Scandinista!

Analyzing TLS Handshake Scans and HTTPS Browsing by Website Category

Andrew Hilts
Citizen Lab, University of Toronto
Open Effect
315 Bloor Street West
Toronto, ON, Canada
andrew@openeffect.ca

ABSTRACT

This paper analyzes whether websites began adopting TLS encryption at a faster rate following the Snowden revelations of mass surveillance. Overall, we find large variations in TLS adoption across different website categories. We furthermore studied HTTPS support in the browser, looking at general support for HTTPS browsing, default upgrades from HTTP to HTTPS, and default downgrades from HTTPS to HTTP. We found overall that the best practice of "HTTPS by default" (redirecting plaintext to encrypted transmissions) is rarely implemented, and that HTTPS downgrades are much more widespread, across all website categories. Findings such as these are important because they may be used for targeted advocacy efforts meant to improve the privacy and security practices of entire categories of websites.

Categories and Subject Descriptors

[**Security and privacy**]: Network Security, Human and societal aspects of security and privacy—*Web protocol security, Privacy protections*; [**Social and professional topics**]: Computing / technology policy—*Surveillance*

Keywords

Privacy, HTTPS, TLS, Encryption, Protocols, Security, Surveillance

1. INTRODUCTION

Two years have passed since former National Security Agency (NSA) contractor Edward Snowden disclosed a massive trove of documents to a group of journalists, documents which have provided insight into a wide range of state surveillance programs. The resulting news stories have led to increased awareness of mass surveillance programs conducted by NSA and its Five Eyes spy alliance partner countries. A report conducted by the Centre for International Governance Innovation and Ipsos[6] estimated that 750 million people are taking greater steps to avoid software and websites that might put their data at risk. Furthermore, the study found that 39% of survey respondents had taken steps to protect their privacy in direct response to the Snowden revelations.

While individuals have taken steps in protecting their privacy, what about organizations operating popular websites? In response being associated with the PRISM program for obtaining user data on targets, large technology companies have taken steps to secure internal data links with HTTPS[15], issue more detailed transparency reports[25], lobby governments to curtail surveillance[21], and generally attempt to position themselves as protectors of their customers' privacy[26]. However, Silicon valley technology giants make up but a small fraction of the Web.

Media reports have highlighted some privacy advocates' concerns regarding the lack of HTTPS support by websites that provide content in exchange for serving targeted advertising to their readers[19]. The insecurity of content-driven websites leaves readers' web browsing habits vulnerable to being captured by mass surveillance infrastructure. While studies have looked at overall HTTPS certificate trends across the IPv4 space[9, 10], there is little data currently published on the adoption of HTTPS by particular categories of websites. The concerns of privacy advocates could either be bolstered or assuaged by data that describes the relative (in)security of content-driven websites compared to other web properties.

1.1 HTTPS and Surveillance

HTTPS is the primary means by which data transmissions between a web browser and server are secured. The communications protocol uses TLS or SSL-based cryptographic protocols to both encrypt the content and verify the authenticity of data transmissions. Certificates issued by servers and signed by certificate authorities help to guarantee that websites are who they claim to be. The HTTPS ecosystem is shared, complex infrastructure[3] that is foundational for a secure web, with security vulnerabilities in shared code can affect millions of websites[8], and subject to a variety of attacks[23, 4].

In recent years, major online services operated by Google[14], Facebook[20], and Twitter[24] have moved to offer their users HTTPS "by default". HTTPS by default is a server configuration that redirects users who request a resource over plaintext HTTP to the same resource served over HTTPS. The

effect of this server behaviour is to enforce encrypted communications without requiring the end user to take any steps to secure themselves. The embrace of privacy-by-default settings reflects behavioural economics research suggesting that end users will often ignore privacy-enhancing options if they are not the default or if they are inconvenient to enable[1].

