Variations in Tracking in Relation to Geographic Location

Nathaniel Fruchter, Hsin Miao, Scott Stevenson

{*nhf,hsinm,sbsteven*}*@andrew.cmu.edu*

***Abstract*—Different countries have different privacy regulatory models. These models impact how countries handle privacy both legally and culturally, specifically in the realms of online tracking and advertising policy. In this paper, we investigate the amount of tracking present on top websites in various countries around the world that utilize these different regulatory models. We found that there are significant differences in tracking activity between different countries using several metrics. We also suggest various ways to extend this study which may yield a more complete representation of tracking from a global perspective.**

I. Introduction

PRIVACY laws have been enacted worldwide with the purpose of protecting internet users' private information. Privacy laws can be divided into four main models [1] that differ in scope, enforcement, and adjudication: the comprehensive model, the sectoral model, the co-regulatory model, and mixed/no-policy model. These models impact how countries handle privacy both legally and culturally, specifically in the realms of online tracking and privacy legislation.

Tracking is implemented in a variety of ways, some of the most popular being third-party cookies and JavaScript tracking code. We discovered that websites utilize a diverse plethora of trackers for various purposes. However, there is currently a lack of information as to how trackers differ between countries that employ these different models. The purpose of this project is to find the relationship between the amount of tracking and various countries that employ different privacy regulatory models.

In this project, we compared the amount of trackers on websites that operate in various countries with different privacy models. We have chosen Germany to represent the comprehensive model, the United States and Japan to represent the sectoral model, and Australia to represent the co-regulatory model. The sites that we are interested in are Alexa Top 500 sites [2] that have domains in multiple countries. We utilized Amazon Web Services to visit and crawl the data from the websites by servers in those countries.

We locate and identify these trackers using 3rd party HTTP requests and cookies. In addition, we identify ads from the websites by using a list provided by AdBlock browser extension [3]. Automation of the process are handled using the OpenWPM [4] tool which allows for synchronization across browsers and virtual machines ensuring that requests will occur at the same time.

In the following sections, we will first review some related literatures. Detail description of our method and experimental results are stated in Section III and IV. Discussion and future works are described in Section V and VI.

II. Related Work

Privacy in the news seems inescapable; a general concern regarding the intrusiveness and pervasiveness of online tracking, advertising, monitoring has caught the public attention. For example, concerns over the activities of social networking sites and advertisers such as Facebook [5] bring up issues of anonymity and tracking in daily life. Similarly, the level of privacy protection put into place by industry giants such as Google has come under scrutiny [6] as jurisdictions with more comprehensive privacy regulations have called the effectiveness of their protections into question.

These worries also demonstrate the large amount of change that the Internet has undergone in a relatively short amount of time. As Mayer and Mitchell note [7], individual instances of web content have evolved from a single-origin affair into a conglomeration of "myriad unrelated 'third-party' websites," each facilitating anything from advertising to social media. This has been demonstrated by Krishnamurthy and Wills [8] in their longitudinal study, demonstrating what they term an ?increasing aggregation of user-related data by a steadily decreasing number of entities.? Furthermore, this explosion of third parties has existed an environment with little to no regulation until very recently [7], with advances only occurring in the comprehensive regulatory environment provided by the European Union.

A. Privacy-related Web Measurement

This tangled web of privacy, regulation, and jurisdiction raises many concerns. Coupled with the increased salience of online privacy concerns, this has led to an explosion of privacy-related web measurement studies in recent years. For example, Engelhardt et al. [9] have identified 32 studies that they categorize as "web privacy measurement studies." This category of study has great breadth, ranging from technical analyses of information leaked by web scripting languages [10] to empirical analyses of search engine personalization [11]. In this vein, numerous comparison-style studies have also been run, touching on diverse subjects such as discrimination in online advertising [12] and the effectiveness of online privacy tools [13].

