Quantifying Web Adblocker Privacy

Arthur Gervais[1(✉)], Alexandros Filios[1], Vincent Lenders[2], and Srdjan Capkun[1]

[1] ETH Zurich, Zurich, Switzerland
arthur.gervais@inf.ethz.ch
[2] Armasuisse, Thun, Switzerland

Abstract. Web advertisements, an integral part of today's web browsing experience, financially support countless websites. Meaningful advertisements, however, require behavioral targeting, user tracking and profile fingerprinting that raise serious privacy concerns. To counter privacy issues and enhance usability, adblockers emerged as a popular way to filter web requests that do not serve the website's main content. Despite their popularity, little work has focused on quantifying the privacy provisions of adblockers.

In this paper, we develop a quantitative framework to compare the privacy provisions of adblockers objectively. For our methodology, we introduce several privacy metrics that capture not only the technical web architecture but also the underlying corporate institutions of the problem across time and geography.

Using our framework, we quantify the web privacy implications of 12 ad-blocking software combinations and browser settings on 1000 websites on a daily basis over a timespan of three weeks (a total of 252'000 crawls). Our results highlight a significant difference among adblockers regarding filtering performance, in particular, affected by the applied configurations. Our experimental results confirm that our framework provides consistent results and hence can be used as a quantitative methodology to assess other configurations and adblockers further.

1 Introduction

Online advertising provides a viable way to support online businesses that offer content free of charge to their users, such as news, blogs and social networks. To achieve targeted and hence more effective advertising however, advertisers and tracking companies record user browsing behavior, e.g. pages viewed, searches conducted, products purchased [8,14,20,25,33]. Such techniques are known as *online profiling* and have raised significant privacy concerns because online user profiles can be used to infer private sensitive information and user interests [10,11,23,27].

Adblockers aim to improve the user experience and privacy by eliminating undesired advertising content, as well as preventing the leakage of sensitive user information towards third-party servers. The most well-known adblocker solutions are browser extensions such as *Ghostery* or *Adblock Plus* which suppress unnecessary requests to third-party advertisements and tracking servers, thereby

S.N. Foley et al. (Eds.): ESORICS 2017, Part II, LNCS 10493, pp. 21–42, 2017.
DOI: 10.1007/978-3-319-66399-9_2

limiting the risk of data leakage towards these servers. Recently, we have experienced a proliferation of adblocker browser extensions in the wild which might be due to users' privacy concerns and awareness about online profiling as well as due to the increasingly intrusive advertisements. According to Mozilla and Google usage statistics [2,4], already more than thirty million surfers are actively using a browser with the Adblock Plus extension enabled. In a recent measurement study [29], researchers show that 22% of the most active users are using the Adblock Plus adblocker while surfing the Web.

Despite the popularity of adblocking tools, surprisingly little research has been performed to understand how well adblocking actually improves the privacy of its users. While the methods employed in advertisement and tracking and their privacy implications have been well researched in the literature [13,19,21,28], the protection that adblockers offer, has not been investigated that much in the literature. Works such as [9,18,29,32] analyze ablockers's performance, however the impact of user privacy is not in the main scope of these studies, as they focus on the effectiveness of the adblocker's implementations and the usage in the wild. Understanding how adblockers affect user privacy is fundamental to their use, because it not only provides feedback to the users, but also helps at correctly using and configuring those systems. Adblockers rely on complex filter configurations in the form of blacklisted URLs and regular expressions, and as we show in this paper, existing adblockers are not necessarily configured by default to provide the best privacy protection to their users.

Our goal in this work is to define a framework and associated metrics to assess the web privacy level that web adblockers provide. We address this problem by developing a quantitative model to compare adblocker filtering performance across various privacy dimensions. Our model includes simple count metrics to third-parties, but also considers more advanced metrics on the level of organizations (legal entities) and countries as well as their relationships. Our primary aim is to provide a methodology, not to conduct a large-scale analysis. While related work with large-scale analysis has focused on one time instant, we provide a daily analysis over several weeks to understand the temporal dynamics of the web.

We have developed a testbed system which allows us to repetitively browse the same Web sites in a systematic way and classify the number of HTTP requests that go to first and third parties without any classification errors. We evaluate 12 different browser profile configurations in our testbed, capturing different adblocker instances and combinations of desktop/mobile user client agents. During three weeks, we surfed on a daily basis the Alexa's top 500 global sites and 500 randomly selected sites and analyzed how different configurations influence these privacy metrics. Because our primary goal is to define the framework to assess the privacy level, we do not attempt to test all extensions, nor to distinguish among anti-tracking tools and adblockers. Our configurations confirm that the metrics provided consistent results and hence can be used to assess other configurations and adblockers further.

Our results show that the usage of adblockers provides a significant improvement in terms of user privacy. However, the degree of protection is highly depending on the configuration. For example, by default Ghostery does not block any third-party requests and Adblock Plus still allows a significant amount of requests to third parties. These results are consistent for the desktop and the mobile user agents. When increasing the level of protection in Ghostery and Adblock Plus however, these tools manage to effectively suppress requests to third-parties and thus improve the privacy. Except for Google Inc. which still receives around 50% of third-party requests because it hosts relevant content not related to advertisement and tracking, the amount of third-party requests towards the other top ten companies in our experiments is only 2.6% of the total amount that would result when surfing without an adblocker.

Our contributions in this paper can be summarized as follows:

- We provide a quantitative methodology to objectively compare the filtering performance of web adblockers.
- We capture the temporal evolution of adblocker filtering performances and study the differences between mobile and desktop devices, as well as the impact of the *do not track* header. Our methodology further allows to measure the influence of other parameters (e.g. third-party cookies) on adblocker filtering performance.
- Beyond the domain of the third parties, our model takes into account the underlying legal entities, their corresponding geographical locations as well as their relationships.
- Using our model, we quantify the privacy of 12 different adblocker browser profile configurations over 1000 different Web sites for repetitive daily measurements over the duration of 3 weeks and discuss the implications in terms of user protection.

The remainder of the paper is organized as follows. In Sect. 2 we illustrate the objective and functionality of adblockers, while in Sect. 3 we outline our privacy metrics. Section 4 outlines our methodology, Sect. 5 discusses the experimental setup and the results. Section 6 presents the related work and Sect. 7 summarizes our work.

2 Web Tracking and Adblockers Background

This section provides relevant background on third-party tracking in the web and how adblocker browser extensions aim at improving user experience and privacy.