HTTPS by default helps protect against both mass and targeted surveillance of communications in transit. For example, web browsing history and insecurely transmitted cookie-based identifiers play a large role in NSA surveillance program MARINA[12]. Identifiers stored in unencrypted cookies enable surveillers to chain web browsing habits within and across different IP addresses to specific individuals. Programs such as these use browsing and cookie data to build pattern-of-life profiles of surveilled individuals, potentially inferring political views, demographics, sexual orientation, health and wellness, among other attributes[11, 18].

By encrypting data in transit, including the path of the requested resource, HTTPS reduces the amount of plaintext information available for collection by network snoops[5]. By validating the identity of the server and authenticity of the data sent to a web browser, HTTPS combats man-in-the-middle attacks[16]. When employed as the default protocol for a website or service, HTTPS brings the above benefits to people who might not otherwise take steps to protect themselves.

1.2 HTTPS in the Wild

Implementing HTTPS is difficult and costly for many websites, though costs are decreasing[17]. Creating certificate signing requests, purchasing a certificate, configuring a web server, and keeping up with the latest attacks on TLS and the broader HTTPS ecosystem all require expertise and technical access that many smaller websites may lack. Forthcoming efforts such as the Electronic Frontier Foundation and partners' "Let's Encrypt" project[22] aim to reduce the steps required in creating certificates and configuring servers.

Content-driven websites, in particular, are typically coupled with dozens of third party resources (primarily advertising tracking scripts or content distribution networks). Websites cannot effectively implement HTTPS on their properties without first ensuring that all third party resources can also be served through the secure protocol. Media reports have claimed that this reliance on third party resources is one of the major barriers in content-driven website's adoption of HTTPS[19].

By comparing TLS handshake adoption across categories, we can determine the degree to which advertising domains' TLS support varies from content-driven sites, such as news media or recreation categories. This information can aid privacy scholars and advocates in identifying barriers to HTTPS adoption.

1.3 Research Questions

This study investigates the following questions: How, if at all, have different categories of websites changed their rate of HTTPS implementation in the months following the Snowden disclosures? How do different website categories vary in terms of their support for HTTPS by default? In answering these questions, we contribute to the discussion about how websites protect their readership from mass surveillance in the post-Snowden era.

2. METHODOLOGY

Our analysis is based on three primary data sources that are compiled into a Postgres database. First, we inserted longitudinal TLS handshake results for each address in the entire IPv4 space. This data was obtained from the University of Michigan's HTTPS Ecosystem project website[1]. Second, we inserted data about the top 100,000 websites in the Alexa ranking system, including a domain's ranking and industry category. Third, we collected data about a website's HTTPS support for basic web browsing by running a test on each categorized domain in the Alexa top 100,000. We also included an additional web domain category, Advertisers, by creating a table using a public list of known ad trackers maintained by the Disconnect privacy company[7].

The University of Michigan data set contains the results of 110 scans of the entire public IPv4 address space conducted between June 2012 and January 2014. The scans collected data about how hosts listening on port 443 responded to TLS handshakes. This data set also included the TLS certificates that were received during the handshakes.

Alexa is a web measurement company that ranks the global popularity of websites using a sample of geographically distributed users who report web browsing data to the company using a browser extension[2]. The company clusters many of these sites into one of 17 industry categories such as News, Computers, Recreation, Adult, and Education.

We measured changes in different industry categories' HTTPS support by relating historical TLS handshake scan results with Alexa categorizations. To relate these datasets, we queried the University of Michigan's certificate table, which contains records of all certificates obtained during its TLS handshake scans. We extracted the hostname from the certificate subject's common name listing and stored that in an associative table alongside the certificate ID. Since we were interested in the typical web browsing behaviour of individuals, we only stored certificates issued for bare domains, the www subdomain, or wildcard certificates. Certificates for other subdomains were not included. This associative table enabled us to run queries on specific domains that link Alexa categorizations with historical TLS handshake results.

