No	Yes	Collect no data
AOL	Yes	Yes	Don't target ads
AudienceScience	Yes	Yes	Collect no data
Batanga (DoubleClick cookie)	No	Yes	Collect no data
Bizo	No	Yes	Stop tracking
BlueKai	No (retains city-level geolocation)	Yes (but not updated)	Collect less data
BrightRoll	No	Yes	
Burst Media	Yes	Yes	
Buysight	No	Yes	Collect no data
Casale Media	No	Yes	Stop tracking
Cognitive Match	Yes	Yes	
Collective	No	Yes	Collect no data
Criteo	Yes	Yes	Don't target ads
Cross Pixel Media	Yes	Yes	
Dapper	No	Yes	Collect less data
DataXu	Yes	Yes	
Dedicated Networks	No	Yes	Collect no data
Dotomi	No	Yes	
eXelate	Yes	Yes	Don't target ads
FetchBack	No	Yes	Don't target ads
Fox Audience Network	Yes	Yes	Don't target ads
Glam Media	No	Yes	Collect no data

Google	No	Yes	Collect less data
interCLICK	No	Yes	Stop tracking
Invite Media	No	Yes	Don't target ads
Lotame	Yes	Yes	Don't target ads
MAGNETIC	Yes	Yes	Don't target ads
Media Innovation Group	No	Yes	
Media6Degrees	No	No	Don't target ads
MediaMath	Yes	Yes	
MediaMind	No	Yes	
Mediaplex	No	Yes	Don't target ads
Microsoft (Atlas)	Yes	Yes	Don't target ads
Microsoft Advertising	Yes	Yes	Don't target ads
Mindset Media	Yes	Yes	Don't target ads
Netmining	Yes	Yes	Collect no data
Pulse360	No	Yes	
Quantcast	No	Yes	Don't target ads
RadiumOne	Yes	Yes	
Red Aril	No	Yes	Collect no data
richrelevance	Yes	Yes	Don't target ads
Rocket Fuel	No	Yes	Stop tracking
Specific Media	Yes	Yes	Don't target ads
TARGUSinfo	Yes	Yes	Don't target ads
TidalTV	No	Yes	Don't target ads
Tribal Fusion	No	Yes	Stop tracking
Turn	Yes	Yes	Don't target ads
Undertone Networks	Yes	Yes	Collect no data
ValueClick	Yes	Yes	Don't target ads
Vibrant Media	Yes	Yes	Collect no data
Wall Street on Demand	Yes	Yes	Stop tracking
Yahoo!	Yes	Yes	
YuMe	No	Yes	Don't target ads

Appendix II – Privacy Policies

The **24/7 Real Media** privacy policy claims that a user may *"opt out of receiving our ad delivery, audience management and behavioral targeting cookies."* We found that opting out deleted the company's tracking cookies, but reloading the content reinstalled the tracking cookies.

AdConian's privacy policy states that a user is *"free to opt out of the Adconion Cookie."* Opting out deleted one of three tracking cookies but left the other two in place. Reloading the content did not update the remaining tracking cookies.

In its privacy policy, **AudienceScience** describes its opt-out option as follows: *"Should you choose to opt-out, we delete all previously collected information from the cookies, and put new information in the cookie which tells us to stop collecting information from that device."* We found that opting out of AudienceScience removes its unique tracking cookie but does not remove a highly unique cookie that represents the user's interests. Subsequent loads of the content updated the interest cookie.

NetMining's privacy policy states that upon opting out *"we will delete your existing ntmng.com or netmining.com cookie(s) and try to place a new cookie that instructs us not to track your future activities when we detect that cookie."* Opting out deleted the Netmining tracking cookie but did not delete a tracking cookie served from a retailer-specific subdomain of netmng.com (and presumably only used on that retailer's site). Reloading the content refreshed the retailer-specific cookie.

The **Undertone** privacy policy notifies users: *"If you would like to opt out of OBA, then we offer 'opt-out cookies' to block the tracking and placement of future Undertone cookies for OBA purposes on your system for five (5) years."* Opting out removed a highly unique cookie that stores the user's interests but did not remove a unique cookie. Subsequent loads of the content updated the unique cookie.

Vibrant Media's privacy policy provides: *"If you'd like to opt-out from having Vibrant Media collect your Non-PII in connection with our Technology, please click here. When you opt out, we will place an opt-out cookie on your computer. The opt-out cookie tells us not to collect your Non-PII to tailor our online advertisement campaigns."* Opting out of Vibrant Media does not remove the network's unique tracking cookie; the cookie remains in place and is updated with subsequent loads of the content.

