Half Baked: The opportunity to secure cookie-based identifiers from passive surveillance

Andrew Hilts[1,2] and Christopher Parsons[1]

[1]Citizen Lab, Munk School of Global Affairs, University of Toronto
[2]Open Effect

Abstract

Documents released by Edward Snowden have revealed that the National Security Agency, and its Australian, British, Canadian, and New Zealand equivalents, routinely monitor the Internet for the identifiers that are contained in advertising and tracking cookies. Once collected, the identifiers are stored in government databases and used to develop patterns of life, or the chains of activities that individuals engage in when they use Internet-capable devices. This paper investigates the extent to which contemporary advertising and analytics identifiers that are used in establishing such patterns continue to be transmitted in plaintext following Snowden's revelations. We look at variations in the secure transmission of cookie-based identifiers across different website categories, and identify practical steps for both website operators and ad tracking companies to take to better secure their audiences and readers from passive surveillance.

1 Introduction

Edward Snowden revealed that the National Security Agency, and its Australian, British, Canadian, and New Zealand equivalents, routinely monitor for identifiers that are contained in advertising and tracking cookies. The identifiers are read as they pass, unencrypted, through key Internet exchanges and access points spread around the world. Once collected, the identifiers are stored in government databases and used to develop "patterns of life", or the chains of activities that individuals engage in when they use Internet-capable devices. This paper investigates the extent to which contemporary advertising and analytics identifiers that are used in establishing such patterns continue to be transmitted in plaintext following Snowden's revelations.

We hypothesize there will be significant variation across different types of websites in terms of whether or not the websites support HTTPS, and the extent to which the websites secure the identifiers being sent to third parties. To contextualize our results, we will focus on three notable categories features in the Alexa website ranking system that have each been featured in media reports relating to the Snowden revelations and mass surveillance. First, we will look at News websites, given that they are well known users of a wide number of ad trackers known to be subject to mass surveillance[23]. Second, Computer[1] (or Technology-oriented) websites are notable given the U.S. technology industry's vocal lobbying against certain programs revealed in the Snowden documents[39]. Finally, we will look at Adult websites, as a topic-specific category whose websites often feature secure payment systems and have been targeted by NSA surveillance infrastructure in at least one program[14].

While the Alexa ranking and categorization system has known issues, including domains being listed in multiple categories, we selected this data set primarily to make our results more comparable with pre-existing work in this area[36, 1, 33]. Furthermore, while signals intelligence agencies are generally interested in people's Internet activities, people's use of popular websites and the attendant identifier transmissions provide the raw data needed to develop "pattern of life" profiles, which we discuss further below.

While earlier work has focused on the legislative responses to Snowden's revelations[8], and how the revelations influenced users to modify their practices[30], we are interested in the current, post-Snowden state of industry practices with regard to how they protect their users' browsing activities from massive government surveillance.

2 HTTP-Based Ad Tracking is a problem

Identifiers, such as those transmitted to advertisers, social networking companies, or analytics companies, are a key component of the contemporary Internet ecosystem. These identifiers are routinely stored in cookies,

which are themselves used to notify websites or services of a user's previous activities. Cookies are regularly used to track the activities of users as well as to authenticate users' activities on websites.

These identifiers have historically been transmitted in plaintext between users' web browsing clients and the servers with which they communicate. Indeed, the Internet is, by default, insecure. Websites and ad networks must take deliberate actions to encrypt the data transmitted from their servers to end users. The process of setting up a web server for HTTPS has historically been difficult and costly, though costs are decreasing[21]. Though costs implementing HTTPS have decreased, a misconfigured HTTPS configuration can cause end user clients to display security warnings, the result of which may cause users to navigate away from the misconfigured site. Moreover, all resources must be served through HTTPS before the main page will be interpreted as secure by clients. Content-driven websites, whose existence is predicated on dozens of different ad trackers being embedded on their pages, face the additional challenge of ensuring all embedded ad trackers support HTTPS. This has been highlighted in media reports as a major blocking factor in the implementation of encryption on major news websites[23]. This motivates us to ask: How prevalent is transmission encryption among top ad trackers, and is there a relationship between tracker security and the embedding pages' security?

2.1 Ad Tracking and signals intelligence

When identifier cookies are transmitted insecurely, third-parties with access to the data transmissions between client and server can log or modify the identifiers. The practice of plaintext cookie transmission has negatively affected users; it has meant, for example, that third-parties could capture and re-use social media credentials to log into social networking accounts[38, 9]. And as we discuss below, the majority of unencrypted cookies that are sent to major advertising and analytics services are regularly tracked by Western signals intelligence (SIGINT) agencies.