While the University of Michigan data set provides us with longitudinal data about basic TLS implementations for web servers it does not help to determine if regular web browsing operations can be served through HTTPS, nor whether or not such browsing is HTTPS by default. TLS-supporting web servers may still redirect users issuing HTTPS requests to insecure equivalents. To address this issue we wrote a Python script to issue HTTPS HEAD requests to the root of each categorized domain in the Alexa top 100,000. We recorded if the request was successful, if it was redirected away from an HTTPS connection to a plaintext HTTP one, and if a plaintext HTTP request is redirected to an HTTPS connection. While the data we collected in early March 2015

[1]Located at https://scans.io/study/umich-https

Table 1: Category distributon of Alexa-ranked domains in our sample

Category	Domains
Adult	155
Advertising	2156
Arts	994
Business	2235
Computers	2360
Games	413
Health	247
Home	284
Kids & Teens	160
News	426
Recreation	526
Reference	1490
Regional	2913
Science	426
Shopping	961
Society	831
Sports	405
World	13592

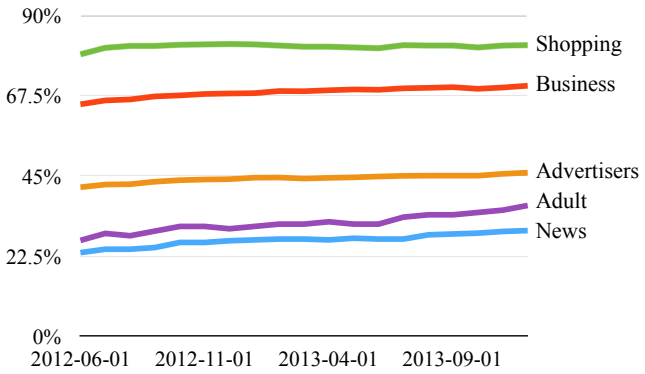

Figure 1: Percentage of Alexa domains responding to TLS handshakes from June 2012 to January 2014

is not longitudinal, they allow us to measure the general ratio between responding to TLS handshakes and actually supporting the downloading of a website over HTTPS.

3. DATA AND RESULTS

We collected data about from the top 100,000 sites in Alexa's ranking system. Unfortunately, many of the websites in the top 100,000 are not categorized. As a result, the sample we worked with contains only 28,418 domains across the 17 Alexa categories. Table 1 shows the category distribution among our data sample.

3.1 TLS Handshake Responses

Every industry category in our categorized data shows a steady increase in the percentage of domains completing a TLS handshake between June 2012 and January 2014. Overall we found that the TLS response rate for Alexa-ranked websites increased from 2.16% to 2.7% over this time period.

As shown in Figure 1, certain website categories have significantly higher TLS handshake response rates than other categories. Shopping sites have the highest TLS handshake response rate at 82% as of January 2014, though this is likely attributable to their e-commerce requirements. These sites are followed by Business at 70%, and Reference (Education) at 67%. At the lower rates, we find Arts at 26%, News at 30%, and Adult at 36%. The relatively low score for Adult sites is surprising given that the industry has a large amount of paid content. Outside of the Alexa categorizations, we find Advertisers at a 45% TLS handshake completion rate.

To assess the relationship between popularity and TLS handshake responses, we divided sites into 10 groupings based on their Alexa rank. We found a clear inverse relationship between TLS handshake responses rates and domain rank grouping (with a slope of -1.44% per 10k ranking bucket), and illustrated in Figure 4. Looking at data from the most recent January 2014 scans, we find 55% of the top 10,000 websites responded to a handshake, 35% of the 40-49k bucket responded, and 32% of 80-89k responded.

The inverse relationship could be explained by the supposition that the more popular a website is, the more resources it has at its disposal to devote to security.

3.1.1 TLS Adoption Rates Before and After Snowden

We looked at the growth of website category TLS support rates before and the Snowden revelations. To do this, we determined the growth rate of the number of domains that respond to TLS handshake requests in both the seven months prior to June 2013 and the subsequent seven months. Seven months was selected as the timeframe because the University of Michigan HTTPS scan data ends in January 2014, seven months after June 2013.

