2016 IEEE 41st Conference on Local Computer Networks

Third-party Tracking on the Web: A Swedish Perspective

Joel Purra and Niklas Carlsson
Linköping University, Sweden

***Abstract*—Today, third-party tracking services and passive traffic monitoring are extensively used to gather knowledge about users' internet activities and interests. Such tracking has significant privacy implications for the end users. This paper presents an overview of the third-party tracking usage. Using measurements, we highlight the current state of the third-party tracking landscape and differences observed across tracking service classes (e.g., advertising, analytics, and content), across domain categories (e.g., popular vs. less popular, and national vs. global domains), and with regards to the organizations that owns many of the tracker services, when using HTTP and HTTPS, respectively. Understanding these differences help answer questions related to the third-party services that track modern web users and their coverage of our browsing.**

***Index Terms*—Third-party tracking; Privacy; HTTP vs HTTPS**

I. Introduction

We are living in an information society in which organizations constantly track our movements on the web, and use the collected information to analyze and gain knowledge about us and our interests. For example, website owners may use such knowledge to present personalized content that may better match user interests, advertisement firms may use the knowledge to sell targeted ads based on user interests [8], [15], [28], [33], media analytics firms may use the data to verify advertisement-related statistics, and data brokers may package and sell the user data inferred from the user interests [3].

Web usage is both actively and passively tracked. With active tracking, third-party scripts and plugins embedded within the visited websites are typically used to extract and collect information about a user's every click, the time spent on each page, and information that can help uniquely identify each browser and client. While some trackers allow advanced fingerprinting through collection of client-side information, much tracking is achieved through server-side tracking of the download of relatively simple third-party contents such as CSS files, customized fonts, or javascript libraries that do not include tracking code themselves [1], [17], [18], [22], [27]. With passive tracking, user activities are typically extracted from traffic logs collected through traffic monitoring within a network or at a cloud-operated data center hosting some of the content. Here, both TCP/IP header information (e.g., IP addresses) and application specific information such as HTTP header information and payload data (including information sent to third-party trackers) from unencrypted HTTP transfers can be used to identify users and track their web sessions.

With website providers and the public becoming increasingly aware of the traffic monitoring by network operators and nation states (e.g., the Swedish government openly allows the National Defense Radio Establishment (FRA) to monitor all traffic passing the countries boundaries[1] and the National Security Agency (NSA) in the USA have received much international press coverage) it is perhaps not surprising that there is an increasing number of services that use HTTPS [21]. However, without measuring the third-party tracking, it is unclear if domains that provide their users with the option to use HTTPS (e.g., to provide users added privacy/security protection against network monitoring) in fact reduces the amount of tracking. While securing the end-to-end communication between clients and servers helps reduce the information that can be passively collected along the network path, it is important to note that the use of HTTPS in itself does not protect against third-party tracking through embedded content or scripts.[2] The degree of third-party tracking that a user is exposed to is instead typically determined by the amount of third-party services that the visited websites themselves leverage in their design.

A. Contributions and Summary of Results

This paper presents a characterization of the current third-party tracking landscape and answer a number of important and un-answered questions. First, we compare the third-party usage across a number of website classes and breakdown the coverage of different tracker types. Second, we present an aggregate analysis that combines the tracker services based on the organizations operating them so to gain insights into the big players aggregate coverage. Finally, throughout our analysis, we try to answer if websites that have adopted HTTPS in fact are more privacy conscious (on behalf of their users) and use less third-party tracking. Combined with our brief description of the current HTTPS adoption, answering these questions also help answer how much privacy can be gained from conscious use of HTTPS.

To help address these and other important questions, we present a novel measurement methodology and analysis of the current third-party tracking usage. The study presents results from both a national (Swedish, in our case) and global perspective. The paper makes two technical contributions.

[1] http://news.bbc.co.uk/2/hi/europe/7463333.stm

[2] HTTPS also does not protect against passive monitoring of DNS lookups [29] and TCP/IP header information [26]; only against deep packet inspection. The design and high accuracy of techniques that identify user actions within HTTPS traffic has been demonstrated [26].