SIGINT agencies such as the American National Security Agency (NSA), Britain's Government Communications Headquarters (GCHQ), and Canada's Communications Security Establishment (CSE) monitor for unencrypted identifiers to track what individuals do online, to ascertain where those individuals are physically located, and to target individuals who have attracted the agencies' interest. These agencies have established listening posts' at core parts of the Internet, such as where undersea cables land (TEMPORA) or landlocked computer servers operated by telecommunications companies (AT&T), upon which programs such as EONBLUE are tasked to filter the bulk of data that is sent through the post'. CSE operates EONBLUE, which is designed to identify and collect pre-defined metadata categories[6]; while unclear what specific identifiers EONBLUE monitors for, CSE has indicated that metadata includes Facebook, Google, or Yahoo! cookie identifiers[7]. These kinds of identifiers let SIGINT agencies detect what files people are downloading[16], what websites people are visiting[13], and when correlated with identifiers linked to other persons with whom people associate.

Collected data is subsequently transferred to long-term metadata storage repositories. One of the NSA's primary long-term metadata databases is MARINA, which holds Internet-related metadata for one year[17]. Other intelligence agencies have their own metadata databases; CSE, as an example, operates the PEITHOS database that retains similar information as MARINA[5] and GCHQ possesses a database codenamed MUTANT BROTH that stores billions of intercepted cookies that are correlated with IP addresses[12]. These agencies have all actively collaborated to ensure that they can share, and use, one another's collected metadata; contemporary programs regularly draw from, or access data in, other nations' metadata (as well as content) databases.

SIGINT agencies run automated queries on the identifiers in the metadata databases to 'link' disparate identifiers that likely are associated with the same person or device; in this way, a Facebook, Gmail, Google PrefID, and Yahoo! identifier can be correlated with one another[32]. Such linking is referred to as enrichment' and is part of developing "pattern of life" profiles of Internet users en masse. Linked information is available to analysts as they monitor the kinds of files that individuals download or websites they visit or communications applications they use. Moreover, by linking these identifiers to IP addresses the agencies can roughly determine where a person is physically located. This rough geolocational awareness can subsequently be enriched if a person is using a mobile phone, and if the agency has access to cellular triangulation information, or if the agency can associate geolocational information sent by an application with the identifier. Cookies are also used by SIGINT agencies to try and identify Tor users[29].

Cookies and other unique identifiers are also used to target specific persons. In the case of the NSA, they may call on their Tailored Access Operations (TAO) unit to fire 'shots' at targets, which divert targets from the legitimate websites they are trying to visit towards ones that the NSA has compromised to install malware, or implants', on the target's device. Specifically, this entails using the QUANTUM system to direct targets towards FOXACID servers[37]. Such tactics have been used to target system administrators who operate major network operation centers[12], target employees of SIM-

card manufacturers[28], and to "degrade/deny/disrupt Tor access."[22]

A range of organizations have proposed, or adopted, new practices in light of the Snowden revelations. The Internet Advertising Bureau[25], has called for the advertising industry to transmit identifiers using HTTPS; the use of HTTPS would prevent third-parties from monitoring web users' Internet traffic while also confirming that the transmitted information was not modified in transit. The Internet Engineering Task Force, a community of technical Internet stakeholders, is developing a working group aimed at encouraging system administrators to adopt HTTPS[40]. The Internet Architecture Board published a "Statement on Internet Confidentiality" that encourages encrypted communications by default[20]. Major web browsers Chrome[3] and Firefox[2] are also exploring user experience modifications that would mark HTTP as deprecated and insecure. Email providers have accelerated their encryption of identifiers, now sending them over HTTPS between the browser and server[19, 26, 24], and StartTLS between servers[11]. Social networking companies as a matter of course authenticate their users over encrypted communications channels[35, 34, 15]. The White House has recently released a proposal exploring the notion that all federal websites should be "HTTPS only"[4].

Despite these efforts it must be asked: to what extent are the digital identifiers which are relied upon by SIGINT agencies to track and target people actually encrypted? It is to this question we now turn, in our evaluation of the percentages of identifiers that are encrypted by advertising and analytics companies.

3 Methodology

We studied the top 500 web pages in every top-level Alexa category and looked at historical support for TLS-encrypted communications and contemporary support for HTTPS in practice. We define support for TLS-encrypted communications as a host's successful completion of a TLS handshake on port 443. If a random sample of a web host's resources in our data set were all able to be successfully completed over HTTPS, we marked the host as supporting HTTPS in practice.

We obtained historical TLS handshake data[2] from the University of Michigan's SSL ecosystem study.[10] That study scanned the entire IPV4 space on a routine basis over 18 months between 2012 and 2013. For each successful TLS handshake in this data set, we extracted the CNAME(s) from the obtained HTTPS certificate and used the CNAME to map found certificates to our Alexa sample. We were then able to query the data set for successful TLS handshakes over time involving an Alexa-ranked hostname.