While the above studies make valuable contributions by taking on tasks like revealing the sources of potential privacy harms, detailing the effects of these third party entities, and taking a user-centric to studying and enhancing privacy, they generally do not explore the impact of industry and country-level policy on the overall incidence of these third parties. Connolly [14] comes the closest, performing an evaluation

of various websites' compliance with the European Union's "Safe Harbor" privacy policy. Finding an astoundingly small subset of companies in compliance with Safe Harbor directives, Connolly discusses the "significant" privacy risk to consumers resulting from noncompliance. Issues like these raise the necessity for a more comprehensive measurement of jurisdictional differences in tracking and advertising activity (see Background and Motivation).

B. Web Measurement Methodology

Web measurement studies are considered challenging for two reasons: causality and automation [9]. These difficulties make researchers design experiments that lead to inconsistency and reinvention. In order to solve the problem, Englehardt et al. conducted a study that reviewed general experimental frameworks and performed methodological analyses of extant web measurement studies. With this framework in mind, the authors developed a platform, OpenWPM [4], that addressed many of the issues of flexibility and scalability surrounding past web measurement studies. The current study will be performed using this platform, as it builds upon proven frameworks such as Selenium [15] and has been validated in several studies [4][9]. We hope to build upon this framework and avoid further problems, especially those surrounding replication of effort and methodological inconsistencies.

III. Method

To source an internet connection point at various locations, we used Amazon Web Services, or AWS. AWS provides cloud-based virtual machines that can be configured in numerous ways. AWS employs a 'pay-for-what-you-use' model, so it is economically convenient for us to use. We installed OpenWPM on these machines and ran our tests from the cloud without having to rely on a proxy to set our location. AWS offers virtual machines in any of the following places: Virginia (US), Ireland (EU), Frankfurt (EU), Oregon (US), California (US), Singapore (Asia), Sydney (AUS), Sao Paolo (South America), and Tokyo (JP) [16]. This covers almost all of the regions we would like to examine – the only regions not represented are Russia and China which are currently not options when using AWS EC2.

A. OpenWPM

Our next step was to collect data on a number of metrics related to tracking, including the number of cookies and HTTP requests. Engelhardt et al.?s OpenWPM platform is a purpose-built web measurement platform that logs a large amount of web session data in a standardized SQLite database format, making this study a perfect environment in which to use the platform. We utilized the most recent publicly available version of OpenWPM for the data collection portion of our study and used the platform?s API to programmatically crawl a list of the top 250 websites as defined by Alexa [2]. OpenWPM?s Firefox backend was used for the crawl with both JavaScript and Flash enabled.

B. Third-party HTTP requests and cookies

Two of our variables of interest are located within different SQLite tables generated by OpenWPM with each crawl: cookies and http_requests. We extracted domains of cookies and URLs of HTTP requests from these two tables by using sqlite3 library in Python. In order to further analyze third-party cookies and HTTP requests, we set a rule to determine whether the URL in a record is related to the website where the record is extracted. To be more specific, if the URL in a record does not contain the domain name of the extracted website, it is a third-party cookie or HTTP request. For example, if a cookie is extracted from amazon.de with the URL fls-eu.amazon.de, it is a first-party cookie because of identical base domain. In contrast, if a cookie also extracted from amazon.de with the URL zanox.com, it is a third-party cookie because of differing domains. By implementing these procedures, we can use statistical tools to analyze the collected data.

C. Heuristic: AdBlock "easylists"

AdBlock Plus [3] is a popular browser extension available for both Firefox and Chrome which allows users to filter and block elements on a webpage according to user-specified rules. As evidenced by the extension name, this capability is most often used in service of blocking advertisements, tracking code, or other content deemed annoying or objectionable. Due to its open source nature and large, international user base, AdBlock Plus provides us with a unique resource: a massive, crowd-sourced list of rules that allows us to detect the presence of advertising or tracking assets within a list of URLs and page elements. These rules are compiled in two "easylists" [17] provided on the AdBlock website, with one focused on ad-blocking rules and the other focused on tracker-blocking rules.

Using a similar approach to the one detailed in the last section, we extracted the full URLs of HTTP requests and responses from the OpenWPM crawl database using Python and the sqlite3 library. We then used the adblockparser [18] Python module to match the extracted HTTP request and response URLs against the two sets of AdBlock rules mentioned above. The number of positive ad or tracker hits (positive pattern matches) were aggregated by domain, country, and rule set in order to produce summary statistics for use in further analysis.

