Analysis of Web Tracking and the Efficacy of Ad-Blockers

Ivan Alvarado
Carnegie Mellon University
imalvara@andrew.cmu.edu

Leo Ip
Carnegie Mellon University
lhi@andrew.cmu.edu

Rob McCarthy
Carnegie Mellon University
rmccarth@andrew.cmu.edu

Marisa Midler
Carnegie Mellon University
mmidler@andrew.cmu.edu

Joshua Vasko
Carnegie Mellon University
jvasko@andrew.cmu.edu

ABSTRACT

Over time many features have been incorporated into websites since the internet's initial launch. Many of today's websites need user interaction to perform their functions and companies track this interaction by collecting information about users and their web-browsing behaviors. Many companies use web cookies to perform this user tracking. With this in mind we conducted research to examine the effectiveness of free ad blocking browser plugins in blocking these tracking cookies to provide average internet users with information on which ad blocker plugin blocks the most cookies and by extension is likely to protect their privacy the best.

We collected the web cookies from the Alexa top 5,000 sites using the OpenWPM framework. We used a Mozilla Firefox web browser to test the effectiveness of the Disconnect, Ghostery, and uBlock Origin browser plugins. We evaluate the number of cookies blocked by each browser extension compared to the control group, and conclude that the Ghostery extension effectively blocked the most cookies. We discuss a few particular reasons as to why certain sites trigger the most cookies, particularly highlighting the value-proposition that exists for sites dealing in free information access and finish with recommending a number of future studies that can expand on the insights gleaned from this research.

Keywords—privacy, web, tracking, ad blockers, cookies, OpenWPM

1. INTRODUCTION

In today's world, surveillance is ubiquitous off and online. Offline surveillance is easier to notice, as it includes devices and methods that are apparent to individuals without specialized knowledge such as cameras or microphones. However, surveillance on the internet is shrouded in mystery to common users as they do not have the knowledge required to understand how their privacy is being invaded by online tracking.

In a study conducted in 2013, it has been found that about half of the population of online users are not willing to share any personal information to companies for the purpose of bettering their user experience online (Leon, et al., 2013). In addition to the lack of willingness to share their personal information with companies, it was also found that fewer than 10% of

internet users were aware of cookies and third-party tracking as a practice (Agarwal, et al., 2013). The fact that companies are collecting user data through their use of the internet constantly along with a lack of users' willingness to share information and a low rate of user awareness creates a dangerous platform that common users do not even know that they are participating in.

It is important that researchers conduct effective investigations on privacy-protecting technologies to inform users of and recommend optimal methods to protect their privacy while utilizing the internet. Our research specifically addresses which ad-blocker is the most effective in protecting users from cookies while searching popular websites on the internet. It does this by querying the Alexa top 5000 websites and tracking the number of cookies that are blocked through the utilization of several privacy-protecting tools: uBlock Origin, Disconnect, and Ghostery.

Our investigation found that Disconnect was by far the least effective privacy-protecting tool that was investigated, even though it blocked over 60% of the cookies present on the Alexa top 5000. Ghostery and uBlock Origin both blocked over 80% of the total cookies that were generated through our analysis, with Ghostery performing the best out of all tested plugins, blocking 85.5% of all cookies across the Alexa top 5,000 sites.

Our research is an attempt to provide common users with the requisite knowledge needed to understand these surveillance technologies and provide data on privacy-protecting tools that may be used within Mozilla Firefox. In order to accomplish the two previously mentioned objectives our research will discuss the following:

- In section 2, a comprehensive literature review of past and present research on the topic of online tracking, internet privacy, and how the ad blockers work.
- Section 3 contains the experiment motivation, rationale, methodology, and process detailing the test of three commonly used privacy-protecting tools on the Alexa top 5000 sites.
- Section 4 includes the statistical analysis of the data collected from the experiment.
- In section 5, the discussion of previous research findings compared to the findings of this experiment, privacy harms of online tracking, and objective analysis of the most effective plugin.

2. LITERATURE REVIEW

History of Online Tracking

Before discussing the history of online tracking, it is important to note that the internet has not changed much since its inception. It has existed in a state similar to what it currently is since 1989 when Tim Berners-Lee first proposed the World Wide Web for research use which subsequently led to The World becoming the first Internet Service Provider (ISP) in history (Gale Group, 2011).

With the creation of websites came the creation of log files that tracked each hit a web page received. As the internet became populated with more users over time, content creators became more concerned with the traffic that their web page received. Recognizing the market need for web analytics WebTrends, the first platform for web analytics and reporting, was created in 1995 (Earnshaw, 2007). WebTrends went public in 1998 stating its aim to "facilitate analysis and reporting of Web site traffic, [and] Internet advertising campaigns" (Webtrends, 1999). Though web tracking initially began as a means to track a web pages popularity, online advertising became the main driving force behind tracking as it was recognized as a less intrusive medium to market products than

telemarketing and more convenient than radio marketing (Wall Street Journal, 2001).

In an attempt to analyze the history of web analytics the University of Washington (UW) analyzed web data from the Wayback Machine, an archive of web page data from 1996, and ran their proprietary tool TrackingExcavator that collects online tracking mechanisms such as cookies and JavaScript (Gale Group, 2016). UW found that in the early 2000s individual web trackers could not be found on more than 5% of web pages and that tracker usage increased to 10% in the late 2000s (Gale Group, 2016). Today the most popular web tracker, Google Analytics, can be found on over a third of all sites analyzed by UW (Gale Group, 2016). The rampant use of web tracking raises the question of privacy concerns for individual consumers when interacting with the Internet.