This analysis let us see if more websites in a given category took steps to secure their properties post-Snowden compared to pre-Snowden. Looking at all categorized domains, we find the TLS handshake response growth rate is 1.75% higher post-Snowden than pre-Snowden, which is low compared to our results for specific categories. Looking at website categories, as illustrated in Figure 2, we found that Adult sites' growth rate had the highest increase, at 14.24%, News sites at 4.12%, and Kids and Teens sites at 3.76%. Health websites actually saw their growth rates decline slightly in the 7 months following the initial Snowden revelations. Upon closer examination, the Adult website's growth rate for this period appears artificially inflated, given the aberrant drop in growth briefly occurring in the months prior to the revelations.

We do not have data to determine whether these growth rate changes have causal relationships with the Snowden revelations. For example, our data cannot reliably demonstrate whether the relatively large change in the Adult website industry's support for TLS handshakes in the Adult category is related to the Snowden disclosures that alleged NSA agents planned to monitor the pornography habits of suspected 'radicalizers'[13]. Such a change could also be explained by increased adoption of paid features, or simply an artifact of the relatively small sample size of Adult sites compared to other categories, as discussed in the Limitations section. A deeper investigation into the increase in

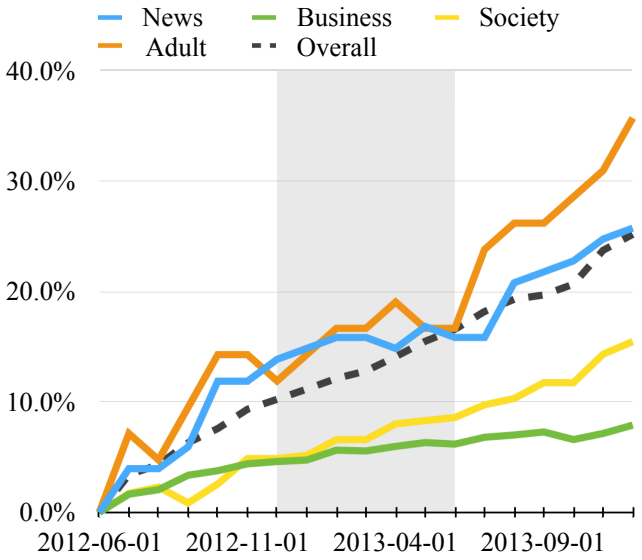

Figure 2: Category TLS Handshake response growth over time. Highlighted area indicates 6 months prior to Snowden revelations.

TLS handshake responses by Adult sites represents an area of future work.

3.2 HTTPS Browsing Support

Our analysis of the University of Michigan HTTPS ecosystem data by Alexa category provided an overview of industry TLS handshake support trends. However, simply responding to a TLS handshake request on port 443 is not the same as supporting HTTPS for end users of a website. For this reason, in early March 2015, we ran our own series of tests for end-user HTTPS support that were described in our methodology section. These findings revealed how many websites supported "HTTPS by default", among other practices.

3.2.1 Support for HTTPS browsing by industry

We found that categorized websites in the Alexa top 100,000 that will respond to a TLS handshake are overall 13% more likely to redirect a user *away* from a secure connection than to let them browse securely. More specifically, 53% of TLS-supporting websites will redirect a user away from a secure connection, while only 6% will redirect a user from an insecure connection to a secure one (HTTPS by default). When looking at all sites in the Alexa data set, not just TLS-supporting ones, as shown in Figure 3, we find that 0.9% of sites support HTTPS by default, 8.5% will redirect a user away from HTTPS, and 7.5% will enable a user to browse HTTPS if they explicitly elect to. Finally, 84% of sites in our data set do not support HTTPS at all, responding to our request with an error.