We loaded each sampled Alexa page in series to capture all network requests associated with a given page in order to develop our contemporary data about HTTPS support in practice. This involved using a combination of the Selenium web browser automation framework – in particular the Google Chrome web driver – and the mitmdump[3] application. That application created a proxy server that intercepted HTTPS requests sent through it and ran inline Python scripts for custom request handling logic. For each request, we captured the request and response headers (including cookies), but not the body. This process resulted in a data set that contained all the HTTP(S) resources sent and received when we accessed a website in our Alexa sample.

We further processed this data set by testing all recorded HTTP(S) resources that included a cookie (both request cookies or response cookies) for the presence of text strings that indicated the presence of a unique identifier. Every cookie consists of a set of properties, with each property having both a name and a value. We searched cookies for: a property name called "`id`" or "`pref`", a match for the regular expression "`.*id$`" (indicating a key that terminates with "id"), the presence of the string "`ident`" in a key, and several other common identifier names found in prior work[31]. We also tested cookie property values for similar strings and for the presence of GUID strings. If any of our searches resulted in a positive match we flagged the cookie as including an identifier.

We further categorized every HTTP(S) resource loaded (and by extension, the resource's cookies) on a given Alexa page as either first-party, known advertisers, or third party hosts[4]. We then tested both the Alexa webpages and a unique list of all collected resource hostnames for HTTPS support using the process described in Figure 1.

While measuring responses to TLS handshake requests on port 443 for these hosts is a common method for testing HTTPS configurations [10, 27, 18] we were more interested in HTTPS in practice in the web browser and the relative accessibility of resources (and particularly those carrying identifying cookies) over HTTPS. Specifically, we asked: Do paths that return 200 OK over HTTP also return 200 OK over HTTPS? Does HTTPS "downgrade" to HTTP and redirect browsers from secure to insecure requests? Or does the host support HTTPS by default and redirect browsers from insecure to secure requests? In light of these questions, we are interested in whether or not web-based resources are accessible using HTTPS as opposed to previous work cited above that has focused on TLS protocol versions, cipher suites, or other security issues.

To test the aforementioned questions, we ran them against a random sample of up to five URL paths for the

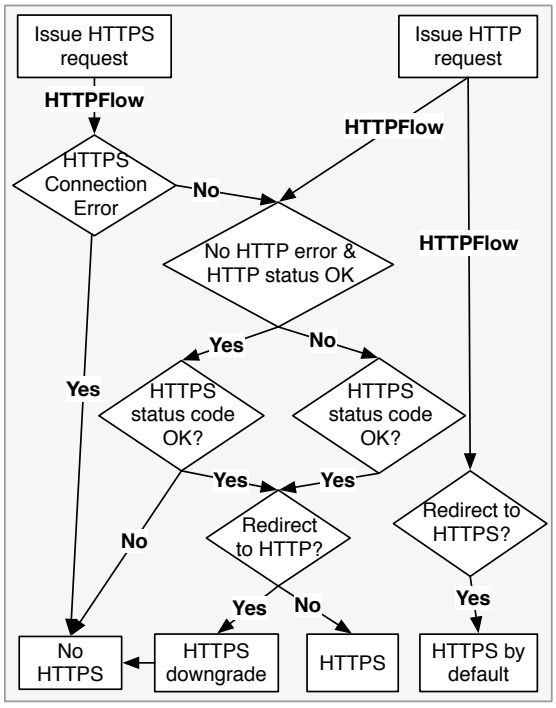

Figure 1: Testing process for determining practical HTTPS support of a given hostname.

given host that we had previously collected. We then reconciled the results for each of these URLs into a final measurement for the hostname, checking for consistent results. Where there was a conflict between different tested URLs we erred towards labeling hosts as not fully supporting HTTPS. We feel this reconciliation process was reasonable given that the HTTPS security model can be compromised if a single resource is served insecurely.

The results of the reconciled measurements let us examine the practical HTTPS support level of all the resources loaded on the top 500 Alexa websites in each category and the relationship between page host security and third party identifier cookie security. We turn to these examinations in the following section.

4 Data

In this section, we first examine TLS support among various Alexa-categorized websites over time, noting which category's support increased the most over time. Next, we look at each website category's levels of HTTPS support in practice, noting which category includes the most (and least) websites that redirect users to HTTPS. We then examine the most commonly-implemented Ad trackers across each category, and noting the proportion of ID transmissions with those trackers that have been secured. Finally, we look at the relationship between a web

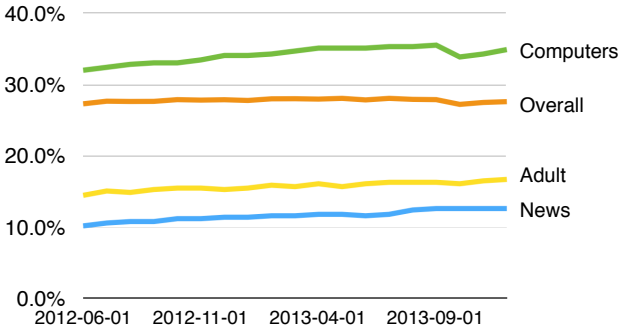

Figure 2: TLS handshake success percentage (2012-06 – 2014-01) by Alexa category

page's first-party security and the security of the third party trackers with which a web page transmits identifiers.