© 2016, Joel Purra. Under license to IEEE.
DOI 10.1109/LCN.2016.14

IEEE computer society

First, we develop and release the software of a measurement framework for automated, repeatable retrieval, and analysis of websites, their third-party usage and HTTPS adoption.[3] The software is built on open source software, uses a "standard" headless browser, and stores results into standardized HAR format. To capture HTTPS adoption and redirections, domains are visited using both HTTP and HTTPS, with and without the `www` prefix. Third-party resource usage is captured through post processing (including comparison with the publicly available tracker lists of the popular privacy tool Disconnect (https://disconnect.me/) and public suffix lists are used to identify domains and subdomains.

Second, we present a characterization and analysis of the third-party tracking usage. Overall, the use of third-party (external) resources is high among all domain categories, regardless if HTTP or HTTPS is used. For example, more than 93% of the globally most popular domains use external third-party resources. These numbers are even higher for the most popular Swedish websites within most websites categories, suggesting that users' activities typically can be tracked through third-party resource usage. In fact, most websites also have at least one *known* tracker present: 53-72% of random domains, 88-98% of top websites, and 78-100% of websites in the nine different Swedish top-categories considered in this paper. Comparing known tracker usage, we have also found slightly higher usage of known trackers on websites using HTTPS and some indications that that there may be a correlation between HTTPS adoption and higher third-party tracking. These results suggest that some users may be given a false sense of privacy when using HTTPS.

While Google (who is at the forefront of HTTPS adoption) has by far the greatest third-party coverage among known tracker companies (e.g., Google has trackers on more than 90% of the websites within the majority of the website categories considered here), we have found that there are many other third-party domains that may fly under the radar, which add to the total tracker coverage. For example, Disconnect's blocking list only detects 10% of the third-party primary domains. With most of these non-blocked third-party entries being third-party content providers, which are known to keep track of users across services, we expect that there will be a continued battle for the knowledge about our web activities.

The remainder of the paper is organized as follows. Our methodology and measurement framework are described in Section II. Section III characterizes the third-party tracking landscape and how it differs between secure and non-secure domains. Finally, Section IV discusses related studies and results, putting our work in the context of these studies, before Section V concludes the paper.

II. Methodology

We first describe our measurement campaign. For a large set of domains, we visit the front page of each domain, using both `http` and `https`. For each case, we access and measure both the domain with and without the `www` prefix, to cover (and investigate) all cases where only one or both of the variations is accessible. Using the data collection tool developed in this project, the front page of each domain is downloaded and parsed using the headless `phantomjs` browser the same way the users' browser would. During this process scripts are executed/processed and every resource used to build the front page is downloaded, including objects such as embedded images, fonts, and scripts. For most modern websites, this step involves downloading many resources spread across different domains. Then, the URL, domain and other HTTP characteristics are extracted for each requested resource. Finally, each object and domain is classified and prepared for post processing and analysis. This step includes separating resources retrieved from the same domain as the front page (internal resources) from resources retrieved from external domains (external resources), as well as identifying resources downloaded from known third-party tracking services.

Swedish perspective: First, our measurements are performed from computers located in Sweden. Second, we leverage the domain lists from the .SE Health Status report [16], identified by the Internet Infrastructure Foundation (IIS) as the most important domains to Swedish internet usage and operation. Combined, these lists include approximately 1,000 domains in the categories: counties (21), domain registrars (146), financial services (79), government-owned corporations (GOCS) (60), higher education (49), ISPs (20), media (33), municipalities (290), and public authorities (282). Third, we consider the 50 global websites most visited by Swedish internet users (reach50), all .se domains (3,364) among the top-million globally most popular domains according to Alexa (www.alexa.com), as well as 100K randomly selected websites within the .se domain zone.

Global baseline: To put our findings into perspective and broaden our conclusions we also consider the globally most popular websites (according to Alexa), all .dk (Denmark) domains (2,637) among the top-million globally most popular domains, and randomly selected websites from the .com, .net, and .dk (Denmark) domain zones. Table I summarizes the domain lists used.

Data collection and parallelization: HTTP/HTTPS traffic metadata such as requested URLs and their HTTP request/response headers have been recorded in the HTTP Archive (HAR) data format. During our data collection, multiple domains have been retrieved in parallel, with parallelism adjusted to fit the available computer capacity. To reduce the risk of intermittent errors, each failed access has been retried up to two times.