The category that supports HTTPS by default at the highest rate is Science, at 15.12% of TLS-supporting sites, followed by Computers, at 12.5%. Shopping websites have the highest rate of redirecting a user away from HTTPS. While shopping sites may use TLS connections for payments processing, they do not prioritize the general browsing security

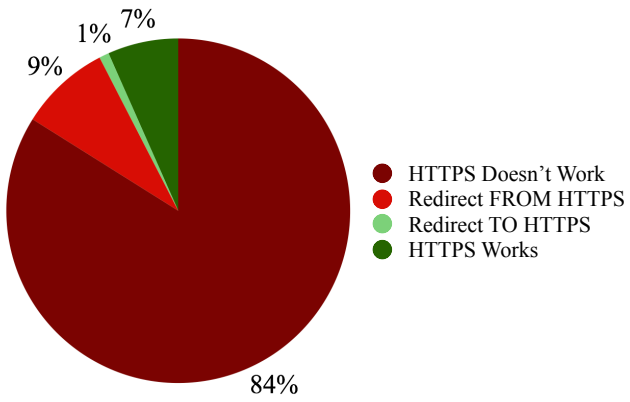

Figure 3: HTTPS browsing support distribution for Alexa-categorized websites in March 2015

of their customers. This has the effect of leaving customer purchase interests available to collection by mass surveillance infrastructure.

When looking at HTTPS support more generally (and excluding those that redirect away from a secure connection), we find that Science, Reference, and Computers categories have the highest amount of support, at 13%, 14%, and 15%, respectively. The average level of HTTPS browsing support for all categorized sites is 7.5%. News websites have the lowest amount of support, at 3%. We find that Advertisers exhibit a slightly higher than average percentage of HTTPS support, at 11%. Adult sites displayed average results, with 8% support for HTTPS browsing.

3.2.2 Support for HTTPS browsing by Alexa rank

Finally, we analyzed whether a website's Alexa rank is associated with their support for HTTPS browsing. To be more comparable to our category-based findings, for this analysis we only included websites which had a defined category in our data set (28,418). While we found that there was an inverse relationship between Alexa rank and TLS handshake response rates (with a slope of -1.44% per 10k ranked sites), we found no relationship[2] between a site's ranking and its support of end user basic browsing activities over HTTPS. As mentioned in the preceding section, we found that sites allow users to browse via HTTPS 7.5% of the time. Across Alexa ranks the distance from this mean never exceeds 1% in either direction, illustrated in Figure 4. We see similar results for HTTPS by default support levels as well as enforced redirects away from HTTPS.

These results illustrate a contrast between TLS handshake responses and support for basic browsing activities. While TLS responses are more prevalent the higher the Alexa rank, support for secure browsing is relatively consistent across rankings. The precise reasons for this warrant further investigation.

4. LIMITATIONS

[2]We measured a slope for HTTPS browsing support rates at -0.07% per 10k domain grouping, or 20x smaller than the growth rate for TLS handshake responses.

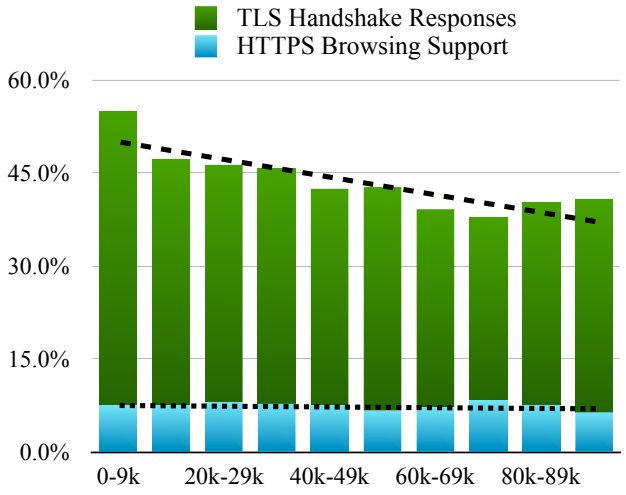

Figure 4: TLS Handshake responses and HTTPS browsing support by Alexa ranking grouping

A limitation of this study is that the list of the top 100,000 Alexa sites we used was retrieved in February 2015. We used this list to categorize the scan results from the University of Michigan study. We did not account for the fact that the scan data originates from mid 2012 to early 2014. Because of this, it is possible that several of the websites in the 2015 Alexa list did not exist throughout the whole scanning period. This limitation is mitigated by the fact that our main findings from these data are about the relative similarity and differences of categories within this data. Noise introduced by the Alexa categorization is likely to be evenly distributed across industry categories.