4.1 Alexa domains

To better contextualize our findings, in this subsection, we present examples from each of the Alexa categories of interest. The top news website in our sample of 500 is: www.reddit.com. The top Computer website is www.google.com. The top Adult site is www.xnxx.com.

4.2 General HTTPS adoption trends

Our findings show that between June 2012 and January 2014[5], the overall percentage of websites in the top 500 of an Alexa category that completed a TLS handshake increased 1.1%, to 27.6% overall.

Looking at individual categories, we find that our categories of interest, News, Adult, and Computers, are the categories with the largest changes over the time period, at 23.5%, 15.3%, and 8.2%, respectively. Looking specifically at the 6 months before and 6 months after the Snowden revelations of June 2013, we find the News category TLS handshake response growth rate increased by 5%, the largest of any category. However, nine of the sixteen surveyed categories saw their TLS growth rates diminish in the months following the Snowden revelations. Both Adult and Computer sites saw slight reductions in their TLS response growth rates.

While the News category had the highest growth rate, in terms of overall TLS support, News had the second lowest percentage, at 12.6% (as of January 2014), below all but Sports, at 8.4%.

4.3 HTTPS support in practice by category

We measured the varying degrees of support for HTTPS in practice in each Alexa category based on the process described in our methodology section.

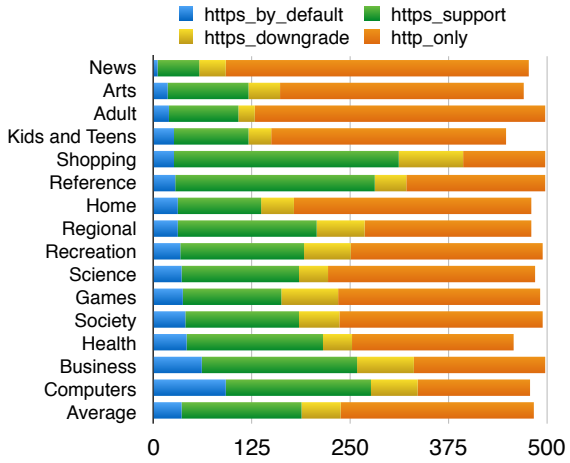

Figure 3: Support for HTTPS in practice by category

We find that 7.2% of websites in our sample have HTTPS by default enabled. In contrast, 10.2% of websites and downgrade visitors from HTTPS to plaintext HTTP. 31.6% support HTTPS if the user explicitly navigates to an HTTPS URL, and 51% have no practical support for HTTPS.

Looking at individual categories, we see that News sites are least likely to direct browsers to HTTPS by default, with only five of 500 websites doing so. Adult sites perform slightly better, with 20 websites enabling the feature. At the other end of the spectrum are Computer websites, with 91 out of 500 encrypting communications with no user intervention required. Over 50% of Shopping websites support HTTPS though this category also has the highest number of sites (82) downgrading HTTPS.

4.4 Ad tracker prevalence by category

Our data demonstrates which ad trackers are the most popular in each category. We made our determinations by performing a database query that counted the number of distinct webpages from which an identifier was transmitted to a unique third party. This query was run for each website category.

Our results show which ad trackers are most popular, whether or not those trackers support HTTPS, and the percentage of pages in our sample that communicated an identifier using HTTPS. We also determined the percentage of webpages communicating with those trackers that support HTTPS themselves.

For example, the most widely-used ad tracker in the News category is `b.scorecardresearch.com` (ScoreCard Research). This tracker exchanges a unique identifier with 291 of the top 500 News sites and does not support HTTPS; at the time of writing, HTTPS communication with the host resulted in an invalid common name certificate error. Scorecard Research is also the most common tracker in the Kids and Teens and Arts categories, and in the top five trackers for ten other webpage categories.

The second most popular tracker in the News category is `cm.g.doubleclick.net` (DoubleClick); this host is operated by Google. While the tracker supports HTTPS only 56% of the 285 pages using the tracker transmitted its identifier using HTTPS. Since only 8.7% of those pages served their first-party content with HTTPS, 47% of the sampled websites, while insecure themselves, often used HTTPS to communicate with DoubleClick.