With about a third of the world's population using the internet, massive amounts of behavioral data are being collected each day (Gale Group, 2011). The amount and scope of the data that companies collect is unknown to average users on the internet (Ur et. al, 2012). In addition to not being aware of this data collection, individuals are unaware that practices such as online behavioral advertising (OBA) are currently being used to market products to internet users (Ur et. al, 2012). After further discussion, opinions from consumers on tracking and its use in OBA were split: on one hand they thought it was an intelligent use of their data, on the other they deemed the practices creepy in nature (Ur et. al, 2012). It is pertinent that individuals, as well as those responsible for developing regulations are aware of online data collection practices as well as the privacy concerns stemming from them in order to craft laws that best protect and inform the individual user.

Another factor that allowed for the rise of online tracking includes the regulation surrounding the practice. In present-day topics concerning privacy, including web tracking, are largely regulated by individual countries for individual countries (Falahrastegar et. al, 2014). Due to these fragmented regulations regions are protected according to the ability of their regulators. The United States and Europe, two of the main consumers of online content, have vastly different regulations that govern privacy practices (Falahrastegar et. al, 2014). The United States' slow-moving regulatory process allowed for the online advertising industry to institute industry friendly practices in an attempt to curb the desire to enact policy on the industry. Even though the United States has the Federal Trade Commission (FTC) that helps protect consumer privacy its power as well as the industry self-regulation ultimately place the United States far behind Europe when it comes to protecting consumer privacy (Falahrastegar et. al, 2014). Europe has instituted data protection authorities (DPA) that are tasked to thoroughly investigate and enforce the application of data protection law. With the introduction of the General Data Protection Regulation (GDPR), Europe has also empowered citizens to utilize these DPAs to help protect their privacy while online (Nicolaidou and Georgiades, 2017). Though Europe has instituted policies to help curb intrusive data collecting practices, the ubiquity of online tracking has made way for a market need to oppose its intrusive practices in the form of ad blockers.

The first commercial ad blocker, Ad Muncher, was introduced in the in the 1990s and was responsible for removing advertisements in all applications (Kudryavtseva, 2017). By 2010 the worldwide population of individuals utilizing ad-blocking software was approximately 21 million and this amount increased tenfold by 2015 (MIT Technology Review, 2016). One possible explanation for the increase in ad blockers is the increased use of advertisements slow down internet connections. From 2009 to 2014 the number of video ads embedded within web pages almost doubled (MIT Technology Review, 2016). Though there's no definite explanation as to why ad blockers are being utilized to a greater extent, however, it can be stated with certainty that the

ads they are trying to disrupt are not losing their relevance anytime soon.

How Different Ad Blockers Work

Currently, browser plugins are the most common method to combat online web advertisements and online tracking, and many of these plugins use lists of known trackers to implement the blocking functionality. There are many ad blocker plugins available but this project focuses on three popular browser plugins, Disconnect, Ghostery, and uBlock Origin. Browser plugins are not the only method to be created in an attempt to prevent tracking. The World Wide Web Consortium (W3C) created a browser setting standard called Do Not Track (DNT) (Hanson, et al, 2018). DNT is a global browser setting that indicates to websites that the user does not want to be tracked; unfortunately, the DNT browser setting has been proven to be ineffective as many websites simply ignore it (Hanson, et al, 2018).

Additionally, not all ad blocker plugins behave the same. This research project was originally going to analyze a fourth browser plugin, Privacy Badger, but due to the how Privacy Badger uses an algorithm to identify trackers as you continue to browse websites and does not block trackers by default, it would not generate any insightful data due to how our experiment is conducted (Boumans and Poll, 2017). Below are the web trackers used in this research project with their advertised features and a brief description of how they work.

Disconnect

The Disconnect browser plugin is available for Google Chrome, Mozilla Firefox, Apple Safari, and Opera browsers (Disconnect, n.d.). It is advertised to "block 2000+ tracking sites, load pages 27% faster," and is available to the public for free (Disconnect, n.d.). Disconnect utilizes a blacklist which includes the "2000+ tracking sites" mentioned above and also incorporates the option to whitelist websites (Boumans and Poll, 2017).

Ghostery

The Ghostery browser plugin is available for seven different browsers: Google Chrome, Mozilla Firefox, Apple Safari, Opera, Microsoft Edge, Microsoft Internet Explorer, and Cliqz for free (Ghostery, n.d.). Ghostery is advertised as "optimizing page performance" while managing trackers that collect user information, it also claims to have the largest tracker database of all the privacy tools offered (Ghostery, n.d.; Boumans and Poll, 2017). Ghostery has a user interface (UI) built into the web browser which allows the user to easily change settings, see information about the trackers on the web page, plugin performance metrics, and customize blocking settings for specific web trackers (Ghostery, n.d.; Hanson, et al, 2018). Ghostery can be configured to block "all or specific third-parties or allow (but not block) all tracking on a particular first-party website" (Hanson, et al, 2018).

uBlock Origin

The uBlock Origin browser plugin is available for Google Chrome, Mozilla Firefox, Apple Safari, and MacOS for free (uBlock, n.d.). This browser plugin is advertised to block advertisements, web trackers, speed up web browsing experience, and even "help protect against some forms of malware" (uBlock, n.d.). uBlock Origin operates with a blocking "list of domains" to block advertisements and trackers (Boumans and Poll, 2017).