Data extraction: A custom-built tool based on the command line JSON processor `jq` is used to extract and transform information about both requests and responses. The extracted data includes protocol, hosts, HTTP status, mime-type, referer and redirect values both for the origin domain's front page and any resources requests by it.

[3]Both the software and datasets are made public with this article, and can be found here: http://www.ida.liu.se/~nikca89/papers/lcn2016-purra.html.

TABLE I
SUMMARY OF DOMAIN LISTS.

Domain lists			Selection		
List(s)	Date	Total size	Type	Size	Unique
.SE Health Report	27/3/14	980	curated (9 categories)	915	915
.se zone	10/7/14	1,318,000	random	100,000	100,000
.dk zone	23/7/14	1,260,000	random	10,000	10,000
.com zone	27/8/14	114,178,000	random	10,000	10,000
.net zone	27/8/14	15,096,000	random	10,000	10,000
Reach 50	1/9/14	50	top	50	50
Alexa top-1M	1/9/14	1,000,000	top	10,000	9,986
–"–	–"–	–"–	random	10,000	9,959
–"–	–"–	–"–	all .se	3,364	3 364
–"–	–"–	–"–	all .dk	2,637	2,637
Total	–	132,852,050	–	156,907	156,045

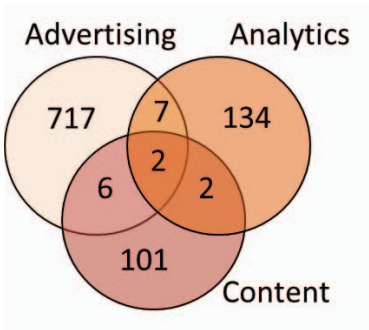


Fig. 1. Organizations using different combinations of known trackers.

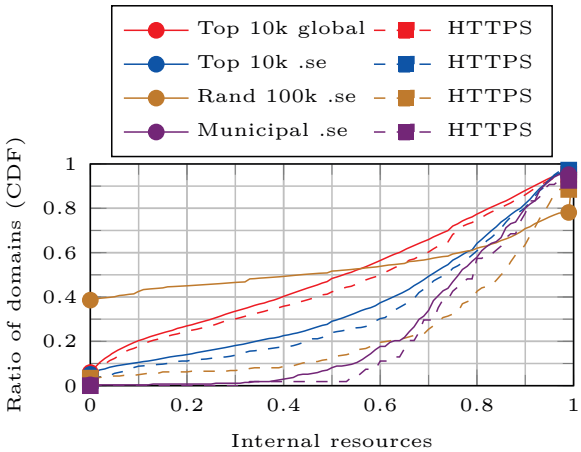


Fig. 2. External third-party resource usage.

Domain identification and classification: Each requested resource is classified across multiple dimensions; e.g., it's mime type, if it was retrieved securely or not, and if it was downloaded from an internal sub-domain or from a known third-party tracker. To determine the primary domain and subdomain(s) we use the public suffix list[4] and match domains against it. This is important to capture the relationships between second-level domains not open for public registration (e.g., .com.br under .br). We refer to resources requested from the front page that are retrieved from the same domain, a subdomain, or a superdomain as *internal* requests. Any other request is deemed as *external*.

Tracker identification: Public lists of known trackers typically are incomplete and out-of-date [13], [17]. Since potentially any external resource can be used for third-party tracking it is difficult to quantify exactly how much third-party tracking takes place. Even static (non-script, non-executable) resources with no capabilities to dynamically survey the users' browser or OS can track users across domains (e.g., using the HTTP referer field or customized URIs).

To avoid missing third-party relationships, our data collection method does not block requests. Instead each resource is classified as either internal or external, and each URL is matched against known trackers during post processing. The external third-party usage provides an upper bound, while known, confirmed, recognized third-party trackers are used for a lower bound analysis. For the tracker analysis, we use the public (open source) tracker list used by the privacy tool Disconnect.me. This list (collected 8/9/14) contains 2,149 domains, each belonging to one of 980 organizations and five tracker categories. Figure 1 shows the number of organizations that use the three most common tracker categories: Advertising, Analytics, and Content. There are also 14 organizations (with 43 total domains) in the social category and 3 organizations (Google, Facebook, and Twitter) in the special "Disconnect" category.

[4]Public suffix list, https://publicsuffix.org/.