In the News category, we found three trackers that supported HTTPS in practice, yet none of the pages communicated securely with those hosts. As shown in Table 1, `sync.mathtag.com`, `ad.turn.com`, and `image2.pubmatic.com` were involved in unique ID transmissions on 215, 194, and 168 of the top 500 News websites, respectively. Our analysis showed that those transmissions can resolve successfully over HTTPS.

We found that the Science category features the highest percentage of pages (59%) that communicate identifiers securely with the category's top ten ad trackers. This finding may be may be related to the fact that the category has the second lowest median number of pages (60.5) that transmit to the top ten identifier hosts. News, in contrast, only communicates securely with trackers 23% of the time and has the highest median number of pages (204.5) that transmit to the top ten identifier hosts in the category.

Across all categories, we saw that an average of 26.5% websites that transmitted unique identifiers with top ad trackers support HTTPS on their own host. While 32% of websites communicate securely with the top 10 ad trackers, 80% of the trackers themselves support HTTPS, indicating that many more websites in the category could be communicating securely with ad trackers. When looking at a specific tracker in the top 10, `apis.google.com`, we can see a stark difference – 100% of communications with the host were secured. This indicates that HTTPS by default, which is practiced by `apis.google.com`, can make a substantial difference in the overall number of pages that securely transmit identifiers to top trackers.

4.5 First party and third party HTTPS

For each category, we grouped websites by their level of HTTPS support in practice; for each grouping we assessed the percentage of identifiers that were transmitted securely. We found that websites that communicated over HTTPS by default had near 100% rates of encrypted third party communications. This finding was expected

Host	HTTPS?	News		Adult		Computers	
		IDs	HTTPS	IDs	HTTPS	IDs	HTTPS
b.scorecardresearch.com	No	291	0.00%	8	0.00%	104	0.00%
cm.g.doubleclick.net	Yes	285	56.14%	14	42.86%	140	42.14%
ib.adnxs.com	Yes	246	6.10%	15	13.33%	130	16.92%
googleads.g.doubleclick.net	Yes	237	72.57%	5	60.00%	102	70.59%
sync.mathtag.com	Yes	215	0.00%	9	0.00%	67	5.97%
ad.turn.com	Yes	194	0.00%	8	0.00%	32	0.00%
match.adsrvr.org	Downgrades[6]	174	37.36%	13	15.38%	46	23.91%
image2.pubmatic.com	Yes	168	0.00%	12	0.00%	58	0.00%
x.bidswitch.net	Yes	124	8.06%	7	14.29%	74	21.62%
apis.google.com	Yes	101	100.00%	40	100.00%	91	100.00%

Table 1: Top trackers in the News, Adult, and Computer categories, by number of pages that transmitted a tracking ID

because the HTTPS security model is predicated on all first- and third-party resources being delivered over an encrypted channel. However, most categories in this group have one to six instances of an HTTPS-enabled webpage insecurely transmitting an identifier to a third party.

For ten of sixteen Alexa categories, websites that did not support HTTPS in practice had a higher rate of securely transmitting identifiers than websites that downgraded browsers from HTTPS. The difference was most pronounced for Science websites, where 70.5% of websites that do not support HTTPS transmitted identifiers securely, while only 43.3% of HTTPS downgrade Science websites transmitted identifiers securely. When examining our results in aggregate we found that the difference is less pronounced, with 67.3% of websites that did not support HTTPS transmitting identifiers securely, while 66% of HTTPS downgrade websites transmitted identifiers securely. Across every category, HTTPS downgrade grouped sites had the greatest number of ad trackers per page compared to the other groupings. HTTPS by default websites, in all but one category, had a greater number of ad trackers per page than HTTP-only websites.

5 Discussion

Our findings reveal that there has been an small increase in the adoption of HTTPS following the start of Edward Snowden's revelations, with the News category increasing by the greatest amount. However, sites with the largest numbers of trackers often also have the lowest rates of securely transmitting identifiers over HTTPS. When we examine the Adult, News, and Computers website categories we find that News, with the most trackers, had the lowest percentage of secure ID transmissions to third parties. This suggests that the mere fact a website category features prominently in discussions around mass surveillance is not enough to predict whether or not it encrypts identifier transmissions to third parties. The number of third party dependencies, as well as the security of the first party itself play major roles, as will be illustrated in our below discussion of the News, Adult, and Computer website categories.

Our results show that while News websites appear to have increased their TLS support over time at a greater rate than all other categories, they have little practical support for HTTPS browsing, leaving their readership open to passive surveillance of their readership habits. Furthermore, the top ten ad trackers in the News category appear on the highest number of websites (on average, 204/500) of all categories, identifier transmissions to which are only encrypted 23% of the time. These sites therefore leave their readers vulnerable to actors with network access to not only monitor their own readership, but to chain their readers' habits across other websites.