2.1 Third-Party Tracking

When visiting an HTTP-based website on a domain (commonly referred to as first party), the web browser sends an HTTP request to the first-party server that hosts the website and loads the content of the first-party domain. The HTML code of the first party is then able to trigger (without the awareness of the user) further HTTP requests to remote servers (commonly referred to as third parties) in order to load further resources that they host. External resources vary in their format and are applied with different objectives, such as the inclusion of external libraries—e.g. jQuery—that are indispensable for the functionality of the website itself. Further reasons include the promotion of advertising content that can be externally loaded and placed at a pre-allocated space on the website.

This third-party content loading mechanism clearly facilitates the development and deployment of dynamic websites because it allows to use different content providers to load resources that do not need to be served from the first party. However, as shown in previous works [8,20,25,33], HTTP requests to third-parties lead to severe privacy implications because third parties can follow the activity of the users and reveal the pages they are looking at while surfing the web. For example, it has been shown in [14] that dominant players in the market such as Google Inc. are embedded as third-parties in so many web sites that they can follow 80% percent of all web activities. Since the web page content and thus user interests can be inferred by the uploaded requests to the third parties, personal profiles of users can easily be derived and potentially used to discriminate people or spy on their interests and habits without getting noticed by the users.

2.2 Adblocker Browser Extensions

To address the aforementioned implications and challenges, numerous software and hardware-based solutions—commonly referred to as *adblockers*—have been proposed in order to remove or alter the advertising and third party content in a web page. Although there exist multiple ad-blocking methods (e.g. DNS sinkholing, proxies run by internet providers (externally) or by an application on the same client machine, special hardware) we focus in this work on one of the most popular solutions: browser extensions, such as *Ghostery* and *AdblockPlus.*

Adblocker browser extensions use one or more lists that describe the content that is to be allowed (whitelists) or blocked (blacklists) and update those on a regular basis. There are two principal methods how adblockers apply these lists to remove ads/third parties from a web page: One is filtering the resource according to the result of an URL-pattern matching, before this resource is loaded by the web browser. The second consists in hiding loaded content with the use of CSS rules (*element hiding*) within the HTML content. In terms of privacy, filtering the resources before they are requested by the browser is the only effective method because these requests are the ones revealing the activity of the users.

Adblocker browser extensions are very popular by users today and their popularity is continuously on the rise [2,4,29]. However, content providers and advertisers see this trend as a risk to their own business models because they regard the application of theses tools as a way for the comsumers to evade "paying for the content". Juniper Research estimates that digital publishers are going to lose over 27 billion dollars by 2020 due to the use of ad blocking services [3]. There is therefore high pressure by these industries on the developers of adblockers to not blacklist their services. For example, Adblock Plus has introduced in 2011 the concept of "non-intrusive advertising", which basically allows third-party advertisements for ads which do not *disrupt the user's natural reading flow* [1]. However, these practices raise concern in terms of privacy because non-intrusive advertisement services may well perform intensive tracking without falling in this category. We therefore argue that it is important to quantify independently the privacy of these tools as we do in this work.

3 Privacy Model and Metrics

In this section, we introduce our privacy model and the metrics we use in order to quantify the privacy provisions of adblockers.

3.1 Threat Definition

A key issue for a threat model in adblocking is to define which third-parties should be considered as a privacy threat to users. In this work, we consider all third-parties as potential threats irrespective of the type and content of the queries towards these third parties. This approach may arguably seem conservative, but it is practically impossible to exclude for sure any third-party from performing tracking and/or profiling given the multitude of possible mechanisms that are available and continuously invented for fingerprinting and tracking user behavior in the web.

In our notion, the privacy objective of the abblocker is therefore to reduce as many requests as possible towards third parties. Notice here the difference of our threat model definition to the slightly different objective that adblockers such as Adblock Plus have. By default, Adblock Plus aims at improving user satisfaction by minimizing the display of intrusive advertisements which annoy the users while third-party requests to non-disturbing advertisements and tracking services for commercial purposes are considered to be acceptable [1].

3.2 User Tracking Model

We model the tracking of a user U through third parties as undirected graph $G = (E, V)$, where E are edges, and V vertices. A vertex V_S represents a web domain and is connected to another vertex V_T through an edge E, if and only if at least one request has been sent from V_S to V_T. In that case, V_S is the *source* of the request and V_T the *target* of the request.

In the following, we use the term *third-party request* (TPR) to denote the requests that are sent to a target domain T that differs from the source domain S and corresponds to a graph edge E between the nodes V_S and V_T. On the contrary, the requests whose source and target coincide are designated as *first-party requests* (FPR) and are not taken into consideration for the construction of G, since no information leaks to third parties and hence they do not bring about further risks for user's privacy[1]. The source and the target domain are referred to as *first-party domain* (FPD) and *third-party domain* (TPD) and correspond to FPD and TPD graph nodes, V_S and V_T, respectively.

Compared to previous works on third-party traffic characterization [9,13], we augment G by incorporating the ownership of third party domains to their corresponding legal entities, i.e. the organizations who own the different TPDs. Two TPD, belong to the same legal entity if they are registered to the same organization (e.g., doubleclick.net and google-analytics.com both belong to Google Inc.) and are thus combined into one vertex, resulting in a hierarchical graph (cf. Fig. 1). Considering the information flow of third-party requests towards legal entities is particularly important for the scope of privacy because legal entities which own multiple domains can fuse the information they collect from their different domains in order to increase their tracking and profiling coverage, thus resulting in a higher privacy threat to the users.

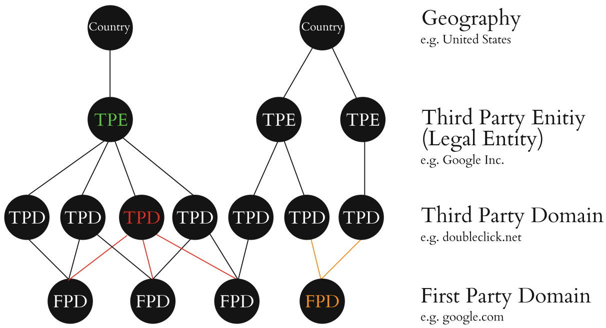


Fig. 1. Graphical representation of our user tracking model. The colored third-party domain (TPD) has a node degree of 3, the colored first-party domain (FPD) has a node degree of 2. The colored third-party entity (TPE) spans all its child TPD nodes and hence has a degree of 3. (Color figure online)

Finally, we further attribute each legal entity to a geographical location (the country where the headquarter of the legal entity is situated) in order to model which countries govern the regulations over which legal entities. This geographical perspective is also of special importance to privacy, because most data privacy laws are specific to local laws of the countries, thus affecting the regulations that apply to the user data that is collected by the legal entities.

3.3 Privacy Metrics

Given the graph representation G of our user tracking model, we evaluate the respective privacy provisions based on the following metrics.

[1] Arguably, users also leak private information to first party domains when they visit and interact with those sites, however, since users are visiting these first parties deliberately, the privacy risks are known to the users and controllable without an adblocker.