Another limitation of this study is the lack of a consistent number of websites across different Alexa categories. Our Alexa data collection method looked at the top 100,000 websites, not the top n websites in each category. The latter method would yield more comparable results across categories.

In general, the study's findings are inconclusive given the lack of statistical tests for significance and robustness. We aim to address this shortcoming in future iterations of this project.

5. DISCUSSION

While the rate at which websites across categories increase basic TLS support appears to be gradually increasing, our findings show that the overall rate by which websites adopt TLS changed little before and after June 2013, when the first Snowden disclosures were revealed. When looking at individual categories, we noticed variance in TLS support growth rates, but having done no formal statistical verification, cannot reliably claim whether or not these variations are significant.

We furthermore found that while a website's Alexa ranking may be associated with its support for basic TLS operations, the ranking did not appear to have a relationship with a site's support for an end user actually browsing a site on HTTPS. What does appear to play a role in a site's support for HTTPS browsing (and not just payment or sign-in operations) is their industry category. Science, Reference (higher education), and Computer websites have the best support for HTTPS browsing.

The fact that an Alexa-categorized website is 9.4 times more likely to redirect a user away from a secure connection than to redirect them to one might be a disheartening statistic for privacy advocates. However, our findings provide useful empirical support for future advocacy efforts aimed at increasing the adoption of security best practices in content-driven websites. Ultimately, however, our results show that content-driven websites may have responded very slightly to the Snowden revelations, and that there is much more to be done in securing readers on the web.

5.1 Future work

In addition to the areas of future work identified above, we intend to continue this line of research by investigating whether the number of ad trackers and degree to which they support HTTPS are correlated with a website's adoption of HTTPS.

This study's results may also be complemented by a longitudinal analysis of the HTTPS support of websites grouped by their country code top-level domains (CCTLD). We intended to include this analysis in our study but determined that we could not perform a reliable analysis of relative adoption amongst CCTLDs given we lack a complete list of registered domains. A data set containing the total number of registered domains by CCTLD over time is needed in order to correlate countries' websites with HTTPS IPv4 scan results.

6. CONCLUSIONS

This study investigated the relationship between a website's Alexa ranking and categorization and its response to TLS handshakes over time. We found that several categories exhibited notable increases in their TLS implementation rates following the Snowden disclosures, though future work is needed to assess whether or not they are statistically significant. We have furthermore found that websites are far more likely to redirect a user away from a secured HTTPS browsing session than to redirect a user to a secured session. Given these findings it is apparent that there is room for industry and advocates to improve the web security ecosystem in order to secure web readers from bulk surveillance of their browsing activities.

7. ACKNOWLEDGMENTS

Thanks to Dr. Christopher Parsons, Masashi-Crete Nishihata, Jeffrey Knockel and the anonymous reviewers for their feedback on this paper.

8. REFERENCES

[1] A. Acquisti, L. Brandimarte, and G. Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015.

[2] Alexa. How are Alexa's traffic rankings determined? `https://support.alexa.com/hc/en-us/articles/200449744-How-are-Alexa-s-traffic-rankings-determined-`.

[3] A. Arnbak, H. Asghari, M. Van Eeten, and N. Van Eijk. Security collapse in the HTTPS market. *Communications of the ACM*, 57(10):47–55, 2014.

[4] M. Bellare, K. G. Paterson, and P. Rogaway. Security of symmetric encryption against mass surveillance. In *Advances in Cryptology–CRYPTO 2014*, pages 1–19. Springer, 2014.