Adult websites have similarly increased their TLS support over time, yet perform poorly in terms of HTTPS browsing in practice, with only 21.6% of websites in the category supporting encrypted browsing. The average Adult website transmitted unique identifiers to 4.7 third parties, the lowest of all categories, and well below News, which transmitted identifiers to the highest number of third party hostnames (11.5). Each of our overall sample's top ten ad trackers appear an average of 12.5 websites among the top 500 Adult sites, with 43% of websites in that category sending identifiers doing so using HTTPS. The relatively low frequency of trackers across websites may be partially explained by the relatively small number of ad trackers embedded on the average Adult website. Additionally, the ten most frequently occurring ad trackers in the Adult category include four trackers not included in the overall top ten list of trackers, which can also help explain the relatively low number of identifier transmissions to the overall top ten trackers. Unlike the News category, the most frequently-occurring

tracker in the Adult category, apis.google.com, supports HTTPS by default, a stark contrast from News' top tracker, b.scorecardresearch.com, which does not support HTTPS at all.

Computer websites historically supported TLS at a higher rate than both News and Adult websites. This differentiation is reinforced when examining HTTPS in practice; our findings show that over 50% of Computer websites support encrypted browsing, and of all categories, Computers had the highest rate of HTTPS by default being enabled. On average, Computer websites transmitted an identifier to 10.3 unique hosts, the lowest result of our categories of Interest. Furthermore, 44% of Computer websites that transmitted IDs to the top ten ad trackers did so securely, higher than both Adult and Business categories. Given that over 50% of Computer websites support HTTPS browsing, this 44% appears relatively underwhelming when compared to Adult websites, with 21.6% HTTPS browsing support, and 43% of ID transmissions being done securely. Given that Computers has the lowest average number of trackers identifiers with whom identifiers were communicated, it may be relatively less complicated for Computer sites to implement HTTPS for their trackers than other categories.

We note that the adoption of HTTPS by advertising, social networking, and analytics services may foster the adoption of securely transmitting identifiers. Currently, major trackers such as b.scorecardresearch.com and dsum.casalemedia.com lack such support. b.scorecardresearch.com is used by Adult, News, and Computers websites on 1.6%, 22%, and 8.2% of sites, respectively. And dsum.casalemedia.com is used by the same website categories on 1.4%, 33%, and 11.6% of the categories' respective pages. As a result, webpage operators wanting to adopt HTTPS for their own pages, but who must rely on these particular trackers for business reasons, may be prevented from encrypting identifiers because of these trackers' failure to adopt HTTPS.

More identifiers may be encrypted if the providers of those identifiers, such as advertisers and analytics companies, supported HTTPS by default. While many of the identifiers associated with the top ten ad trackers can be encrypted but are not necessarily encrypted, were companies to provide HTTPS-only APIs and gradually phase out non-HTTP APIs the encryption rates would presumably increase over time. In the interim, webpage operators could re-write embed codes for many of the top trackers, such as image2.pubmatic.com, ad.turn.com, and cm.g.doubleclick.com, to force the transmission of identifiers to use HTTPS. Were such practices adopted by tracker companies alongside webpage operators, a mutually reinforcing signal that HTTPS adoption was a significant business concern would be emitted, reducing the number of tracking identifiers that were transmitted in plaintext while emphasizing that the advertising and analytics industries regard the security of identifiers as a basic business best practice.

6 Limitations

Our data collection method using Selenium and mitmproxy was bound to a single user-agent at a single network vantage point in Toronto, Canada. The third party resources loaded on given pages will vary based on those factors, but we did not account for this in our study. In the future, we hope to conduct the study using both Firefox and Chrome Selenium drivers (and perhaps others), as well as from multiple network vantage points.

Our automated method for testing HTTPS support in practice is a work in progress. The method has an error rate of 6%. This error rate has a 95% confidence level and 10% margin of error. We determined this error rate by randomly sampling 100 tested hostnames and manually checking their HTTPS support in practice, and comparing those to our automated results. For the top 10 ad trackers in our study, 2 ad trackers were incorrectly marked as not supporting HTTPS, when in fact they did. We corrected this in our analysis after manually checking the hostname and a few random paths.

Our process for testing cookies for the presence of unique identifier strings does not currently analyze cookies that employ data obfuscation through base64 or other encoding methods. Examining common obfuscation techniques in advertiser cookies would be an interesting area of future work.

Finally, our use of the University of Michigan SSL ecosystem data introduces noise into our findings. While that project routinely scanned every public IPv4 address in the world, the scans only picked up one certificate per IP address. It is likely that many certificates associated with shared IP addresses were missed, and we cannot be sure to what extent this noise is distributed across categories. Future work will mitigate this limitation, as discussed below.