IV. Results

A. Evaluation Metric: Third-Party Cookies and Requests

The goal of our project is to discover the variation in trackers between different countries. In our experimental design, the independent variable is country. It is a categorical variable with 4 levels if we compare the number of trackers in different countries. If we compare the trackers in different regulation models, the level of the variable is 3 because Japan and the United States both belong to sectoral model.

There are some dependent variables for further analyses. First, we analyzed the number of third-party cookies and HTTP requests, which is closely related to online trackers.

Second, because the number of third-party cookies and HTTP requests are dependent to the number of first-party ones, we looked at the proportion of third-party and first party cookies and HTTP requests to see whether the ratios are identical in different countries. Moreover, the number of first-party cookies or HTTP requests are analyzed because some sites (e.g., Google) are both an analytics provider and a service provider, they may use other methods besides third-party cookies to track the information of users.

Due to the categorical independent variable and quantitative dependent variables, the one-way ANOVA test is suitable for the analyses. There are some assumptions for the test which include normally distributed data, equal variances, and independent error. We ensure that the independent error assumption is correct because we collected the data by using servers in different countries. However, the other two assumptions could be violated because we collected HTTP requests and tracking cookies from websites in different countries. Therefore, we used Kruskal-Wallis test for the analyses.

B. Third-party HTTP requests

We compared the number of third-party HTTP domain requests among different countries. Table ? shows the average rank for each country in Kruskal-Wallis test. We found that the difference of the numbers of third-party domain of HTTP requests among our four countries are significant (?2 = 43.863; df = 3; p ¡ 0.0005). We also found that there are more third-party HTTP requests in the US compared to Germany and Australia (?2 =10.752, df=1, p=0.001). The differences between Germany and Australia were not significant. Moreover, there were more third-party HTTP requests in Germany and Australia compared to Japan (?2 =39.709, df=1, p¡0.0005).

C. Cookies

We also compared the number of third-party and first party cookies among different countries. Table ? shows the average rank of number of third-party cookies for each country in Kruskal-Wallis test. The results show that although the difference of numbers of first-party cookies is not significant, the difference of number of third-party cookies is significant (?2=13.147; df=2; p=0.004). We found similar results compared to the number of domains in third-party HTTP requests. There are more third-party cookies in the US compared to Germany (?2 = 4.111; df = 1; p = 0.043) and Australia. Also, the difference between Germany, Australia, and Japan is not significant.

D. Correlation between HTTP requests and cookies

Table I shows the correlation between the number of third-party domain for HTTP requests and third-party cookies. We found that in these countries these two variables are strongly correlated, providing an indicator for the validity of the measure.

TABLE I
CORRELATION BETWEEN HTTP REQUESTS AND COOKIES

Country	r
AU	0.691
DE	0.634
JP	0.778
US	0.715

TABLE II
SUMMARY STATISTICS FOR ALL TRACKING-RELATED HTTP REQUESTS

N	Mean Requests (SD)	Mean Hits (SD)	Mean Proportion Hits (SD)
1931	111 (116)	6.54 (7.7)	0.06 (0.05)

E. Metric: AdBlock rules

1) Origin-dependent tracking activity : One crucial phenomenon to test for is the presence of origin-dependent tracking activity, something that we will term geographic tracker churn for short. The essence of the question is simple: if user A visits example.com from country A and user B also visits example.com at the same time, but from country B, will they receive the same type and number of trackers? Evaluating the presence of this churn for several reasons. First, the presence or absence of the churn will help us determine how heavily geographic factors need to be controlled for in this (and other) studies. Second, the presence of churn could indicate interesting, adaptive behavior by tracking companies that could warrant further investigation.

To this end, we crawled Alexa?s list of the top 500 global sites from all four of our server locations at identical times and compared matches against Adblock?s tracking EasyList. Controlling for outliers (mean +/- 3 s.d.), nonparametric tests of both the absolute number of hits by country and the proportions of hits by country show no significant difference (n hits: ?2=0.805; df=3; p¿0.84, proportion: ?2=0.172; df=3; p¿0.98). Because of this, we can conclude that geographic tracker churn will not be a significant factor for us within the scope of our experiment.