[5] S. M. Bellovin. By any means possible: How intelligence agencies have gotten their data. *Security & Privacy, IEEE*, 12(4):80–84, 2014.

[6] CIGI and Ipsos. Global survey on internet security and trust. `https://www.cigionline.org/internet-survey`, 2014.

[7] Disconnect. List of ad trackers and their hostnames. `https://services.disconnect.me/disconnect.json`.

[8] Z. Durumeric, J. Kasten, D. Adrian, J. A. Halderman, M. Bailey, F. Li, N. Weaver, J. Amann, J. Beekman, M. Payer, et al. The matter of Heartbleed. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, pages 475–488. ACM, 2014.

[9] Z. Durumeric, J. Kasten, M. Bailey, and J. A. Halderman. Analysis of the HTTPS certificate ecosystem. In K. Papagiannaki, P. K. Gummadi, and C. Partridge, editors, *Internet Measurement Conference, IMC'13, Barcelona, Spain, October 23-25, 2013*, pages 291–304. ACM, 2013.

[10] P. Eckersley et al. SSL observatory. *Invited Talk, Usenix Security*, 2011.

[11] S. Englehardt, D. Reisman, C. Eubank, P. Zimmerman, J. Mayer, A. Narayanan, and E. W. Felten. Cookies That Give You Away: The Surveillance Implications of Web Tracking. In *World Wide Web Conference 2015*, May 2015.

[12] B. Gellman. U.S. surveillance architecture includes collection of revealing Internet, phone metadata. *Washington Post*, June 13, 2013.

[13] R. G. Glenn Greenwald, Ryan Grim. Top-secret document reveals NSA spied on porn habits as part of plan to discredit 'radicalizers'. *Huffington Post*, November 26, 2013.

[14] E. Kao. Making search more secure. `http://googleblog.blogspot.ca/2011/10/making-search-more-secure.html`.

[15] N. Lidzborski. Staying at the forefront of email security and reliability. `http://gmailblog.blogspot.co.uk/2014/03/staying-at-forefront-of-email-security.html`.

[16] M. Marquis-Boire. Schrodinger's cat video and the death of clear-text. `https://citizenlab.org/2014/08/cat-video-and-the-death-of-clear-text/`, 2014.

[17] D. Naylor, A. Finamore, I. Leontiadis, Y. Grunenberger, M. Mellia, M. Munafò, K. Papagiannaki, and P. Steenkiste. The cost of the S in HTTPS. In *Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies*, pages 133–140. ACM, 2014.

[18] Office of the Privacy Commissioner of Canada. Metadata and privacy: A technical and legal overview. `https://www.priv.gc.ca/information/research-recherche/2014/md_201410_e.asp`, October 2014.

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

[20] S. Renfro. Secure browsing by default. `https://www.facebook.com/notes/facebook-engineering/secure-browsing-by-default/10151590414803920`.

[21] T. Romm. NSA surveillance creeps onto tech's lobbying agenda. `http://www.politico.com/story/2013/10/nsa-surveillance-creeps-onto-techs-lobbying-agenda-98706.html`, 2013.

[22] S. Schoen. Let's encrypt. In *31st Chaos Communcations Congress*, 2014.

[23] C. Soghoian and S. Stamm. Certified lies: Detecting and defeating government interception attacks against ssl (short paper). In *Financial Cryptography and Data Security*, pages 250–259. Springer, 2012.

[24] Twitter. Securing your Twitter experience with HTTPS. `https://blog.twitter.com/2012/securing-your-twitter-experience-with-https`.

[25] T. Ullyot. Facebook releases data, including all national security requests. `http://newsroom.fb.com/news/2013/06/facebook-releases-data-including-all-national-security-requests/`, June 2013.

[26] D. Yardon and D. Paletta. Cybersecurity summit exposes silicon valley's privacy fears. *Wall Street Journal*, February 13, 2015.