Previous Research

A couple of studies have tested the efficacy of ad blockers on controlling OBA in addition to the amount of online tracking that is being conducted in general. A team of researchers from Carnegie Mellon University first studied this topic in 2012 by comparing ad blockers such as Ghostery to other privacy-protecting

tools such as Do Not Track headers (Balebako et al., 2012). The team first trained their clean browser to visit sites based around a general topic, such as pregnancy or bicycling, to build a browser history around these topics. After the training data was collected websites with large audiences that utilized text ads, such as nytimes.com and chicagotribune.com were visited to test if not only ads appeared, but if the ads were marketing products that had been researched to generate the training data (Balebako et al., 2012). Their research showed that for 4 of the 5 topics that the browser history was trained on exhibited OBA on the news websites that were visited (Balebako et al., 2012). As for the efficacy of privacy-protecting technologies, the research showed that ad blocking technologies blocked the advertisements from appearing on the news websites and were the most effective at blocking cookies accruing in the browser history when compared to other tools that did not block cookies altogether (Balebako et al., 2012).

A team of researchers created FPDetective, a framework to measure web fingerprinting, and tested the presence of fingerprinting trackers on 1 million of the top websites (Acar et al., 2013). FPDetective utilizes PhantomJS and Chromium to detect JavaScript, Plugin, and Extension based fingerprinting techniques. Their research showed that these fingerprinting techniques are present on less than a percent of the websites that were tested (Acar et al., 2013). The team also tested the efficacy of anti-fingerprinting techniques and found that the Tor browser was the only effective anti-fingerprinting tool and that Do Not Track headers were completely ignored by fingerprinters on the websites tested (Acar et al., 2013).

Another study from a team of researchers based at Princeton University utilized a proprietary tool, OpenWPM, to crawl websites and track the amount of fingerprinting-based and cookie-based online tracking that was occurring in addition to testing the efficacy of ad-blocking technologies (Englehardt and Narayanan, 2016). The team of researchers measured 1 million websites' contents in a FireFox browser utilizing Chromium to automate the process. Their research found that a minority of sites utilized fingerprinting to track users, but that the majority of sites that utilized this technique were among the most popular sites within the 1 million that were tested (Englehardt and Narayanan, 2016). Cookie-based tracking is more prevalent on websites and is present on a majority of sites that were tested. Among the trackers found only 1 third-party tracker, Google Analytics, was present on a majority of the 1 million sites tested (Englehardt and Narayanan, 2016). As for the ad-blocking technologies, it was found that Ghostery was the most effective tool, but that it struggled to block obscure cookie-based tracking when compared to the more popular methods of this type of surveillance.

3. Experiment

Background

Motivation

The present deployment of web pages makes it virtually impossible for users of the internet to browse the internet without encountering websites that conduct tracking and advertising as a facet of its user experience. Several Fortune 500 companies, such as Alphabet (Google) and Facebook, generate a bulk of their revenue by collecting user data and providing a medium to other companies for marketing and research purposes (Fortune, 2018). Due to the profitability of online tracking, companies have started to conduct increasingly intrusive research on individuals. The tracking and subsequent research that is being conducted it can be argued that these companies are violating the privacy of users that access online content.

Experimental Rationale

Privacy-minded users are faced with choices to make about which browser extensions to use natively. The

effectiveness of individual extensions vs the competition is unclear, so users often revert to word-of-mouth to help them decide. Our research was conducted with the hope that if users can be presented with clear key performance indicators between privacy extensions, they can make a more rational choice.

Web privacy measurement framework OpenWPM was utilized in order to test the effectiveness of Ghostery, uBlock Origin, and Disconnect against the top 5000 of the Alexa top 1 million websites. OpenWPM utilizes the Mozilla Firefox browser for automating HTTP requests, optionally storing JavaScript, profile (JavaScript + HTTP Request cookies), and Flash cookies in an SQLite database. (Princeton, 2018).

Tools and Procedure

Virtual Machine and Testing Setup

In order to ensure that the ad blockers were all tested under identical conditions, a pristine Ubuntu virtual machine (VM) was created for each iteration of testing. Ubuntu version 16.04 was selected because it is the latest version of Ubuntu that ensures functionality with the OpenWPM tool. In addition to using the common VM amongst the control and variable groups, we used a Carnegie Mellon University (CMU) IP address to ensure that each test was conducted as a member of the same network. After identical VMs were created and connected to the CMU network, each plugin was configured in an identical manner utilizing the procedure described by OpenWPM (Princeton, 2018).

Operating System. We chose to use Ubuntu (16.04 Release) as the operating system for our experiment. The OpenWPM framework only fully supported on Ubuntu versions 14.04 and 16.04 and we wanted to use the latest version to support our research efforts.

Application Selection. We considered other tools for analyzing third-party cookies but eventually decided on OpenWPM because of its integrated use of ad blocking plugins for use with the Mozilla Firefox web browser.

Browser Selection. We chose to use Mozilla Firefox for Ubuntu version 63.0.3 for our experiment as OpenWPM best supports this browser. Additionally, PrivacyTools.io recommends the browser as it is "open source and respects [user's] privacy" (PrivacyTools.io, n.d.).