III. THIRD-PARTY TRACKING

A. External third-party resource usage

To upper bound the amount of third-party tracking, we first analyze the use of external resources, each of which can enable third-party tracking. Figure 2 shows the cumulative distribution function (CDF) of the ratio of external resources used by each domain (x-axis), with 0% and 99% internal resources marked. Overall, the external resource usage is very high. For example, 93% of the domains among the top domains (e.g., "Top-10K global") use at least some external resources, and for 70% most resources are external.

Interestingly, with the exception of random websites (e.g., "Rand-100K .se"), the difference between the HTTP and

HTTPS datasets is generally small, suggesting that third-party providers potentially may have as good insight into the users activities on secure sites as they do on insecure sites. In fact, for most "popular" categories we have observed slightly higher external resource usage for the HTTPS datasets. The high (40%) usage of entirely external resources seen for the random HTTP domains appears to be partially due to "parked domains" (purchased domains with placeholder contents), which often loads all their resources from an external domain belonging to the domain retailer who sold the domain. For example, post processing shows that 28% of the "Rand-100K .se" domains obtained at least one external resource from (to us) known retailer domains.

B. Known Trackers

To glean some insights into who does the tracking and what type of tracking is being done we leverage Disconnect's tracker list, which categorizes 2,149 known and recognized tracker domains and maps these trackers to the organizations that own them. When interpreting these results, we note that these known trackers make up less than 10% of the third-party domains observed serving external resources in our measurements. As third-party tracking can be done effectively by any of these domains, the following analysis therefore serves as a lower bound, and we may be even more tracked than suggested by the numbers presented.

Figure 3 (top row) breaks down the observed tracker usage of five tracker categories (separate bars) for each website category when using HTTP. The white thick bars show the combined coverage of the union of all known trackers, regardless of tracker category. The Disconnect category (pale yellow) has almost the same coverage as the union of all categories, suggesting that the top players (Google, Facebook, and Twitter) have very large coverage on their own. The second largest category is content. Disconnect.me does not block this category by default, suggesting that users running Disconnect's software would still be tracked by known trackers on at least 60-70% of the websites; likely more, given the limitations of blocking lists [17]. Advertising and analytics are particularly common among Swedish media domains and globally popular domains ("Top-10K globally").

Overall, the relative coverage observed for each tracker type typically differs more between category types of different popularity (e.g., popular vs random) than across domain categories with different locality (.se, .dk, or global, for example). This is particularly apparent for advertising services, which is likely to target the popular domains.

While we have presented the results for HTTP, the results for HTTPS show slightly higher tracker usage. This is illustrated in Figure 4, which shows the correlation of known tracker coverage seen with HTTP and HTTPS. Here, coverage is measured on a domain-category basis, and results are shown separately for Swedish (Figure 4(a)) and Global (Figure 4(b)) domain categories.

The small differences between the coverage seen when using HTTP or HTTPS is an important and interesting finding, as it suggests that users are as tracked by known third-party trackers when using HTTPS as they are when using HTTP. In fact, if anything, we have found that the tracker coverage (especially by the content category) is slightly higher among website that implements HTTPS. This is exemplified by the larger number of points above the parity reference line (added to show the case if there was no difference in coverage). A possible explanation may be that the websites that implements HTTPS (e.g., to satisfy their users' emerging desire for privacy protection against network monitoring) also are the websites that are most interested in knowing their users' interests. This is consistent with the observation that both the HTTPS adoption and tracker usage are highest among the most popular domains, which may be most influenced by recommendations by other professionals. However, it may also be due to HTTPS sites simply serving different content, which rely more on third-party content and advertisements.

The big players: Motivated by the high coverage by the Disconnect category, we take a closer look at the coverage of the three dominating third-party players in our dataset: Google, Facebook, and Twitter. Figure 3 (bottom row) shows the coverage for these organizations as observed across both national (left-hand side) and international (right-hand side) domain categories. We also include a marker ($\times$) for domains falling into the Disconnect category, while noting that Google and Facebook own domains both within and outside this special category.