Degree of First Party Domain. The degree of a FPD node of graph G refers to the number of TPDs that it has sent at least one third-party request to when loading the web page from the FPD. That is, the more edges a FPD node has—or, equivalently, the more third parties loaded by a first-party—the more third parties are able to track the web-browsing history of a user. The FPD node degree is a metric that is commonly used to evaluate the adblocker's performance [32]. However, it is alone not a sufficient metric to capture the impact on user privacy, as it does not represent the structure behind the relationships between FPD and TPD. The following metrics therefore aim at capturing these relationships.

Degree of Third Party Domain. The degree of a TPD node can be directly translated to the number of first-party websites that a particular third party exchanges information with and potentially tracks. Clearly, the more often a third party is accessed over the user's series of websites S_U, the less privacy the user experiences from this particular third party. To exemplify this statement, let's assume that a third party is requested by only one of the first-party websites S_U visited by U. This third party will in this case learn that the user has accessed the respective first party, but has a limited view of their browsing behavior. If the third party, however, is requested by over 80% of the user's visited websites, S_U, the third party will likely be able to recover up to 80% of the web behavior of U.

Degree of Legal Entity. Instead of focusing on domain degrees, the degree of a legal entity reflects the number of third-party domains that belong to a legal entity. Third-party domains such as doubleclick.net and google.com for example are both owned by the same entity Google Inc. Their collusion therefore seems more likely, and affects the privacy of a web user U more significantly, than if both were belonging to two different legal entities. By incorporating the legal relation among third party domains, we therefore capture a more realistic privacy leakage through user web surf activity.

Geographical Location. After having mapped the TPD's to legal entities, we further assign a geographical location to the TPD. This allows our model to capture the geographical distribution of the TPDs and thus infer which geographical countries have for instance the most TPD. The geographical location of a legal entity is defined by the country in which its headquarter resides. Alternatively, we could consider the particular location of the servers as derived from the IP address, but content retrieved from web services is often hosted on distributed caches and content distribution networks and hence the server IP address does not necessarily reflect the country to which the user data is finally sent to. By choosing the headquarter's location, we thus aim at modelling the country in which the privacy laws and regulations will apply to the user data as collected by the third-party.

Graph Density. In addition to the degree metrics outlined above, we consider a metric based on the graph density of G. Since an edge on the graph G represents a partial tracking relationship between a third and a first party, we expect that the denser the graph G, the more information can be retrieved by third parties/can leak to third parties with respect to the browsing behavior of the user. We observe that the more dense G is, the more third parties are likely able to track the user U. The graph density therefore allows to reason about the possible privacy improvements by the respective ad-blocking software. We rely on a common definition of the graph density as:

$$D = \frac{2|E|}{|V|(|V|-1)} \tag{1}$$

Note however that we cannot achieve the maximum density of 1, because the first parties in G are not directly connected (cf. Definition in Sect. 3.2).

4 Evaluation Methodology

In order to compare the privacy of different adblockers, as well as the influence of different browser settings on their adblocking efficiency, we create different browsing configurations without adblockers, with the Ghostery, and with the Adblock Plus browser extenstions installed in the Firefox browser.

4.1 Considered Browser Profiles

All our experiments are performed on "Linux (Release: Ubuntu 14.04.4 LTS, Version: 4.2.0-35-generic GNU/Linux)" with the version 45.0.1 of the Firefox browser. For Ghostery, we use the browser plugin version 6.1.0 and for Adblock Plus the plugin version 2.7.2. The different protection levels, *Default* or *Max-Protection*, for the two adblockers *AdblockPlus* and *Ghostery* respectively, are achieved through the use of a different combination of blacklists. AdblockPlus and Ghostery store their respective blacklists in the form of URL and CSS regular expressions. The blocking options of AdblockPlus are set through the direct inclusion of blacklists to be applied, while Ghostery's blacklist configuration consists in the selection among a multitude of tracker categories to be blocked. An overview of these configurations is presented in Table 1.

Table 1. AdblockPlus blacklist combination for default and maximal protection level. Ghostery's default and maximal protection correspond to the selection of none and all tracker categories, respectively.

Protection level	Lists			
	AdServers	EasyList	EasyListChina	EasyPrivacy
Default		✓		
Maximal	✓	✓	✓	✓

Modern web browsers such as Firefox further allow to set the *do not track* HTTP header option, to express their personal preference regarding tracking to each server they request content from, thereby allowing recipients of that preference to adjust tracking behavior, accordingly [31]. It remains the sole responsibility of the web server to respect the request of its clients. Almost 10% of the Firefox users have enabled this option on their desktop browsers in 2014 [5]. In order to evaluate to which extend the DNT header has an influence on our proposed metrics we as well include the DNT option in our evaluation.

The usage of mobile devices for web browsing has recently witnessed a steady growth [6]. As a consequence, an ever increasing number of websites has been adapting to the demands of the mobile user agents. Because of the dimensions and the reduced-bandwidth requirements of the mobile devices, the structure and content of the web pages has to be adjusted accordingly and the advertising content could not remain unaffected by these limitations. To investigate the effects of user agents from a privacy-related perspective, we consider this parameter in the design of the experimental evaluation and evaluate several mobile-device instances by setting the HTTP header *User-Agent* accordingly.

Faking a user agent is trivially detectable. Even if some hosts, however, recognized this fact, our results indicate a clear difference between the privacy levels for desktop and mobile user agents and confirm the validity of our metrics, which is the primary goal of our study.

Based on above mentioned criteria, we create 12 browser profiles, U as described in Table 2. Each configuration is defined as a combination of the following parameters:

– Adblocker: No adblocker, Ghostery, or Adblock Plus
– Block policy: maximum or default protection

Table 2. Overview of browser profiles examined

Browser profile	Adblocker	Block policy	DNT	User agent	Legend
Ghostery_Default	Ghostery	Default	No	Desktop	
Ghostery_MaxProtection	Ghostery	Max	No	Desktop	
Adblockplus_Default	AdblockPlus	Default	No	Desktop	
Adblockplus_MaxProtection	AdblockPlus	Max	No	Desktop	
NoAdblocker	None	-	No	Desktop	
NoAdblocker_DNT	None	-	Yes	Desktop	
Ghostery_Default_MUA	Ghostery	Default	No	Mobile	
Ghostery_MaxProtection_MUA	Ghostery	Max	No	Mobile	
Adblockplus_Default_MUA	AdblockPlus	Default	No	Mobile	
Adblockplus_MaxProtection_MUA	AdblockPlus	Max	No	Mobile	
NoAdblocker_MUA	None	-	No	Mobile	
NoAdblocker_DNT_MUA	None	-	Yes	Mobile	

– User agent: mobile or desktop
– Do Not Track (DNT): header enabled or disabled

Throughout the remaining of the paper, we use the following conventions for each browser profile U (cf. Table 2):

– The *color* denotes the adblocker installed.
– The *line width* indicates the protection degree—i.e. default, maximum protection or DNT header.
– Profiles with Mobile User Agent are plotted in *dashed lines*.