7 Conclusions

Our findings suggest that though numerous prominent news stories have been released showing how signals intelligence agencies use identifiers to monitor, geolocate, and target web users, many website operators have not chosen to secure their readers from this mode of surveillance. The result is that Western signals intelligence agencies retain much of their surveillance capacities despite surveys showing that many users claim to have

modified their own practices to try and limit the extent of state surveillance to which they are subject[30]. Moreover, even though many trackers now support HTTPS transmissions many webpage operators themselves do not support secure transmissions of identifiers; the result is that web users are only somewhat better protected from mass surveillance relying on identifiers, today, compared to six months prior to the Snowden revelations breaking in June 2013.

The technical testing infrastructure developed for this project will make it simple to run subsequent measurements on a regular basis. This will enable us to have more control over our data set than simply relying on the University of Michigan data set, which was not collected with our research questions in mind. We believe a longitudinal look at HTTPS support in practice across website categories and the levels of secure transmission of unique identifiers will be a valuable resource in understanding the movement towards encrypting the web.

References

[1] AGER, B., MÜHLBAUER, W., SMARAGDAKIS, G., AND UHLIG, S. Comparing DNS resolvers in the wild. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement* (2010), ACM, pp. 15–21.

[2] BARNES, R. Deprecating non-secure HTTP. https://blog.mozilla.org/security/2015/04/30/deprecating-non-secure-http/, April 2015.

[3] CHROMIUM. Marking HTTP as non-secure. https://www.chromium.org/Home/chromium-security/marking-http-as-non-secure.

[4] CIO.GOV. The HTTPS-only standard. https://https.cio.gov/.

[5] COMMUNICATIONS, T. L. CSE codewords and abbreviations. http://electrospaces.blogspot.ca/p/cse-codewords-and-abbreviations.html, April 2015.

[6] CSE. CSEC SIGINT cyber discovery: Summary of the current effort. https://www.christopher-parsons.com/Main/wp-content/uploads/2015/02/cse-csec-sigint-cyber-discovery.pdf, 2010.

[7] CSE. LEVITATION and the FFU hypothesis. https://www.christopher-parsons.com/Main/wp-content/uploads/2015/02/cse-presentation-on-the-levitation-project.pdf, 2012.

[8] DAVIES, S. A crisis of accountability: A global impact of the impact of the snowden revelations. https://citizenlab.org/wp-content/uploads/2014/06/Snowden-final-report-for-publication.pdf, June 2014.

[9] DAVIS, B. Unencrypted cookies make Wordpress accounts vulnerable over open networks. http://www.neowin.net/news/unencrypted-cookies-make-wordpress-accounts-vulnerable-over-open-networks, May 2014.

[10] DURUMERIC, Z., KASTEN, J., BAILEY, M., AND HALDERMAN, J. A. Analysis of the HTTPS certificate ecosystem. In *Internet Measurement Conference, IMC'13, Barcelona, Spain, October 23-25, 2013* (2013), K. Papagiannaki, P. K. Gummadi, and C. Partridge, Eds., ACM, pp. 291–304.

[11] ECKERSLEY, P. New Gmail data shows the rise of backbone email encryption. https://www.eff.org/deeplinks/2014/06/new-gmail-data-shows-rise-backbone-email-encryption, June 2013.

[12] GALLAGHER, R. Operation Socialist: The inside story of how british spies hacked belgium's largest telco. https://firstlook.org/theintercept/2014/12/13/belgacom-hack-gchq-inside-story/, December 2014.

[13] GREENWALD, G. Xkeyscore: Nsa tool collects 'nearly everything a user does on the internet'. http://www.theguardian.com/world/2013/jul/31/nsa-top-secret-program-online-data, June 2013.

[14] GREENWALD, G., GRIM, R., AND GALLAGHER, R. Top-secret document reveals NSA spied on porn habits as part of plan to discredit 'radicalizers'. www.huffingtonpost.com/2013/11/26/nsa-porn-muslims_n_4346128.html, 2013.

[15] HICKS, M. A continued commitment to security. https://www.facebook.com/notes/facebook/a-continued-commitment-to-security/486790652130, January 2011.

[16] HILDEBRANDT, A., PEREIRA, M., AND SEGLINS, D. CSE tracks millions of downloads daily: Snowden documents. *CBC News* (April 2015).

[17] ISRAEL, T. Foreign intelligence in an inter-networked world: Time for a re-evaluation. In *Law, Privacy and Surveillance in Canada in the Post-Snowden Era*, M. Geist, Ed. Ottawa University Press, Forthcoming.

[18] LEVILLAIN, O., ÉBALARD, A., MORIN, B., AND DEBAR, H. One year of SSL internet measurement. In *Proceedings of the 28th Annual Computer Security Applications Conference* (New York, NY, USA, 2012), ACSAC '12, ACM, pp. 11–20.