Dataset. We obtained the top 5,000 websites from Amazon Alexa's Top 1 million database. This database was made available to us through one of our professors at CMU, Dr. Timothy Libert. We created a CSV file and referenced the list for the entirety of our project. The number one website on the list at the time we pulled the data was *google.com* and the 5,000th website was *torlock.com*. For our study, we utilized the cookie counts stored under the profile_cookies table as it contained both cookies added by Javascript and by HTTP Responses. The profile_cookie table also allowed us to discard the websites that did not produce any cookies due to timeouts or failures to load.

Application Preparation. We modified the *demo.py* by enabling the capture of cookies as per the OpenWPM documentation. Additionally, we structured the demo.py file to import the col_b from urls.py to provide a list of URLs to direct the crawl. In order to increase efficiency and reduce loads placed on our crawling devices, we disabled browser headers, disabled flash, and added a system time-print in order to ascertain the total length in hours of the crawls. urls.py was developed to reference and parse our list of 1 million URLs stored in urls.csv.

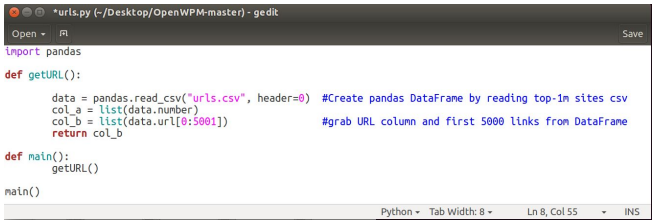

Figure 3.1 - Python Script Referencing Alexa Top 5,000 URLs

We modified the *default_browser_params.json* file for each plugin we tested:

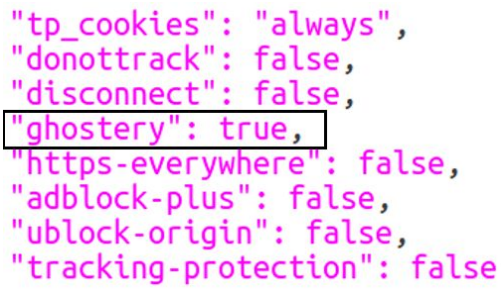

Figure 3.2 - Modified Settings to Accommodate Plugins

Network Configuration. We utilized the Carnegie Mellon University network to conduct our experiment:

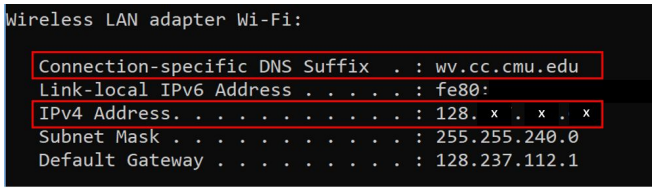

Figure 3.3 - Standard Network Settings for Study

Ad Blocker Testing

The default_browser_params.json file was only changed to specifically enable each individual extension for its respective data collection crawl. The individual extensions were not customized or provided any additional code or blacklisted URLs prior to the crawls.

Browser_commands.py was configured to log the load time for each URL in a loadtime.txt file. This data was imported manually to our central data repository. Following browser automation, the database images were collected and assembled into a Microsoft Access database where they were collated and exported into Tableau.

As OpenWPM executes, it displays configuration / settings to the user:

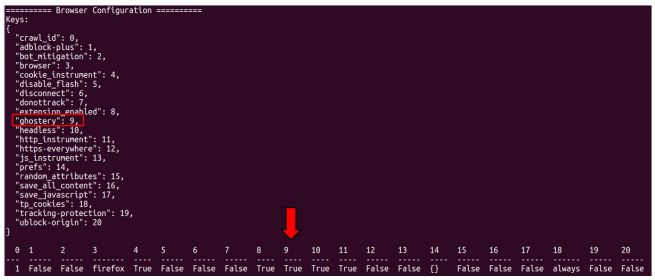

Figure 3.4 - Script Displays Settings to User

The OpenWPM collection can be viewed in the operating system terminal. Two distinct commands are executed for each website: “GET” and “DUMP_PROFILE_COOKIES.” The GET command opens the Mozilla Firefox web browser and visits each of Alexa’s top 5000 URLs (websites). It obtains the cookies and modifies the database. The “DUMP_PROFILE_COOKIES” command closes the tab and moves on to the next URL to repeat the collection process.

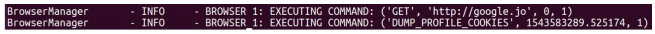

Figure 3.5 - OpenWPM Running in Terminal Window

Statistical Methodology

During the crawl, approximately 12% of the sites failed to load. These sites, their cookies, and their load times were eliminated from their respective data sets for the extensions in order to provide a fair sampling. An additional 233 sites were eliminated from the samples due to the control group not loading any cookies at all. Since the efficacy of an extensions blocking capability could not be tested against this control, it was stricken. These data points which retrieved no cookies were not stricken from the analysis present in figure 5.1.

A two-sample t-test was performed on each subset of data pertaining to each blocker, compared to the control group. The two-sample t-test was chosen as it will adequately address whether or not two samples' means are statistically significant from one another. A 99% confidence interval was chosen since the N was large for each sample population. Finally, the uBlock Origin plugin was directly compared to the Ghostery plugin for both load-time and cookie-blocking efficacy.

Load-time data was stored by OpenWPM in milliseconds (*ms)*. This time was converted for readability purposes to *s* before analysis was conducted.