4.2 Experimental Setup

The distinction between FPRs and TPRs is crucial in our attempt to precisely quantify the filtering capability for each browser profile, since they define the exact topology of the derived graph G. Passive classification of HTTP requests into first-party and third party requests is not a trivial task given the complex and dynamic structure of Web pages [29]. For this reason, we rely in this work on an active approach in which we collect our own synthetic web surfing traffic with automated web surfing agents. To create a realistic and representative dataset, the agents visits Alexa's top 500 web sites (the 500 domains with the highest incoming traffic in the web) and 500 web sites which are sampled uniformly among Alexa's top 1 million most-visited domains. The motivation for including less popular web sites is to avoid the risk of favoring an adblocker optimized to perform best for the most popular web sites, eventually biasing the experimental results. The overall sample set S of 1000 URLs is retrieved once and kept unchanged throughout the evaluation period, so as to de-correlate any variations of the results between different days.

Since nowadays most web applications are based on asynchronous calls to fetch data, it is insufficient to wait for the DOM to finish rendering to record all resource requests sent from the website to any first or third parties. To collect the complete data and better evaluate the common user browsing behavior, our agent therefore waits 20 s on each website of our sample set S and records any requests sent, before closing and proceeding to the next domain. We visit the same set of web sites every day during three weeks from 28/04/2016 until 19/05/2016. To decouple the experimental conditions from the influence of any time- or location-related effects—i.e. variations of the served content, locale-based personalization—all browser profiles U execute the same crawling routine simultaneously, whilst running on the same machine, thus behind the same IP address, browser and operating system. However, some of the instances are configured to send their requests with a User-Agent HTTP header that corresponds to a mobile device (iPhone with iOS 6[2]), in order to extend our observations for mobile users.

[2] User Agent: `Mozilla/5.0 (iPhone; CPU iPhone OS 6_0 like Mac OS X) AppleWebKit/536.26 (KHTML, like Gecko) Version/6.0 Mobile/10A5376e Safari/8536.25`.

In order to record all HTTP requests, we rely on the *Lightbeam* plugin. However in contrast to [32], we do not use Lightbeam to determine the source domain that a request is initiated from and to classify it accordingly as a FPR or a TPR because Lightbeam relies on heuristics that are too error-prone for our purpose. More precisely, the classification of Lightbeam is not always in accordance with our definitions of FPR and TPR, as introduced in Sect. 3. By examining the request logs after a complete crawl cycle and comparing the estimated source to the actual visited domain, two types of false-positive cases (cf. Table 3) arise in Lightbeam:

- **Unrecognized TPRs:** The request is mistakenly considered to be a FPR according to the Lightbeam heuristics, this way "hiding" a TPR edge from the graph.
- **Misclassified TPRs:** The request is correctly found to be a TPR, but not for the correct FPD node, i.e. the one corresponding to the actually crawled domain. The inaccuracy introduced to the graph results from the potential introduction of a bogus FPD node, as well as the false number of TPR edges starting from the correct and the bogus FPD nodes.

As results from the experimental evaluation on the data of one full crawl cycle (1000 visited first parties) and 12 different browser profiles, the misclassified and unrecognized TPRs make up for 2.0%–12.0% and 4.0%–11.0% of the total requests, depending on the respective browser profiles that we define in the following.

Table 3. Examples of misclassified and unrecognized TPRs

	Visited domain	Estimated source	Target
Recognized	wp.pl	wp.pl	facebook.com
Misclassified	wp.pl	facebook.com	fbcdn.net
Unrecognized	wp.pl	facebook.com	facebook.com

We thus modify Lightbeam to account for the currently visited first-party as a priori known by the agent which triggers page visits.

Note that we do not simulate an interaction with the website, e.g., mouse mouvements, scrolling or keystrokes and leave this for future work.

4.3 Classification of Domains to Legal Entities and Locations

We infer the legal entities' domains and locations by inspecting the WHOIS database. The WHOIS database provides information about the holders of Web domains. For each domain, we look up the legal entity that is registered as holder and the country of the holder's address. Note that only a part of the considered domains—accounting for about 60%—could be assigned to a legal entity and followingly to a country. One reason is that WHOIS does not provide sufficient information for all of the domains loaded. Moreover, our parser that allowed for the automated extraction of the entity information depends on a relatively uniform format of the WHOIS documents and as a result, deviations from this format causes information loss.

5 Evaluation

We examine the impact of the configuration parameters on the achieved privacy level using our privacy metrics from Sect. 3.

5.1 Effectiveness of Adblockers at Suppressing Third-Party Requests

Baseline Without Adblocking. Before investigating the effect of the different adblockers, we characterize the FPD node degree with the *NoAdblocker* and *NoAdblocker_MUA* browser profiles as a baseline. Figure 2 shows the cumulative distribution function (CDF) of the FPD node degree of both profiles on a single day (28/04/2016) for the top-ranked 500 domains and the 500 uniformly-selected ones. As can be seen, in both the top 500 and the uniformly selected domains, almost 20% of the websites did not load any third-parties at all. These domains do therefore not impose a privacy risks to the users. On the other hand, more than 80% of the visited domains generate requests to third parties. In general, we can say that the top 500 domains tend to generate more requests to third-parties than the uniformly selected domains, indicating that advertisement and tracking is more likely to happen on popular domains. However, even the randomly selected domains have a quite significant number of third-party requests. While the mean FPD node degree for the top 500 domains and uniformly selected domains are around 17 and 12 respectively, both FPD node degree distributions has a quite long tail. We observe a significant number of FPD node degrees above 100 with one domain in the top 500 exhibiting a degree of 180. These sites raise serious concerns in terms of privacy since each individual third-party request could potentially leak personal information of the visiting users to these third parties.