2) More trackers than ads: However, there were significant differences in type of hit (trackers vs. advertisements) within the same top 500 sites. The proportion of requests associated with trackers was significantly higher than the proportion associated with advertisements (?2=45.1; p¡0.0001). A pairwise comparison across the top 500 sites showed that trackers accounted for approximately 2% more requests than advertisements (95% CI [0.015, 0.021]). This is significant considering the overall proportion of requests for both ads and trackers is 5.4% (SEMean = 0.0009, 95% CI [0.052, 0.056]).

TABLE III
SUMMARY STATISTICS BY COUNTRY FOR TRACKING-RELATED HTTP REQUESTS

Country	N Rows	Mean(Number_Requests)	Mean(Number_Hits)	Mean(Proport
AU	494	99.19	6.83	0.06
DE	492	121.04	5.70	0.05
JP	451	103.15	4.10	0.05
US	494	120.59	9.34	0.08

TABLE IV
PAIRWISE COMPARISONS BETWEEN COUNTRIES FOR TRACKING HITS

Country A	Country B	Z	p	95% CI For Change
US	JP	10.42	<.0001	[0.028, 0.040]
US	DE	7.77	<.0001	[0.018, 0.031]
US	AU	2.57	<.02	[0.001, 0.014]
JP	DE	-3.64	<.0005	[-0.013, -0.002]
DE	AU	-5.29	<.0001	[-0.021, -0.009]
AU	AU	-8.33	<.0001	[-0.031, -0.019]

3) Differences by country: While we couldn?t draw conclusions about the models themselves, we do find interesting results when examining each individual country in a series of pairwise comparisons between the top 250 sites in each country. Differences in the proportion of total HTTP requests associated with trackers differs significantly and may imply the presence of significant variation beyond what can be explained on the country or model level.

4) More tracking-related requests in the United States: A pairwise examination of the proportion of HTTP requests related to tracking activity (operationalized as the proportion of requests that matched an Adblock rule) show that United States has significantly more tracking activity compared to all of our other countries. While the differences varied by country, each comparison showed a significantly greater (at least p¡0.02) percentage of tracking requests, ranging from less than 1% (US-AU) to more than 3% (US-JP). Table ? displays these pairwise tests, along with confidence intervals, in more detail.

5) Differences within the sectoral model: It is especially interesting to note the comparisons between our two sectoral model countries, the United States and Japan. Even though they ostensibly have the same regulatory model, the United States showed a significantly greater (all p¡0.02) amount of tracking-related HTTP requests?anywhere from 2.8% to 4% more. Considering the average number of requests per page is over 100 (see table ?), even a 4% increase in tracking-related requests could indicate the loading of 4 to 5 more tracking elements or scripts.

6) Does origin matter?: When this current data set, based on the top 250 sites from each country, is compared to the top 500 global sites data set used earlier, an interesting possibility presents itself. Looking at the series of pairwise comparisons for the top 500 sites (see Table ?), none of the differences between countries are significant (all p¿0.71). This indicates that it may be the website?s country origin, not the user?s, that matters in terms of tracking activity present. However, there may be other factors that account for this difference in variation, something that will be expanded on in our discussion below.

V. DISCUSSION

Software issues In the infancy of this study we attempted using Tor as a proxy to set our location. Tor allows users to specify exit nodes using 2-letter country codes and restricts exiting traffic to nodes in that country. This is done using a simple command-line flag when starting Tor. Tor utilizes the SOCKS5 (Socket Secure) protocol when it is operating as a proxy [19]. The OpenWPM tool we are using to collect our results utilizes and relies upon a Python library called MITMProxy. This library allows OpenWPM to collect all requests and cookies at the proxy and store the results, including those that would normally be encrypted (hence the man-in-the-middle nature of the proxy). Our proposed architecture would have required the proxy created by OpenWPM to communicate with Tor directly using SOCKS. Unfortunately, MITMProxy has known problems communicating with upstream SOCKS proxies and we could not obtain results using this combination of technologies [20] Due to this, we were forced to look for an alternative method.