Google has by far the greatest coverage and has on its own greater coverage than the Disconnect category.[5] Google, with its 271 tracker domains included in Disconnect, is particularly popular among top domains (approximately 90% coverage), but has consistently good coverage across all domain categories (above 70%). Facebook and Twitter, who have the next best coverage among all 980 considered organization, are far behind. For example, with exception of Swedish media sites (Facebook 60%), Facebook does not reach more than 40% for any category. Twitter peaks at 25% (Swedish media and top-10K global websites). While lower absolute coverage compared to Google, both Facebook and Twitter see bigger differences in coverage between the top domains and the random domains.

Category comparison: Overall, we have seen the greatest tracker coverage among the most popular domains. For example, among the top-10K global websites, we have found that regardless if using HTTP or HTTPS, more than 95% of the domains use at least one known tracker, 70% use tracker domains from more than one external organization (each with at least one tracker domain at the front page of the website), 10% allow more than 12 known tracker organizations to track their users, and 1% allow more than 48 known tracker organizations to track visitors. Given that we only analyzed the front page of the domain and the Disconnect list (as we have seen) covers less than 10% of the external domains, these

[5]The Google trackers not included in the Disconnect category are known content trackers, including both YouTube and Google branded search tools, and unbranded font resources and script hosting.

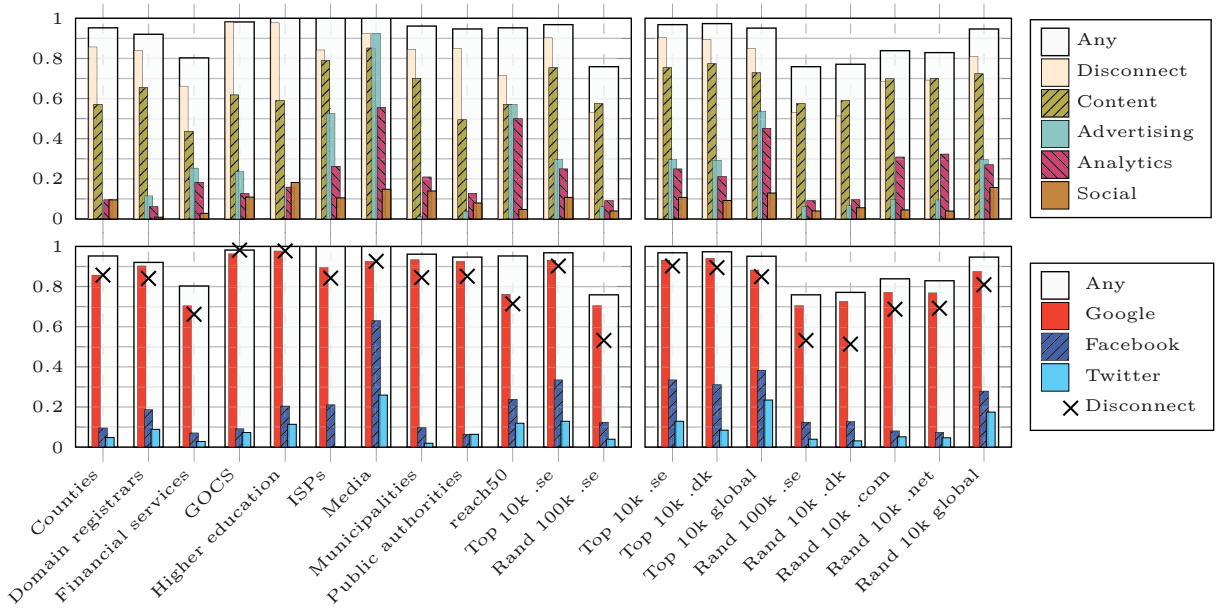
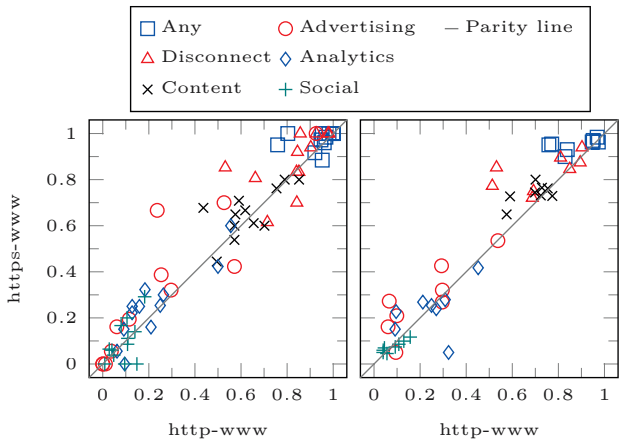


Fig. 3. Domain coverage of different categories of known tracker services (top row) and companies (bottom row), as per Disconnect's blocking list.