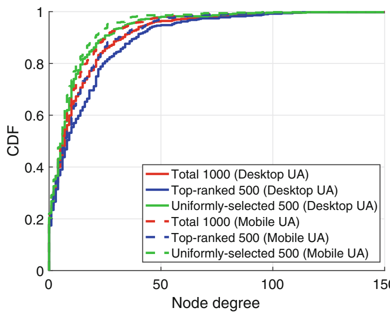
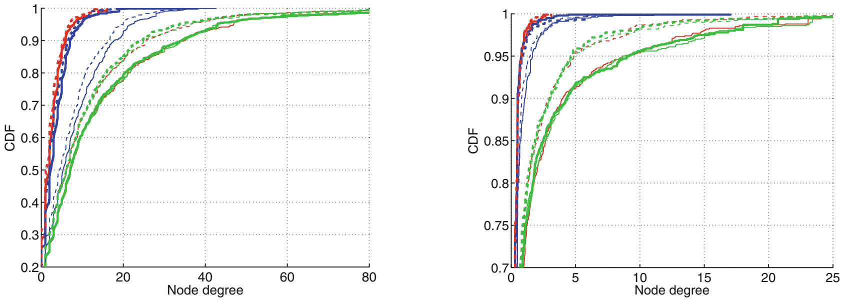


Fig. 2. FPD node degree for the browser profiles *NoAdblocker* (solid line) and *NoAdblocker_MUA* (dotted line) on 28/04/2016.

Comparison of the Different Browser Profiles. To understand the effectiveness of the different adblockers and browser profiles at suppressing requests to third-parties, we plot in Fig. 3 the FPD node degree distribution for all domains as a CDF. Figure 3a shows the node degree distribution averaged over the different days while Fig. 3b represents the standard deviation of the node degree over the same days. Our results indicate the following findings.



(a) CDF of the average node degree over 3 weeks.

(b) CDF of the standard deviation of the node degree over 3 weeks.

Fig. 3. FPD node degree distribution for all browser profiles. Legend is provided in Table 2.

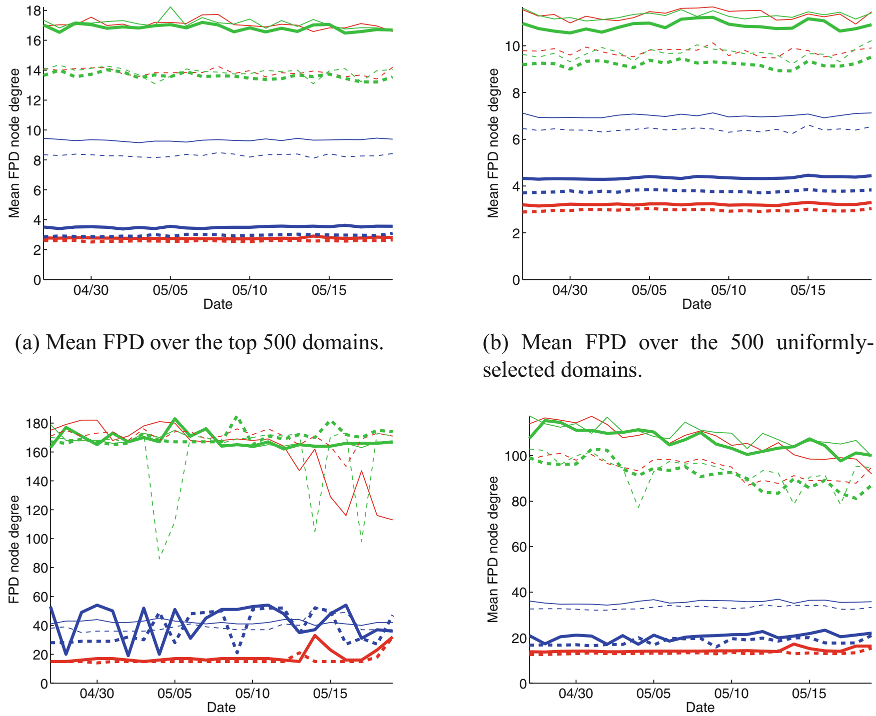
The worst filtering performance is achieved with the *do not track* HTTP header options (*NoAdblocker_DNT* and *NoAdblocker_DNT _MUA*) and Ghostery in default mode (*Ghostery_Default* and *Ghostery _Default_MUA*). With these browser profile configurations, almost none of the third-party requests are blocked. AdblockPlus (*Adblockplus_Default* and *Adblockplus_Default_MUA* with its default settings has a FPD node degree that is significantly lower than the aformentioned cases, i.e., the browser profiles with the DNT header enabled and Ghostery in its default configuration. Unsurprisingly, the browser profiles that filter the most third parties are those with adblockers configured to a maximum protection level. We observe that *Ghostery_MaxProtection* decreases the mean FPD node degree by approximately 80% compared to *NoAdblocker*. On the other hand, the FDP node degree of Adblock Plus (*AdblockPlus_MaxProtection*) is reduced by almost 75% which is slightly behind the performance of Ghostery, but still significantly better than the default configuration option.

Interesting to note here is the large difference in blocking performance between the different configurations of the same adblockers. This result suggests that the privacy of the users is highly affected by a good configuration of the tools and that by default, these tools still permit a significant portion of the third-party requests.

The standard deviation of the FPD node degree over all domains is shown in Fig. 3b. As we can see, the profiles which have a large FPD node degree tail such as *NoAdblocker*, *NoAdblocker_MUA*, *NoAdblocker_DNT*, *NoAdblocker_DNT_MUA*, and *Ghostery_Default* also exhibit this tail in the standard deviation. However, the profiles which tend to have a small FDP node degree feature a small standard deviation as well.

Temporal Dynamics. To capture the temporal dynamics of third-party requests, we plot in Fig. 4 the FPD over time in the considered period of 3 weeks. Figure 4a and b show the mean FPD for the top 500 domains and the uniformly selected domains respectively. We observe a quite stable temporal evolution over the individual days for both datasets. In particular, in none of the datasets, we can observe a change in relative order between the different browser profiles. We can therefore conclude that in general, the privacy of the users is not sensitive to web site or blacklist optimizations that happen at shorter time scales.

To check whether this conclusion also translates to individual domains, we take a closer look at the domains with the highest FPD in Fig. 4c and d. Figure 4c shows the evolution of the FPD for the domain with the highest FPD in any of the dataset while Fig. 4d represents the mean of the FPD over the ten domains



(a) Mean FPD over the top 500 domains.

(b) Mean FPD over the 500 uniformly-selected domains.

(c) Maximum FPD over all visited domains.

(d) Mean FPD over the 10 domains with the highest FPD from all visited domains.

Fig. 4. Evolution over time of the first party node degree (FPD). Legend is provided in Table 2.

with the highest FPD. We make two interesting observations here. First, the domains with the largest FPDs tend to exhibit a higher variation over different days. In particular, for *Ghostery_Default* and *Ghostery_Default_MUA* in Fig. 4c, the filtering of third-party requests shows a larger fluctuation over time. Also, *AdblockPlus_MaxProtection* and *AdblockPlus_MaxProtection_MUA* has a significantly higher fluctuation for the top domain than on average. Second, the filtering performance of the different browser profiles is more clustered than it is was on average for all the domains. For example, on most days, the performance of *Ghostery_Default* and *Ghostery_Default_MUA* is almost identical to *NoAdblocker*, while those two profiles where significantly outperforming the *NoAdblocker* profile in Fig. 4a and b. These two observations indicate that these domains with a high FPD score could be more active at circumventing blocking strategies by adblockers.