Outliers We are also interested in outliers about tracking behaviors in the websites. In the US, nydailynews.com has the most number of third-party cookies in top 250 websites. There are 6,546 third-party cookies in that website. Other news websites including foxnews.com, sfgate.com, drudgereport.com and nypost.com all have more than 900 third-party cookies. We found that news websites also play important roles of third-party cookies in Japan and Australia. theaustralian.com.au has 1,819 third-party cookies on its website, which is the third most in top 250 websites. In addition, in Japan, reuters.com has 1,827 third-party cookies, which is the most in top 250 websites. In contrast, websites with the most numbers of third-party cookies are not news websites. The finding is interesting because it implies that news websites rely on third-party cookies heavily in the US, Japan, and Australia. However in Germany, the tracking behaviors are not similar to other three countries because most of third-party cookies are in shopping websites instead of news websites.

Other factors Currently, we do not have enough data to conclusively say whether the different privacy regulatory models are actually statistically different from one another in practice. The reason for this is that our sample size isn?t large enough to provide a good representation. Something we did find evidence of though is that regulatory model may not indicate the level of technological privacy users get. We noticed that the US had many more tracking indicators than Japan overall yet they both follow the sectoral model. We are unsure of exactly why this is the case but we suspect that it may be due to cultural differences or perhaps the types of websites that are popular. It could be the case that the popular sites in Japan fall under a particular sector that is more regulated than those in the US. Another possibility is that tracking, advertising, and the sale of customer data is not the most popular business model for websites in Japan which could lead to the difference in tracking we saw.

Future work Although our study is fairly comprehensive in terms of what we are looking for, we are lacking in a few areas. The most prominent is that we lack a node in China and Russia, and therefore have no direct representation of the no privacy model regions. We initially thought that this would directly affect our ability to measure tracking properly but our results have shown that where we connect to a particular website from may have very little to do with tracking. This result is based on a small sample (just the US and Japan) so we would also like to verify this fact over a longer period of time and with more countries prior to making a concrete

conclusion. China may be an exception to this finding since they have the ?Great Firewall of China? in place which may distort our results. We are unsure of how much different the internet is inside and outside of China?s firewall.

Russia is another interesting case and doesn?t have the complication of a national firewall. Russia?s government has been taking an increasingly aggressive interest in the internet, recently going so far as commandeering the Vkontakte, the 'Facebook of Russia' [21]. Extending our study to incorporate these two countries seems very promising since it could yield results that are very different from what we have seen in our current study. In the case of China, AWS EC2 is currently in open beta (for Chinese residents only) for nodes in Beijing so setting up a node there will be possible in the near future. This would also give us the ability to measure and compare tracking in China from inside and outside the firewall.

It may be valuable to conduct this study again in the future as well. Doing so would allow comparison of tracking throughout time and it may be fruitful to link certain privacy-related events with changes in tracking. For instance, if Do Not Track becomes a widely-accepted standard (like the US government is pushing for), how different will the tracking landscape look? Would tracking increase or decrease for people not utilizing Do Not Track?

Another extension to our study would be to look deeper into other methods of tracking. Third-party cookies, third-party HTTP requests, and AdBlock rules don?t tell the whole story. For example, even though Google has very few third-party cookies or requests, they are probably tracking users more than other websites that have many third-party cookies or requests. In a similar vein, many major service providers like Google are also their own analytics providers. We do not account for this possibility in our study, but developing methods for doing so may reveal a more complete picture.

The final thing we would like to investigate more is why we see the results we do. What is causing the US to have much more tracking than Japan even though they are both sectoral countries? As of now we have some assumptions that it may be cultural or that the popular websites in Japan differ in category from those in the US or that the popular sites in Japan may fall under the regulation of a stricter sector. None of this can be confirmed by our data so additional research would have to be done to confirm or deny these assumptions.