(a) Sweden (b) Global

Fig. 4. Correlation of tracker coverage seen with HTTP and HTTPS.

number only provides a lower bound. The large number of potential trackers is particularly concerning as the use of many trackers makes it more difficult to control the flow of personal information and where it ends up [30].

Nationally, the tracking was most extreme among Swedish media domains, with 50% sharing information with more than 7 known tracker organizations, and one sharing information with 38 organizations. A single visit to the front page of each of the 27 investigated media websites leaked information through over 3,800 external requests and to at least 57 known tracker organizations. Clearly, there are many organizations that collect information about our online interests, regardless if we use HTTP or HTTPS. We are constantly tracked and there are many organizations that need limited-to-no guesswork to determine which news articles a user reads.

Additional HTTPS observations: When performing this study we observed a fairly skewed HTTPS adoption. The overall HTTPS usage is lowest among random domains (e.g., less than 0.6% for "Rand 100k .se" and less than 1% for the other random categories) and highest for popular domains (e.g., 15-50% for the different popular and important Swedish categories, 10-30% for the globally popular websites, and 53% for Reach50). To make things worse, many resources are accessed insecurely even when the main page is downloaded over HTTPS.

Furthermore, although the HTTPS redirects consistently (on a per-category basis) result in more secure redirects, the majority of redirects still results in an insecure connection. In fact, many domains that implement HTTPS redirect clients to a preferred variant of their domain name (usually the `www` subdomain) rather than to their secure domain. It is even less common that domains proactively redirect clients from `http` domains to secure `https` domains. For example, out of the redirects from `http` domains only two categories have more than 15% such secure redirects: Swedish ISPs (30%) and Reach-50 (34%).

IV. Related Studies and Results

We are not the first to characterize various aspects of the third-party tracking landscape [7], [9], [12], [18], [19], [25], [27]. First was perhaps Krishnamurthy and Wills [12], who use longitudinal measurement snapshots to shows that the third-party tracker usage was increasing significantly already between 2005 and 2008. Their work provides a nice overview of how some organizations (e.g., Google) have increased their tracking coverage both by increased usage of some of the third-party domains that they own, and through active acquisition of new domains (e.g., doubleclick) that provide

ad services, analytics, and tracking. Since their most recent Sept. 2008 dataset, we note that much have happened and that today's coverage of the top players is even greater. For example, their Sept. 2008 data suggests that Google had a coverage of just over 40% among the most popular domains. Today, we observe a Google coverage of 90% among a similar top-10K set. This is even greater than the 81% (upper) coverage bound provided by all third-party domains combined in Sept. 2008 (or 53% in Oct. 2005).

More recently, Falahrastegar et al. [7] use active measurements to identify both global giants (e.g., Google) and many local (language dependent) third-party websites. In their study they compare the most frequently observed third-party services observed on the front pages of the top-500 pages in a few western (US, UK, Australia) and eastern (China, Iran, Egypt, and Syria) countries. Metwalley et al. [19] use passive measurements inside two ISPs to provide a complementary and very interesting angle to third-party tracking. Their measurements show that 77% of the users face trackers just one second into their online activity, 100% within 100 seconds of their first web request, and that most users (more than 80%) do not use any privacy enhancing browser plugin (e.g., AdBlockPlus and DoNotTRackMe/Blur). Similarly, Pujol et al. [25] leverage functionalities in AdBlockPlus to characterize the ad traffic in a major European ISP.

Others have considered the tracking when following the (non-sponsored) links provided by popular search engines [9], suggested taxonomies for third-party tracking [27], and developed tools to measure and protect against third-party tracking [17], [18]. In contrast to the above works, we leverage Disconnect's classification of trackers, group domains by organizational ownership, and compare the third-party resource usage and known trackers observed across different services when using HTTP or HTTPS.

Other works have studied the use of finger printing [2], [23], tracking cookies [1], and other techniques that allows users can be uniquely identified based on their browsers [6]. In this work we assume their existence and instead focus on the the third-party resource usage with and without HTTPS.