5.2 How do Adblockers Reduce the Tracking Range of Third-Party Domains?

In order to understand the extend to which individual third-parties are able to track users while surfing across different domains, we look next at the degree of third-party domains (TPD). The TPD degree reflects how many visits to different first-party domains an individual third-party can observe. We observe, that the TPD is highly skewed. Only 10% of the third-parties have a TPD of more than 10 for the *NoAdblocker* profile while the largest TPD degree we observe is 486 (*None* column of Table 4). In general, we can therefore say that a small number of third-party domains are able to capture the vast majority of the visits to first parties.

Table 4. Top-loaded TPDs for browser profile *NoAdblocker* and the corresponding values for Ghostery and AdblockPlus with maximum-protection settings (browser profiles *Ghostery_MaxProtection* and *AdblockPlus_MaxProtection*) on 28/04/2016

Third-party domain	Legal entity	TPD degree		
		None	Ghostery	AdblockPlus
doubleclick.net	Google Inc.	486	0	1
google-analytics.com	Google Inc.	476	4	0
google.com	Google Inc.	383	93	144
facebook.com	Facebook Inc.	318	5	164
gstatic.com	Google Inc.	308	226	235
googlesyndication.com	Google Inc.	204	0	0
google.ch	Google Inc.	189	0	0
fonts.googleapis.com	Google Inc.	185	145	141
adnxs.com	AppNexus Inc.	159	0	0
facebook.net	Facebook Inc.	157	0	140

Considering the effect of the different browser profiles, we observe a similar trend as for the FPD degree. The *Ghostery_MaxProtection* and *AdblockPlus_MaxProtection* profiles manage to effectively reduce the TPD node degree of all domains. However, in their default settings, AdblockPlus and Ghostery have only a noticeable effect on the domains with a small TPD degree, while these profiles have almost no impact on the filtering performance of domains with a large TPD node degree. Again, the browser profiles with the Do Not Track option enabled result in similar TPD node degrees as without the option.

In Table 4, we list the 10 domains with the highest TPD node degree (when no adblocker is applied) and compare how these numbers decrease with the *Ghostery_MaxProtection* and *AdblockPlus _MaxProtection* browser profiles. Ghostery achieves generally better performance, although AdblockPlus outperforms Ghostery slightly for two domains. Interesting to notice here is that some third-party domains from this list still exhibit a high TPD node degree with any of the adblockers enabled. These are the domains google.com, gstatic.com, and fonts.googleapis.com. These domains provide important content to render the web pages of the first parties and can therefore not be blocked. The other domains relate to advertisements, tracking, and social media and their TPD degrees are effectively reduced by Ghostery. AdblockPlus is not so effective at reducing the TPD degree of domains such as facebook.com and facebook.net.

5.3 How do Adblockers Reduce the Tracking Range of Legal Entities?

As we have seen in Table 4, the TPD degree of many domains was effectively reduced with adblockers, but some domains still remain with a high TPD node degree, mostly in order to provide useful content when rendering the page of the FPD. As a next step, we aim to understand how adblockers reduce the tracking range at the level of legal entities. A legal entity may acquire multiple domains and therefore still receive a lot of third-party requests despite some of its domains being blocked by the adblockers.

Table 5 summarizes the 10 legal entities with the highest TPD node degree, i.e. that were present on most of the visited URLs when the default Browser settings were applied (*NoAdblocker*). As the data suggests, domains owned by *Google Inc.* are loaded by 674 out of the 1000 URLs visited, thus having the most frequent presence among the rest of the third-party entities. Followed by Google Inc. are Facebook Inc., AppNexus Inc., and TMRG Inc. with node degrees of 328, 159, and 143 respectively. The degree of the following domains then quickly drops below 100.

Also presented in Table 5 is the node degree of the top 10 legal entites with the *Ghostery_MaxProtection* and *AdblockPlus_MaxProtection* browser profiles enabled. Except for Google Inc., Ghostery is able to suppress the node degree of all top 10 legal entities below 10. Google Inc. however remains with a node degree of 328, meaning that despite using Ghostery, Google Inc. is able to track more than 30% of the page visits to the FPDs. AdblockPlus is significantly less

Table 5. Legal entities with the highest TPE node degree for browser profile *NoAdblocker* and the corresponding values for Ghostery and AdblockPlus with maximum-protection settings (browser profiles *Ghostery_MaxProtection* and *Adblock-Plus_MaxProtection*) on 28/04/2016.

Legal entity	Degree		
	None	Ghostery	AdblockPlus
Google Inc.	666	328	354
Facebook Inc.	328	6	211
AppNexus Inc.	159	0	0
TMRG Inc.	143	0	4
Twitter Inc.	137	9	87
Oracle corporation	123	2	39
Adobe systems	107	6	32
Yahoo! Inc.	99	7	5
AOL Inc.	88	3	3
OpenX technologies	88	0	0

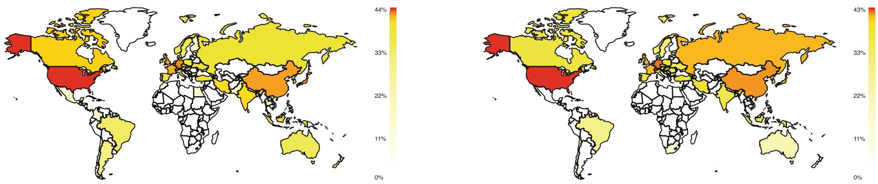
effective than Ghostery even in the maximum protection mode. Still, it reduces significantly the TPD node degree for most TPDs.

5.4 Geographical Considerations

Another key privacy dimension is the geographical location to which third-party requests are transferred to since local regulations govern what legal entities may do with the personal data that they collect about users. Table 6 lists the 10 countries with the highest number of legal entities acting as first party in our

Table 6. Countries hosting the highest percentage first-party entities

Country	First-party entities
United States	35.7%
Canada	7.4%
Japan	4.8%
Switzerland	4.0%
Germany	3.8%
India	3.5%
Great Britain	3.0%
Russia	2.6%
France	2.6%
Panama	2.0%



(a) Third-party legal entity locations (b) Third party server locations

Fig. 5. World map depicting the locations of the legal entities and the servers for the third parties loaded during our experiments. (Color figure online)

traces. The country with the most first parties is the United States (35.7%) followed by Canada (7.4%) and Japan (4.8%). Figure 5a visualizes the relative number of legal entities acting as third parties in each country. The darkest regions (red) are the countries with the most TPEs loaded, while the white ones host none of the TPEs found in our graphs. As we would expect, the USA hosts most of the first and third-party domains, while regions such as Africa or Latin America contain very few TPEs.