4. Results

The OpenWPM script records the captured cookies into a database. Figure 4.1 shows an example of the cookie information and associated data which includes the creation time, expiration date, name, and associated host website.

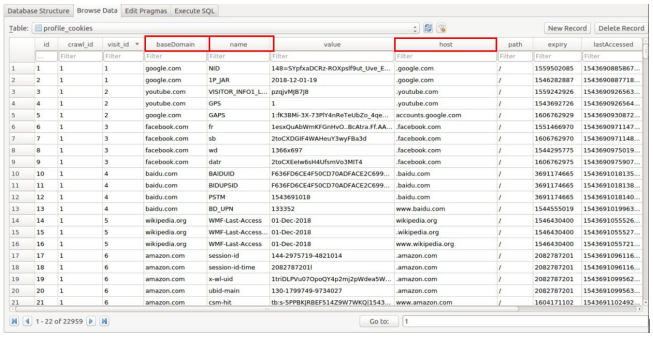

Figure 4.1 - Example of Cookie Data

Figure 4.2 shows that all the tested plugins were effective (blocked more than 51%) and that the Ghostery extension removed more than 85% of cookies. However, there is a difference in effectiveness between these tracking blockers as shown in Figure 4.3. Evaluating the top 10 websites where plugins performed best shows again that Ghostery was most successful. Both Disconnect and uBlock Origin outperformed each other in four of the top 10 evaluated sites. It is also worth noting that their difference was measured in fractions of a percentage point.

Plugin Utilized	# Cookies Not Blocked	% Cookies Blocked
N/A (Control)	157,934	-
Disconnect	61,218	61.2%
Ghostery	22,931	85.5%
uBlock Origin	25,882	83.6%

Figure 4.2 - Percentage of Cookies Blocked by Each Plugin

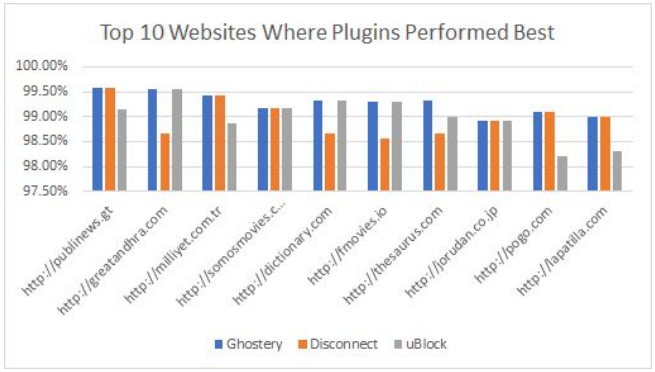

Figure 4.3 - Websites Where Plugins Performed Best

Two-sample t-tests and confidence intervals were conducted for each browser extension and the control group. The data was paired in the following format for both load-time analysis and cookie-blocking efficacy analysis: Disconnect v. Control, uBlock Origin v. Control, Ghostery v. Control.

Our analysis found that each plugin was effective in blocking a statistically significant number of cookies over the scope their individual sample sizes. All two-sample t-tests returned a p-value of 0.000 over a 99% confidence interval for both load-time and

cookie-blocking efficacy. Ghostery and uBlock Origin were directly compared for both load-time and cookie-blocking efficacy since they were the clear top-performers by a large margin in both categories.

All plugins on average increased site load-times. Disconnect increased load-time compared to the control by approximately 53 seconds 99% CI [47.86,58.25]. uBlock Origin increased load-time compared to the control by approximately 29.45 seconds 99% CI [25.27, 33.64]. Ghostery increased load-time by approximately 72.35 seconds 99% CI [67.35, 77.36]. The control group was the fastest website to load out of the data set, 79.72% of the time. This was followed by uBlock Origin which was fastest to load 9.95% of the time, and then Disconnect and Ghostery which were the fastest to load 7.44% and 2.88%, respectively.

Disconnect reduced the number of cookies loaded in comparison to the control by approximately 22, 99% CI [-25.127, -20.473]. uBlock Origin reduced the number of cookies loaded in comparison to the control by approximately 30, 99% CI [-33.75, -27.25]. Ghostery reduced the number of cookies loaded in comparison to the control by approximately 33, 99% CI [-35.4, -31.2].

Ghostery, when compared to uBlock Origin, had a statistically significant higher load-time by approximately 42.90 seconds, 99% CI [38.21, 47.59]. With respect to the number of cookies blocked, Ghostery was found to block approximately 1 additional cookie, 99% CI [.331, 1.129]. When comparing the load-time to the amount cookies blocked, there is a weak positive correlation (0.02 R-squared) for all three blockers. This implies that the decrease in time efficiency is not necessarily caused by the number of cookies blocked. Since the p-values returned were all 0.000 for each two-sample t-test, a skewness analysis took place to determine if the dataset was sufficiently representative of the population, of which we suspected it would not be (5,000 vs all sites on the world wide web). In the event of a properly sized sample, we would observe a normal distribution in the sample according to the Central Limit Theorem of statistics. Following the skewness analysis, we find that the vast majority (approximately 1,690 sites have between 1 and 12 cookies).