There has also been works that look at how the tracking data may be used. Researchers have analyzed how websites leverage personal data for targeted personalized ads, improve web search results, and social network update feeds [5], [8], [15], [24], [28], [33] and improve web search results and social network update feeds [5], [24]. E-commerce websites have also used indicators such as geographic location, hardware platform/browser for price steering and price discrimination [10], [20]. While many sites may keep the information they obtain, the market for tracking data resale is expected to grow, as the amount of data increases and quality improves [3]. The current information landscape is further complicated by cross-site information leakage (of personal information) authorized through third-party identity management agreements [31] and the observation that users are connected to their identities, even after logging out from a service [11], [14].

V. Conclusion and Discussion

This paper has presented a measurement-driven analysis of the current third-party tracking usage across different domain categories. The paper first described our measurement framework for automated, repeatable data collection and analysis of websites. Second, we presented an analysis of the third-party tracker usage when using HTTP and HTTPS, respectively. All analysis was performed from both a national (Swedish) and global perspective. Across both national and international domain categories, our results show that the use of third-party resources is often at least as high among HTTPS domains as it is among the corresponding HTTP domain categories, with some indications that domains that have higher HTTPS adoption may be more likely to have known third-party trackers. Users who choose HTTPS for the purpose of increased *privacy* may therefore gain very little of it. While Google has by far the greatest third-party coverage (more than 90% for most website categories), there are many other third-party domains and organizations that may fly under the radar, which add to the total tracker coverage.

Given the large coverage of the most frequent third-party tracking organizations and the potential cross-site information sharing involving these and other third-party organizations, it is difficult even for privacy conscious users to completely protect themselves from third-party tracking. Possible approaches to reduce the value obtained by the third-party trackers include altering and obfuscating the traffic patterns associated with users (e.g., by altering requests, adding dummy requests, or blocking requests to known third-party tracking services), explicitly removing user identifiers and cookies from the requests, or through the use of anonymity services such as Tor [4]. However, these services typically come with significant overhead and may in some cases not always be feasible/allowed. In addition, users are increasingly often asked to login and use services online that require them to use their true identity. This include both government operated services and commercially operated third-part authentication services [31], [32]. With usage patterns often being possible to connect across services and profiles [34], we foresee increasingly less privacy on the internet unless the websites themselves take a stand and refrain from using third-party services. However, as of today, the commercial interest of the websites, including some actors treating personal data as a commodity, for example, are pushing in the opposite direction, leaving users increasingly tracked.

References

[1] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz. The web never forgets: Persistent tracking mechanisms in the wild. In *Proc. ACM CCS*, 2014.

[2] G. Acar, M. Juarez, N. Nikiforakis, C. Diaz, S. Gurses, F. Piessens, and B. Preneel. FPDetective: Dusting the web for fingerprinters. In *Proc. ACM CCS*, 2013.

[3] S. Armour. Data brokers come under fresh scrutiny: Consumer profiles marketed to lenders. *Wall Street Journal*, Feb. 2014.

[4] R. Dingledine, N. Mathewson, and P. Syverson. Tor: The second generation onion router. In *Proc. USENIX Security Symposium*, 2004.

[5] Z. Dou, R. Song, and J.-R. Wen. A large-scale evaluation and analysis of personalized search strategies. In *Proc. WWW*, 2007.

[6] P. Eckersley. How unique is your web browser? In *Proc. PETS*, 2010.

[7] M. Falahrastegar, H. Haddadi, S. Uhlig, and R. Mortier. The rise of panopticons: Examining region-specific third-party web tracking. In *Proc. TMA*, 2014.

[8] P. Gill, V. Erramilli, A. Chaintreau, B. Krishnamurthy, K. Papagiannaki, and P. Rodriguez. Follow the money: Understanding economics of online aggregation and advertising. In *Proc. IMC*, 2013.

[9] R. Gomer, E. M. Rodrigues, N. Milic-Frayling, and M. C. Schraefel. Network analysis of third party tracking: User exposure to tracking cookies through search. In *Proc. IEEE/WIC/ACM WI-IAT*, 2013.

[10] A. Hannak, G. Soeller, D. Lazer, A. Mislove, and C. Wilson. Measuring price discrimination and steering on e-commerce web sites. In *Proc. IMC*, 2014.