A more detailed view of the number of TPEs hosted by the top 10 countries is presented in Table 7. For each row, the absolute numbers refer to the TPDs that were recognized and assigned to a TPE for the specific country, while the percentages refer to the ratio of these TPEs over the total number of TPEs that

Table 7. Countries hosting the highest percentage TPEs when no adblocker is used (browser profile *NoAdblocker*), and the corresponding percentages when Ghostery and AdblockPlus are used under maximum protection settings (browser profiles *Ghostery_MaxProtection* and *Adblockplus_MaxProtection*) on 28/04/2016.

Country	Third-party entities		
	None	Ghostery	AdblockPlus
United States	784 (45%)	483 (42%)	500 (45%)
Germany	106 (6%)	40 (4%)	34 (3%)
China	82 (5%)	70 (6%)	67 (6%)
Japan	80 (5%)	62 (5%)	61 (6%)
Great Britain	77 (4%)	43 (4%)	44 (4%)
France	69 (4%)	33 (3%)	31 (3%)
Canada	49 (3%)	33 (3%)	28 (3%)
India	46 (3%)	38 (3%)	38 (3%)
Panama	41 (2%)	32 (3%)	25 (2%)
Turkey	32 (2%)	27 (2%)	27 (2%)
Total	2908	1866	1812
Found	1748 (60.1%)	1140 (61.1%)	1097 (60.5%)

were recognized by our automated script. In this table, we compare the TPEs hosted by each of these countries (column *None*) to the number of TPEs loaded when the adblockers Ghostery and AdblockPlus are deployed under maximum-protection settings (columns *Ghostery* and *AdblockPlus*).

Interesting to note here is the difference in rank between countries in terms of legal entities that act as first and third parties. For example, China does not appear in the top ten list of countries for first parties, but ranks third in the ranking for legal entities that act as third-parties. This indicates that China hosts in relation to the other countries more third-party domains than first-party domains. The opposite is true for Switzerland and Russia which rank 4th and 7th in the ranking for first-party entities but don't appear in the top ten of third-party entities. Regarding the effect of the Ghostery and AdblockPlus, we can see that these adblockers do not significantly affect the overall distribution and ranking of the third-party legal entities. All countries experience a diminishing number of third-party legal entities that is in proportion relatively equal.

5.5 Graph Density

When grouping the TPD nodes according to the legal entities they belong to, we observe a considerable reduction of the mean FDP node degree, asserting that the number of legal entities potentially collecting information about the user is indeed less than that of the actual third-party domains tracking them.

On the contrary, the mean TPD node degree, as well as the graph density do not present any significant variation, which leads us to the conclusion that the various legal entities have on average access to roughly the same first parties, although controlling multiple third-party domains.

6 Related Work

Privacy Concerns: Many works in the literature have been dedicated to the privacy concerns as a consequence of tracking and fingerprinting by third-party domains [8,13,20,23,28,33]. Castelluccia *et al.* [10] showed that the user's interests can be inferred by the ads they receive and their whole profile can be reconstructed. This can lead to discriminations of the users according to their profile details and configurations, as shown in [11,27].

Countermeasures: As a result, several methods have been proposed that enable targeted advertisements without compromising user privacy [16,17,22,34]. Additionally, there have been a lot of attempts for the detection of tracking behavior and ad-blocking blacklist ehnancements [15,24,35], while some studies have proposed further mitigation techniques [18,30].

Comparison of Mitigation-Techniques: Balebako *et al.* [7] propose a method to measure behavioral targeting and the effect of privacy-protection techniques—e.g. disabling of third-party cookies, Do-Not-Track header, ad-blocking tools—in the limitation of the behavioral-targeted character of the advertising content, while Krishnamurthy *et al.* [19] compare different privacy-protection techniques against the trade-offs between privacy and page quality. Leon *et al.* [21] investigate and compare the usability of some existing tools designed to limit advertising. Pujol *et al.* [29] aim to infer the use or no use of an adblocker by examining the HTTP(S) requests sent by a browser, using the ratio of the ad requests and the downloads of filter lists as indicators. Ruffell *et al.* [32] analyze the effectiveness of various browser add-ons in mitigating and protecting users from third-party tracking networks. However, the time evolution of these metrics is not examined and no legal-entity details are taken into consideration for the graph creation. Mayer and Mitchell [25] implemented the tool FourthParty—an open-source platform for measuring dynamic web content—as an extension to Mozilla Firefox. Englehardt and Narayanan [13] use OpenWPM [12], a web privacy measurement platform that can simulate users, collect data and record observations, e.g. response metadata, cookies and behavior of scripts. Although their study spans across 1 million sites, their measurement provide insights on one time instant, while we provide insights on the daily temporal evolution of web privacy. Recently, Merzdovnik *et al.* [26] performed a large scale study on more than 100,000 popular websites. Unlike our study, their work does not perform an observation of the temporal evolution.

7 Conclusions

The emerging trend of web advertising as well as the earning potential that it has to offer have turned it into the driving force for the development of a broad spectrum of websites and businesses. However, this practice is in direct conflict with privacy matters of the end-user, since the protection of their personal information is at stake through fingerprinting and online-profiling techniques whose objective is to optimize the efficiency of the web advertisements. Adblockers aim to counter these risks by removing advertising content and preventing third-party tracking.

Our analysis provides a quantitative methodology to compare the filtering performance of different adblockers. After the inspection of multiple browser profiles—i.e. combinations of ad-blocking software and configurations—for desktop and mobile devices, we show that the usage of an adblocker can indeed increase the privacy level and restrain the leakage of information concerning the browsing behavior of the user towards third-party trackers. The most important factor that can determine the achieved privacy level is according to our experiments the selection of blacklists, whilst the activation of the *do not track* HTTP header only has a minor effect. Our findings indicate that the best-performing adblockers are Ghostery and then AdblockPlus, when both are set to a maximal-protection level, whilst the highest privacy risks exist when no adblocker or Ghostery with its default blacklist settings is used.