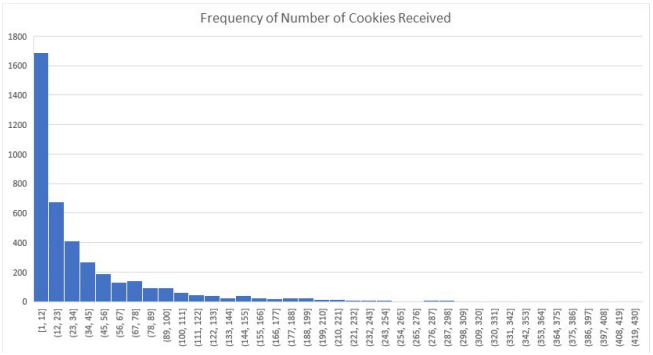

Figure 4.4 - Frequency of Number of Cookies Received (Control)

Since this data is not normally distributed, best practice would be to take a random sampling of available internet sites such that the distribution becomes normally distributed. Another alternative would be to take small, randomly selected samples from our collected sample and to use them for targeted analysis. Since time-constraints and hardware limitations were present that allowed us to crawl approximately 5,000 websites, it can be safely assumed that a more powerful, robust, and timeless experiment could be designed to acquire a dataset that is much more representative of the cookies found on the internet as a whole.

Overall, the data is still applicable for the majority of users who rarely interact with a website outside of the top 5,000 URLs present on the Alexa 1 million, but the significance levels of our findings should be interpreted and extrapolated quite conservatively.

Less than 5% of the time we received more cookies with an extension enabled for a particular website than the control group received for that same site. We hypothesize why this may be the case in the discussion

section of this paper and suspect strongly that this is being accomplished by potentially-malicious domains scanning the browser for particular extensions, utilizing JavaScript cookies, a technique known as "JavaScript Hacking".

Ranking by Avg. Block Rate	TLD	Average Block Rate
1	gt	98.60%
2	il	95.67%
3	nz	95.47%
30	net	75.45%
35	uk	73.90%
36	com	72.98%
59	org	66.29%
106	edu	40.61%
108	gov	43.68%

Figure 4.5 - Percentage of Cookies Blocked by Top-Level Domain

Within the 182 top-level domains (TLD) used by Alexa's 5000, all three blockers are most effective against websites with unpopular TLD, such as nz, tr and cz, reaching up to a 98% blocking rate. For more popular TLD, such as com and net, the blocker's average effectiveness is significantly reduced. Even when broken into specific TLD, the effectiveness of each individual blockers are still consistent with the aggregated data.

5. Discussion

It was concluded by the authors of the OpenWPM framework that news sites generally had more web trackers (Englehardt, S., & Narayanan, 2016). Although this was not the focus of this paper, our conclusions in some respects were similar. For example, when the analysis was conducted on a sample of 5 popular news site, we found that these sites exceeded the average number of cookies for the top 5000 sites by more than three times. It follows that news website would gain a significant amount of information from understanding where its visitors browse-to after visiting their website, and which websites they come from. This information could be valuable to both news-distribution sites (for their own research) as well as social media sites which are concerned in recent years with combating fake news reports being posted on their sites. In this context, we begin to see the value in third-party tracking. Third-party tracking provides a significant value-proposition to webmasters which allows them to gain substantial compensation in exchange for the information treasure troves that their sites have become to other organizations. The only entity that loses out in this proposition is the user who has his or her private data tracked, revealed, and sold to the highest bidder. It seems to follow that those sites that provide a service of *free information* have little in the way of funding acquisition from that same business model. They are pigeon-holed by their stakeholders to generate revenue; this, in turn, may drive incentive to implement extensive and invasive web-tracking on their sites. Similar behavior is seen by medical-information sites such as WebMD; private information about who and what ails a user is extensively collected and sold for profit (Libert, 2015).

In light of these potential privacy harms, it is the collective opinion of the authors of this paper that it is often difficult for a consumer to pick a privacy tool to rely upon, even when the need for such a tool is clear. This difficulty may arise from the challenge in assessing which tool is clearly an objective "best". Individual behavior online is incredibly unique; this is highlighted in the accuracy and precision of simple analytics techniques such as browser-fingerprinting. A

word of advice from a trusted confidant is a far cry away from scientific objective consensus. We should rely on cumulative statistical data, relying on objectively measurable key performance indicators to determine which extensions are "best" and provide recommendations to consumers. The two key performance indicators that we measured in this study were the number of cookies loaded and the load-time of a page, both on a per-URL basis. Not only did our findings make some fair headway in providing a recommendation to the consumer, it also suggested additional research that can be done to provide an even more refined analysis of the extensions available on the marketplace as of date. Although there existed limitations to our data gathering capabilities, our research was overall successful in providing a first-glimpse into potential recommendations that can be made by policymakers and privacy-advocates. Understanding which extensions the marketplace offers that reduce the maximum number of third-party trackers while simultaneously maintaining quick page-load speeds is a good first step. Additional steps that can be taken are discussed in the future work section of this paper. This includes doing analysis on extensions which utilize algorithmic tracker-blocking rather than solely the blacklist-blocking extensions that our paper explored; future work should also crawl a higher number of sites, perhaps the Alexa top 1 million, and crawl each site in-depth and multiple times.

During our analysis, it was noted that less than 5% of the time, a website would successfully introduce more cookies into the browser with an extension installed than was present in the control group. We hypothesized that this may be because of "Javascript hacking", the process of loading a Javascript element on the page that checks which extensions are installed on the browser and reports back to the site to have more, or different, cookies loaded in the response. It is possible that these particular sites approximately 200 in total (control group) utilized this technique; however it would take further investigation, and likely further automation to crawl these particular sites and look for Javascript like "/manafest.json" which is where unique extension IDs are located and can be scraped by malicious JS code (Kotowicz, 2012). There were more significant, less technical limitations that were experienced during the course of the project as well.