[19] MCCULLAGH, D. How Web mail providers leave door open for NSA surveillance. http://www.cnet.com/news/how-web-mail-providers-leave-door-open-for-nsa-surveillance/, June 2013.

[20] MORGAN, C. IAB statement on internet confidentiality. https://www.iab.org/2014/11/14/iab-statement-on-internet-confidentiality/, November 2014.

[21] NAYLOR, D., FINAMORE, A., LEONTIADIS, I., GRUNENBERGER, Y., MELLIA, M., MUNAFÒ, M., PAPAGIANNAKI, K., AND STEENKISTE, P. The cost of the S in HTTPS. In *Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies* (2014), ACM, pp. 133–140.

[22] NSA. Tor stinks. https://edwardsnowden.com/wp-content/uploads/2013/10/tor-stinks-presentation.pdf, 2007.

[23] PETERSON, A. News sites could protect your privacy with encryption. here's why they probably won't. *Washington Post December 11* (2013).

[24] PETERSON, A. Yahoo implemented SSL late. and they didn't even do it well. http://www.washingtonpost.com/blogs/the-switch/wp/2014/01/09/yahoo-implemented-ssl-late-and-they-didnt-even-do-it-well/, January 2014.

[25] RIORDAN-BUTTERWORTH, B. Adopting encryption: The need for HTTPS. http://www.iab.net/iablog/2015/03/adopting-encryption-the-need-for-https.html, March 2015.

[26] RISEN, T. Nsa surveillance spurs tech giants to add encryption. http://www.usnews.com/news/articles/2013/11/01/nsa-surveillance-spurs-tech-giants-to-add-encryption, 2013.

[27] RISTIC, I. State of SSL. *Talk at InfoSec World* (2011).

[28] SCAHILL, J., AND BEGLEY, J. The great SIM heist: How spies stole the keys to the encryption castle. https://firstlook.org/theintercept/2015/02/19/great-sim-heist/, May 2015.

[29] SCHNEIER, B. Attacking Tor: how the NSA targets users' online anonymity. http://www.theguardian.com/world/2013/oct/04/tor-attacks-nsa-users-online-anonymity, October 2013.

[30] SCHNEIER, B. Over 700 million people taking steps to avoid nsa surveillance. *Lawfare* (Decmber 2014).

[31] SOLTANI, A., CANTY, S., MAYO, Q., THOMAS, L., AND HOOFNAGLE, C. J. Flash cookies and privacy. In *AAAI Spring Symposium: Intelligent Information Privacy Management* (2010).

[32] SOLTANI, A., PETERSON, A., AND GELLMAN, B. NSA uses Google cookies to pinpoint targets for hacking. http://www.washingtonpost.com/blogs/the-switch/wp/2013/12/10/nsa-uses-google-cookies-to-pinpoint-targets-for-hacking/, December 2013.

[33] STEBILA, D. Reinforcing bad behaviour: the misuse of security indicators on popular websites. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction* (2010), ACM, pp. 248–251.

[34] TUNG, L. Twitter enforces SSL encryption for apps connecting to its API. http://www.zdnet.com/article/twitter-enforces-ssl-encryption-for-apps-connecting-to-its-api, January 2014.

[35] TWITTER. Making twitter more secure: Https. https://blog.twitter.com/2011/making-twitter-more-secure-https, March 2011.

[36] VAN GOETHEM, T., CHEN, P., NIKIFORAKIS, N., DESMET, L., AND JOOSEN, W. Large-scale security analysis of the web: Challenges and findings. In *Trust and Trustworthy Computing*. Springer, 2014, pp. 110–126.

[37] WEAVER, N. Our government has weaponized the internet. here's how they did it. http://www.wired.com/2013/11/this-is-how-the-internet-backbone-has-been-turned-into-a-weapon/, November 2013.

[38] WHALEN, T. Fleeced by firesheep? http://blog.priv.gc.ca/index.php/2010/11/04/fleeced-by-firesheep, 2010.

[39] WILHELM, A. Tech giants pile on in support of the NSA-curtailing USA FREEDOM Act. *TechCrunch* (May 2015).

[40] YORK, D. Introducing a new deploy360 topic: TLS for applications. http://www.internetsociety.org/deploy360/blog/2014/02/introducing-a-new-deploy360-topic-tls-for-applications/, February 2014.

Notes

[1] Sites categorized as "Computers" by Alexa tend to focus on digital technology in some fashion.

[2] See https://scans.io/study/umich-https for the data.

[3] mitmdump is a tool included with mitmproxy, free software available at: https://mitmproxy.org/

[4] We tested hostnames against the Disconnect list of known online trackers: https://services.disconnect.me/disconnect.json

[5] The duration of the University of Michigan's study

[6] `match.adsrvr.org` appears to be a redirection portal, and downgrades HTTPs in some circumstances.