The privacy policy on **Wall Street on Demand**'s advertising platform claims: *"By clicking here, the unique cookie used by this system/domain and stored locally by your browser will be changed to 'OPT_OUT'. By creating a generic cookie id instead of a unique cookie id - it is even more impossible to track your history."* Opting out deleted Wall Street on Demand's unique cookie, but left in place a seemingly highly unique cookie that appears to store user interests. Refreshing the content renewed the interests cookie.

We identified one additional company with a privacy policy that may be interpreted to prohibit its current business practices. The **TARGUSinfo AdAdvisor** opt-out page explains that *"[t]he AdAdvisor opt-out works by replacing the existing AdAdvisor cookie with a new cookie that clearly indicates that the user has elected to opt-out of the Services."* Opting out left TARGUSinfo's unique tracking cookie in place. Refreshing the content did not update the tracking cookie.

Appendix III – Updated Privacy Policies and Clarifications

AudienceScience reached out to clarify its practices. Its cookies store a compressed and encrypted data structure. When a user opts out, AudienceScience removes all interest segments and the unique ID from the data structure, but it continues to update the last time the browser contacted its servers. We have confirmed that AudienceScience now entirely removes its data structure after opting out.

Netmining has updated its privacy policy:

> *If you select the "opt out" button there for Netmining, we will delete your existing netmng.com or netmining.com online behavioral advertising cookie(s) and try to place a new cookie that instructs us not to track your future activities for the purposes of serving online behavioral advertising when we detect that cookie.*

Vibrant Media submitted the following statement:

> *We drop a user ID cookie when a user initiates engagement with one of our ad units. This collects non-personally identifiable information on keywords a user has engaged with. If the user doesn't visit a site in our network for 10 days, we delete this data. If someone opts out, we add a do-not-track cookie.*
>
> *We had been deleting any data associated with the user ID, but had not been deleting the cookie itself (this is acceptable for NAI compliance). When we encounter someone with a do-not-track cookie, we completely ignore the user ID and therefore don't use their information to serve ads. Although the cookie was remaining, we do not reference or use the ID in any way and we completely delete all data, be it in logs or storage devices for that particular user ID. Going forward, in order to prevent any misunderstanding we will also be deleting that cookie.*
>
> *We have always been vigilant about adhering to industry best practices and NAI compliance policies.*

***Wall Street on Demand** has updated its privacy policy:*

> *Online Behavioral Advertising (OBA) is the process of targeting specific advertisements to each individual user, based on browsing history. If you opt out of OBA from our service by clicking the link below, the OBA cookie we use to contain this information will be emptied and changed to a placeholder signaling that you have done so. . . . Opting out does not necessarily delete or replace all cookies from our domain; others may remain which are used for aggregate reporting on the performance of the advertisements we serve.*

References

1. Alliance, N.A. (2004) *Web Beacons – Guidelines for Notice and Choice.*
2. Mayer, J. *FourthParty Web Measurement Platform.* 2011; Available from: http://fourthparty.info/.
3. Mozilla. *Mozmil* 2010; Available from: https://developer.mozilla.org/en/Mozmill.
4. Amazon. *Free traffic metrics, search analytics, demographics, and more for websites...* 2011; Available from: http://www.alexa.com/.
5. Mayer, J. *Tracking the Trackers: Early Results.* 2011; Available from: http://cyberlaw.stanford.edu/node/6694.
6. McDonald, A.M., *Footprints Near the Surf: Individual Privacy Decisions in Online Contexts*. 2010, Carnegie Mellon University.
7. Komanduri, S., et al., *AdChoices? Compliance with Online Behavioral Advertising Notice and Choice Requirements*. 2011, Carnegie Mellon University.
8. Mayer, J. and A. Narayanan. *Do Not Track: A Universal Third-Party Web Tracking Opt Out.* Internet-Draft 2011; Available from: http://datatracker.ietf.org/doc/draft-mayer-do-not-track/.
9. Abine. *IE9 Tracker Protection List.* 2011; Available from: http://www.abine.com/tpl/.
10. EasyList. *The Official EasyList Website.* 2011; Available from: https://easylist.adblockplus.org/en/.
11. PrivacyChoice. *TrackerBlock for Internet Explorer 9.* 2011; Available from: http://www.privacychoice.org/trackerblock/ie9.
12. TRUSTe. *Safer IE9 Browsing with TRUSTe.* 2011; Available from: http://tracking-protection.TRUSTe.com/.
13. Abine. *Do Not Track Plus.* 2011; Available from: http://www.abine.com/dnt/.
14. Alliance, D.A. (2011) *Digital advertising alliance announces first 100 companies participating in self-regulatory program for online behavioral advertising.*
15. Hernandez, J., A. Jagadeesh, and J. Mayer. *Tracking the Trackers: The AdChoices Icon.* 2011; Available from: http://cyberlaw.stanford.edu/node/6714.
16. Goldstein, A. *WebPage ScreenShot.* 2011; Available from: http://www.webpagescreenshot.info.