Limitations

While load-time is an effective measurement for efficiency, this measurement is subjected to internal factors that are costly to control. These limitations include physical hardware, network speed, and routing. External factors such as server load, weather and maintenance can have a strong effect on load-time and cannot be easily controlled.

Our sample size was a limited representation of the World Wide Web. While these 5000 sites are the most frequency visited in the world, the sample is too limited for analysis to extrapolate the implications of our findings to the web was a whole.

Our research did not consider the ease of use for each of the web tracking blockers. While their effectiveness is important, consumers must install these extensions onto their browser to provide a positive browsing experience. To better understand their privacy impact on web tracking, a holistically study should evaluate the accessibilities of these blockers.

Future Work

Extensions onto the work done in this paper are quite possible. For one, a larger sample size could be obtained from the Alexa top 1 million sites, which would form a more thorough representation of the world wide web's advertisement and web-tracking market-penetration. Secondly, the current top 5,000 sites can be sliced into key areas of concern (e.g. site category, popularity, search result rank) to reveal potentially un-observed trends. Future research may also include more extensive visits to crawled sites.

This may be in the form of crawling the individual site for it's most frequently occurring links, and traversing to those sites instead of only the root domain.

From an application implementation standpoint, the crawl could be performed on more browsers, extensions, and operating systems. Researchers may also consider performing the analysis using a combination of extensions to perhaps find a "sweet-spot" with implementation due to the currently tested ad blockers utilizing blacklists. Due to the ad blockers that were tested utilizing a blacklist architecture, it would also be worthwhile to test algorithmic architecture ad blockers such as Privacy Badger. A thorough comparison of all that is previously mentioned would be worth the effort, but to create a proper methodology coordinating the operating systems, browsers, and extensions would be an arduous task.

6. Conclusion

A world with ubiquitous surveillance deserves accessible knowledge to inform users on methods to avoid surveillance. In order to properly educate the masses on the most effective tools for combatting third-party tracking and advertisement, objective research should first be conducted. This research should also be ongoing as the nature of privacy, tracking, surveillance, and advertisement are ever changing in a technological world. Not only does our research replicate a small subset of the research done by the creators of the OpenWPM framework, but it also provides a number of suggestions as to what can be done for future research, as well as how the existing research can be expanded upon.

The initial results of our experimentation reveal a strong correlation between the implementation of *any* cookie-blocking software and a reduction of installed cookies. However, our research has yet to show a statistical significance between types of browser extensions. Additionally, we did not test a variety of browsers due to technical limitations, but if this work were to be expanded it would be highly recommended by the researchers to do so. Further research into the domain of JavaScript hacking in order to determine browser extensions is suggested and may prove to be useful in the privacy arms-race. Finally, repetition of the experiment that was performed would shore up the conclusions drawn by these authors, as the Alexa top 1 million sites change fairly frequently and thoroughly.

As consumers are bombarded by a deluge of more and more advertisements each year, they are becoming more privacy-aware and oriented. Motivation is not enough, we must arm the consumers with the capability to make meaningful and effective choices in the actions they take to fight browser-fingerprinting, third-party cookies, and JavaScript hacking. A further extension of this research would do much to accomplish this goal.

The project repository can be found at: https://www.github.com/rmccarth/extEval.

The OpenWPM framework can be found at: https://github.com/mozilla/OpenWPM.

7. Bibliography

Acar, G., Juarez, M., Nikiforakis, N., Diaz, C., G ¨urses, S., Piessens, F., and Preneel, B. FPDetective: dusting the web for fingerprinters. In *Proceedings of CCS*. ACM, 2013

Agarwal, L., Shrivastava, N., Jaiswal, S., and Panjwani, S. (2013). Do Not Embarrass: Re-Examining User Concerns for Online Tracking and Advertising. SOUPS 2013. Retrieved from http://cups.cs.cmu.edu/soups/2013/proceedings/a8_Agarwal.pdf

Alreck, P. L., & Settle, R. B. (2007). Consumer reactions to online behavioural tracking and targeting. Journal of Database Marketing & Customer Strategy Management, 15(1), 11-23.

Associated Press. (2018, September 14). Apple, Mozilla Web Browsers Thwart Facebook, Google Ad Tracking. Retrieved from CBS San Francisco: https://sanfrancisco.cbslocal.com/2018/09/14/apple-mozilla-browsers-thwart-facebook-google-web-tracking/

Balebako, R., Leon, P., Shay, R., Ur, B., Wang, Y., & Cranor, L. (2012). Measuring the Effectiveness of Privacy Tools for Limiting Behavioral Advertising. Retrieved from http://lorrie.cranor.org/pubs/EffectivenessBA.pdf

Boumans, W., & Poll, E. (2017, April 7). Web Tracking And Current Countermeasures. Retrieved from https://www.cs.ru.nl/bachelors-theses/2017/Willem_Boumans___4337166___Web_tracking_and_current_countermeasures.pdf

Cahn, A., Alfeld, S., Barford, P., & Muthukrishnan, S. (2016). An Empirical Study of Web Cookies. Retrieved from http://pages.cs.wisc.edu/~pb/www16_final.pdf

Disconnect. (n.d.). Disconnect. Retrieved from https://disconnect.me/disconnect

Earnshaw, Aliza. (2007, November 1). WebTrends Sale Not Imminent. Bizjournals.com, www.bizjournals.com/portland/stories/2007/10/29/daily39.html.