VI. Conclusion

Going into this experiment, we assumed that there would be a significant difference in tracking between countries employing the different privacy models. We expected to see the most tracking in the no model and sectoral model countries, less in the co-regulatory model, and even less in the comprehensive model. We were also interested in determining if the country the website is based in, versus the country we are connecting from, plays a role in the amount of tracking. We were able to conclude that there were significant differences in tracking activity between different countries using several metrics. Due to our limited sample size, though, we were not able to draw strong conclusions regarding the models themselves. However, we were able to quantify many interesting variations in tracking behavior between countries and provide several directions for relevant future work to further investigate these variations.

References

[1] Peter P. Swire and Kenesa Ahmad, *Foundations of Information Privacy and Data Protection: A Survey of Global Concepts, Laws and Practices*. IAPP, 2012.

[2] Alexa, "The top 500 sites on the web. [online]. available: http://www.alexa.com/topsites."

[3] "Adblock Plus. [online]. available: https://adblockplus.org/en/about."

[4] OpenWPM, "OpenWPM. [online]. available: https://github.com/citp/OpenWPM."

[5] J. Marshall, "Facebook Extends Reach With New Advertising Platform. [online]. available: http://online.wsj.com/articles/facebook-extends-reach-withad-platform-1411428726."

[6] G. Sterling, "EU Seeking Numerous Google Privacy Disclosures, Policy Changes. [online]. available: http://marketingland.com/consent-google-analytics-one-many-privacy-changes-sought-europe-101495."

[7] Jonathan R. Mayer and John C. Mitchell, "Third-Party Web Tracking: Policy and Technology," *Security and Privacy (SP), 2012 IEEE Symposium on*, 2012.

[8] Krishnamurthy, Balachander and Wills, Craig, "Privacy Diffusion on the Web: A Longitudinal Perspective," *Proceedings of the 18th International Conference on World Wide Web*, 2009.

[9] Steven Englehardt, Christian Eubank, Peter Zimmerman, Dillon Reisman, Arvind Narayanan, "Web Privacy Measurement: Scientific principles, engineering platform, and new results," *Manuscript*, 2014.

[10] Jang, Dongseok and Jhala, Ranjit and Lerner, Sorin and Shacham, Hovav, "An empirical study of privacy-violating information flows in JavaScript web applications," *Proceedings of the 17th ACM Conference on Computer and Communications Security*, 2010.

[11] Hannak, Aniko and Sapiezynski, Piotr and Molavi Kakhki, Arash and Krishnamurthy, Balachander and Lazer, David and Mislove, Alan and Wilson, Christo, "Measuring Personalization of Web Search," *Proceedings of the 22nd International Conference on World Wide Web*, 2013.

[12] L. Sweeney, "Discrimination in Online Ad Delivery," *Queue*, 2013.

[13] Rebecca Balebako, Pedro G. Leon, Richard Shay, Blase Ur, Yang Wang, Lorrie Faith Cranor, "Measuring the effectiveness of privacy tools for limiting behavioral advertising," *In Web 2.0 Workshop on Security and Privacy*, 2012.

[14] Chris Connolly, "The US Safe Harbor - Fact or Fiction? [online]. available: http://www.galexia.com/public/research/articles/research_articles-pa08.html," *Galexia*, 2008.

[15] SeleniumHQ, "SeleniumHQ. [online]. available: http://www.seleniumhq.org/."

[16] "Regional Availability of AWS EC2. [online]. available: http://aws.amazon.com/about-aws/global-infrastructure/regional-product-services/."

[17] "Easylist. [online]. available: https://easylist.adblockplus.org."

[18] Scrapinghub, "Adblockparser python module. [online]. available: https://github.com/scrapinghub/adblockparser."

[19] C. Hansen, "How to use Tor as a Socks5 proxy. [online]. available: http://www.deepdotweb.com/2014/05/23/use-tor-socks5-proxy/."

[20] "Bug Report: Issues with MITMProxy and upstream SOCKS proxies. [online]. available: https://github.com/mitmproxy/mitmproxy/issues/211."

[21] Amar Toor, "How Putin's cronies seized control of Russia's Facebook. [online]. available: http://www.theverge.com/2014/1/31/5363990/how-putins-cronies-seized-control-over-russias-facebook-pavel-durov-vk," *The Verge*, 2014.