References

1. Allowing acceptable ads in adblock plus. https://adblockplus.org/en/acceptable-ads
2. Google chrome adblock plus. https://chrome.google.com/webstore/detail/adblock-plus/cfhdojbkjhnklbpkdaibdccddilifddb/support?hl=en-GB
3. Juniper research. http://www.juniperresearch.com/press/press-releases/ad-blocking-to-cost-publishers-$27bn-in-lost-reven
4. Mozilla statistics for adblock plus. https://addons.mozilla.org/en-US/firefox/addon/adblock-plus/statistics/?last=30
5. The state of do not track in firefox. https://dnt-dashboard.mozilla.org/
6. Number of mobile-only internet users now exceeds desktop-only in the U.S, April 2015. http://www.comscore.com/Insights/Blog/Number-of-Mobile-Only-Internet-Users-Now-Exceeds-Desktop-Only-in-the-U.S
7. Balebako, R., Leon, P., Shay, R., Ur, B., Wang, Y., Cranor, L.: Measuring the effectiveness of privacy tools for limiting behavioral advertising. Web (2012)
8. Barford, P., Canadi, I., Krushevskaja, D., Ma, O., Muthukrishnan, S.: Adscape: harvesting and analyzing online display ads. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 597–608. ACM (2014)
9. Butkiewicz, M., Madhyastha, H.V., Sekar, V.: Understanding website complexity: measurements, metrics, and implications. In: Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference, pp. 313–328. ACM (2011)
10. Castelluccia, C., Kaafar, M.-A., Tran, M.-D.: Betrayed by your ads!. In: Fischer-Hübner, S., Wright, M. (eds.) PETS 2012. LNCS, vol. 7384, pp. 1–17. Springer, Heidelberg (2012). doi:10.1007/978-3-642-31680-7_1
11. Datta, A., Tschantz, M.C., Datta, A.: Automated experiments on ad privacy settings. Proc. Priv. Enhanc. Technol. **2015**(1), 92–112 (2015)
12. Englehardt, S., Eubank, C., Zimmerman, P., Reisman, D., Narayanan, A.: Web privacy measurement: scientific principles, engineering platform, and new results (2014). http://randomwalker.info/publications/WebPrivacyMeasurement.pdf
13. Englehardt, S., Narayanan, A.: Online tracking: a 1-million-site measurement and analysis. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 1388–1401. ACM (2016)
14. Gill, P., Erramilli, V., Chaintreau, A., Krishnamurthy, B., Papagiannaki, K., Rodriguez, P.: Best paper - follow the money: understanding economics of online aggregation and advertising. In: Proceedings of the 2013 Conference on Internet Measurement Conference, IMC 2013, pp. 141–148. ACM, New York (2013)
15. Gugelmann, D., Happe, M., Ager, B., Lenders, V.: An automated approach for complementing ad blockers' blacklists. Proc. Priv. Enhanc. Technol. **2015**(2), 282–298 (2015)
16. Guha, S., Cheng, B., Francis, P.: Privad: practical privacy in online advertising. In: USENIX Conference on Networked Systems Design and Implementation, pp. 169–182 (2011)
17. Haddadi, H., Guha, S., Francis, P.: Not all adware is badware: towards privacy-aware advertising. In: Godart, C., Gronau, N., Sharma, S., Canals, G. (eds.) I3E 2009. IAICT, vol. 305, pp. 161–172. Springer, Heidelberg (2009). doi:10.1007/978-3-642-04280-5_14
18. Kontaxis, G., Chew, M.: Tracking protection in firefox for privacy and performance. arXiv preprint arXiv:1506.04104 (2015)

19. Krishnamurthy, B., Malandrino, D., Wills, C.E.: Measuring privacy loss and the impact of privacy protection in web browsing. In: Proceedings of the 3rd Symposium on Usable Privacy and Security, pp. 52–63. ACM (2007)
20. Krishnamurthy, B., Wills, C.: Privacy diffusion on the web: a longitudinal perspective. In: Proceedings of the 18th International Conference on World Wide Web, pp. 541–550. ACM (2009)
21. Leon, P., Ur, B., Shay, R., Wang, Y., Balebako, R., Cranor, L.: Why johnny can't opt out: a usability evaluation of tools to limit online behavioral advertising. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 589–598. ACM (2012)
22. Levin, D., Bhattacharjee, B., Douceur, J.R., Lorch, J.R., Mickens, J., Moscibroda, T.: Nurikabe: private yet accountable targeted advertising (2009, under submission). johndo@microsoft.com
23. Libert, T.: Exposing the invisible web: an analysis of third-party http requests on 1 million websites. Int. J. Commun. **9**, 18 (2015)
24. Ma, J., Saul, L.K., Savage, S., Voelker, G.M.: Beyond blacklists: learning to detect malicious web sites from suspicious URLs. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1245–1254. ACM (2009)
25. Mayer, J.R., Mitchell, J.C.: Third-party web tracking: policy and technology. In: 2012 IEEE Symposium on Security and Privacy (SP), pp. 413–427. IEEE (2012)
26. Merzdovnik, G., Huber, M., Buhov, D., Nikiforakis, N., Neuner, S., Schmiedecker, M., Weippl, E.: Block me if you can: a large-scale study of tracker-blocking tools. In: 2nd IEEE European Symposium on Security and Privacy, Paris, France (2017)
27. Mikians, J., Gyarmati, L., Erramilli, V., Laoutaris, N.: Detecting price and search discrimination on the internet. In: Proceedings of the 11th ACM Workshop on Hot Topics in Networks, pp. 79–84. ACM (2012)
28. Nikiforakis, N., Acar, G.: Browse at your own risk. IEEE Spectr. **51**(8), 30–35 (2014)
29. Pujol, E., Hohlfeld, O., Feldmann, A.: Annoyed users: ads and ad-block usage in the wild. In: Proceedings of the 2015 ACM Conference on Internet Measurement Conference, pp. 93–106. ACM (2015)
30. Roesner, F., Kohno, T., Wetherall, D.: Detecting and defending against third-party tracking on the web. In: Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, pp. 12–12. USENIX Association (2012)
31. Singer, D., Fielding, R.T.: Tracking preference expression (DNT). Technical report, May 2015
32. Ruffell, M., Hong, J.B., Kim, D.S.: Analyzing the effectiveness of privacy related add-ons employed to thwart web based tracking. In: 2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC), pp. 264–272. IEEE (2015)
33. Soltani, A., Canty, S., Mayo, Q., Thomas, L., Hoofnagle, C.J.: Flash cookies and privacy. In: AAAI Spring Symposium: Intelligent Information Privacy Management, vol. 2010, pp. 158–163 (2010)
34. Toubiana, V., Narayanan, A., Boneh, D., Nissenbaum, H., Barocas, S.: Adnostic: privacy preserving targeted advertising. In: Proceedings Network and Distributed System Symposium (2010)
35. Tran, M., Dong, X., Liang, Z., Jiang, X.: Tracking the trackers: fast and scalable dynamic analysis of web content for privacy violations. In: Bao, F., Samarati, P., Zhou, J. (eds.) ACNS 2012. LNCS, vol. 7341, pp. 418–435. Springer, Heidelberg (2012). doi:10.1007/978-3-642-31284-7_25