[11] G. Kontaxis, M. Polychronakis, A. D. Keromytis, and E. P. Markatos. Privacy-preserving social plugins. In *Proc. USENIX Conference on Security Symposium*, 2012.

[12] B. Krishnamurthy and C. Wills. Privacy diffusion on the web: A longitudinal perspective. In *Proc. WWW*, 2009.

[13] B. Krishnamurthy and C. E. Wills. Cat and mouse: Content delivery tradeoffs in web access. In *Proc. WWW*, 2006.

[14] B. Krishnamurthy and C. E. Wills. Characterizing privacy in online social networks. In *Proc. WOSN*, 2008.

[15] B. Liu, A. Sheth, U. Weinsberg, J. Chandrashekar, and R. Govindan. Adreveal: Improving transparency into online targeted advertising. In *Proc. HotNets*, 2013.

[16] A.-M. E. Löwinder and P. Wallström. Health status 2012. Technical report, .SE The Internet Infrastructure Foundation, 2012.

[17] D. Malandrino, A. Petta, V. Scarano, L. Serra, R. Spinelli, and B. Krishnamurthy. Privacy awareness about information leakage: Who knows what about me? In *Proc. ACM WPES*, 2013.

[18] J. R. Mayer and J. C. Mitchell. Third-party web tracking: Policy and technology. In *Proc. IEEE S&P*, 2012.

[19] H. Metwalley, S. Traverso, M. Mellia, S. Miskovic, and M. Baldi. The online tracking horde: A view from passive measurements. In *Proc. TMA*, 2015.

[20] J. Mikians, L. Gyarmati, V. Erramilli, and N. Laoutaris. Crowd-assisted search for price discrimination in e-commerce: First results. In *Proc. ACM CoNEXT*, 2013.

[21] D. Naylor, A. Finamore, I. Leontiadis, Y. Grunenberger, M. Mellia, M. Munafo, K. Papagiannaki, and P. Steenkiste. The cost of the "S" in HTTPS. In *Proc. ACM CoNEXT*, 2014.

[22] N. Nikiforakis, L. Invernizzi, A. Kapravelos, S. V. Acker, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna. You are what you include: Large-scale evaluation of remote javascript inclusions. In *Proc. ACM CCS*, 2012.

[23] N. Nikiforakis, A. Kapravelos, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna. Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. In *Proc. IEEE S&P*, 2013.

[24] E. Pariser. *The filter bubble : what the Internet is hiding from you*. Penguin Press, New York, 2011.

[25] E. Pujol, O. Hohlfeld, and A. Feldmann. Annoyed users: Ads and ad-block usage in the wild. In *Proc. IMC*, 2015.

[26] G. Rizothanasis, N. Carlsson, and A. Mahanti. Identifying user actions from HTTP(S) traffic. In *Proc. IEEE LCN*, 2016.

[27] F. Roesner, T. Kohno, and D. Wetherall. Detecting and defending against third-party tracking on the web. In *Proc. NSDI*, 2012.

[28] D. Saez-Trumper, Y. Liu, R. Baeza-Yates, B. Krishnamurthy, and A. Mislove. Beyond CPM and CPC: Determining the value of users on OSNs. In *Proc. ACM COSN*, 2014.

[29] S. Sanders and J. Kaur. A graph theoretical analysis of the web using DNS traffic traces. In *Proc. IEEE MASCOTS*, 2015.

[30] H. J. Smith, T. Dinev, and H. Xu. Information privacy research: An interdisciplinary review. *MIS Quarterly*, 35(4):989–1015, Dec. 2011.

[31] A. Vapen, N. Carlsson, A. Mahanti, and N. Shahmehri. Information sharing and user privacy in the third-party identity management landscape. In *Proc. IFIP SEC*, 2015.

[32] A. Vapen, N. Carlsson, A. Mahanti, and N. Shahmehri. A look at the third-party identity management landscape. *IEEE Internet Computing*, 20(2):18–25, Mar/Apr. 2016.

[33] N. J. Yuan, F. Zhang, D. Lian, K. Zheng, S. Yu, and X. Xie. We know how you live: Exploring the spectrum of urban lifestyles. In *Proc. ACM COSN*, 2013.

[34] R. Zafarani and H. Liu. Connecting users across social media sites: A behavioral-modeling approach. In *Proc. ACM SIGKDD*, 2013.