Englehardt, S. & Narayanan, A. (2018, November 27). Mozilla/OpenWPM. Retrieved from https://github.com/mozilla/OpenWPM

Englehardt, S., & Narayanan, A. (2016). Online Tracking: A 1-million-site Measurement and Analysis. Retrieved from http://randomwalker.info/publications/OpenWPM_1_million_site_tracking_measurement.pdf

Falahrastegar, M., Haddadi, H., Uhlig, S., & Mortier, R. (2014). Anatomy of the Third-Party Web Tracking Ecosystem. Cornell University Library. https://arxiv.org/abs/1409.1066

Fortune. (2018). Fortune 500 Companies 2018: Who Made the List. Retrieved from http://fortune.com/fortune500/list/

Gale Group. (2011, August 15). A brief history of the internet. Retrieved from http://go.galegroup.com.proxy.library.cmu.edu/ps/i.do?p=AONE&u=cmu_main&id=GALE|A266346805&v=2.1&it=r&sid=AONE&asid=25207400

Gale Group. (2016, September 4). UW computer scientists reveal history of third-party web tracking. Retrieved from http://link.galegroup.com.proxy.library.cmu.edu/apps/doc/A461803143/AONE?u=cmu_main&sid=AONE&xid=091f438f

Ghostery. (n.d.). Ghostery Makes the Web Cleaner, Faster and Safer! Retrieved from https://www.ghostery.com/

Hanson, M., Lawler, P., & Macbeth, S. (2018, May). The Tracker Tax: The Impact of Third-party Trackers on Website Speed in the United States. Retrieved from Ghostery: https://www.ghostery.com/wp-content/themes/ghostery/images/campaigns/tracker-tax/Ghostery_Study_-_The_Tracker_Tax.pdf

Hellman, E. (2018). FieldReports| privacy with google analytics. Library Journal, 143(5), 14. Retrieved from https://search-proquest-com.proxy.library.cmu.edu/docview/2014338939?accountid=9902

Kotowicz. (n.d.). Intro to Chrome addons hacking: Fingerprinting. Retrieved from http://blog.kotowicz.net/2012/02/intro-to-chrome-addons-hacking.html

Kudryavtseva, L. (2017, May 31). *"Sing, Goddess, the wrath": a history of ad blocking, part one*. Retrieved from Ad Guard: https://adguard.com/en/blog/adblocking_history_1/

Leon, P.G., Ur, B., Wang, Y., Sleeper, M., Balebako, R., Shay, R., Bauer, L., Cristodorescu, M., and Cranor, L.F. (2013). What Matters to Users? Factors that Affect Users' Willingness to Share Information with Online Advertisers. SOUPS 2013. Retrieved from

http://cups.cs.cmu.edu/soups/2013/proceedings/a7_Leon.pdf

Libert, T. (2015, March). *Privacy Implications of Health Information Seeking on the Web* [PDF]. Philadelphia. https://timlibert.me/pdf/Libert-2015-Health_Privacy_on_Web.pdf

Mayer, J., & Mitchell, J. (2012). Third-Party Web Tracking: Policy and Technology. Retrieved from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6234427&isnumber=6234400&tag=1

Melicher, W., Sharif, M., Tan, J., Bauer, L., Christodorescu, M., & Leon, P. G. (2016). (Do Not) Track me sometimes: users' contextual preferences for web tracking. Proceedings on Privacy Enhancing Technologies, 2016(2), 135-154. https://www.researchgate.net/publication/291418720_Do_Not_Track_Me_Sometimes_Users'_Contextual_Preferences_for_Web_Tracking

MIT Technology Review. (2016, January). The Fast Rise of Ad Blockers. MIT Technology Review, 119(1), 20. Retrieved from http://search.ebscohost.com.proxy.library.cmu.edu/login.aspx?direct=true&db=buh&AN=112089611&site=ehost-live

MIT Technology Review. (2014, September 19). The Murky World of Third Party Web Tracking. Retrieved from https://www.technologyreview.com/s/530741/the-murky-world-of-third-party-web-tracking/

Nicolaidou, I. & Georgiades, C. (2017). The GDPR: New Horizons, in *EU Internet Law, Regulation and Enforcement*.

PrivacyTools.io. (n.d.). Privacy Tools | Encryption against global mass surveillance. Retrieved from https://www.privacytools.io/#browser

uBlock. (n.d.). uBlock - A Fast and Efficient Ad Blocker. Easy on CPU and Memory. Retrieved from https://www.ublock.org/

Ur, B., Leon, P., Cranor, L., Shay, R., & Wang, Y. (2012). Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising. Retrieved from http://cups.cs.cmu.edu/soups/2012/proceedings/a4_Ur

Wall Street Journal. (2001, Apr 23). E-commerce (A special report): Openers --- great expectations: Some high and low points in the history of online advertising. Retrieved from https://search-proquest-com.proxy.library.cmu.edu/docview/398806273?accountid=9902

Webtrends. (1999). The IPO Reporter, 1-13. Retrieved from https://search-proquest-com.proxy.library.cmu.edu/docview/222507865?accountid=9902