Variations in Tracking in Relation to Geographic Location

Nathaniel Fruchter, Hsin Miao, Scott Stevenson

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

I. Introduction

Privacy 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. Besides, 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 number of trackers and different countries.

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. 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. [8] 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 [9] to empirical analyses of search engine personalization [10]. In this vein, numerous comparison-style studies have also been run, touching on diverse subjects such as discrimination in online advertising [11] and the effectiveness of online privacy tools [12].

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 [13] 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 [8]. 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 [14] and has been validated in several studies [4][8]. We hope to build upon this framework and avoid further problems, especially those surrounding replication of effort and methodological inconsistencies.

III. METHOD

In our proposal, we discussed 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 [15]. 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 [16]. Due to this, we were forced to look for an alternative method.

The best alternative we found was Amazon Web Services (AWS) EC2. 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) [17]. 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

The next step is 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 25 websites as defined by Alexa [2]. OpenWPM?s ?headless? Firefox backend was used for the crawl with both JavaScript and Flash enabled.

B. Heuristic: First vs. Third Party Origins

Two of our variables of interest are located within different SQLite tables generated with each OpenWPM crawl: cookies and http_requests. We extracted the domains of cookies and the full URLs of HTTP requests from these two tables by using the sqlite3 library in Python. In order to further analyze first-party and third-party cookies and HTTP requests, we set a rule to determine whether the hyperlink in a record is related to the website where the record is extracted. To be more specific, if a hyperlink contains the domain name of the extracted website, it is a first-party cookie or HTTP request; otherwise, it is a third-party one. For example, if a cookie is extracted from amazon.de with the domain fls-eu.amazon.de, it is a first-party cookie because

of an identical base domain. In contrast, if a cookie also extracted from amazon.de with the domain zanox.com, it is a third-party cookie because of the 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" [18] 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 a modified version of the adblockparser 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 were aggregated by domain, country, and rule set in order to produce summary statistics for use in further analysis.

IV. Results

A. Evaluation Metric

The goal of our project is to discover the variation of trackers in 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, we use one-way ANOVA test for the analyses. There are some assumptions for the test: the distribution of the dependent variables follows a Normal distribution, variances of the dependent variables are identical, and the error of the dependent variables are independent. We ensure that the independent error assumption is correct because we collected the data by using servers in different countries. However, in order to obtain significant results, we should check whether other two assumptions are also valid when using ANOVA test to analyze the data.

B. HTTP requests

Table I, II, and III show mean and standard deviation of number of third party requests, ratio of third and first party requests, and number of first party requests in different countries. We found that the mean value of third-party and first party requests in Germany is more than other countries obviously. Although the mean values for HTTP requests in Germany are higher than other countries, the p-value for these variables are all larger than 0.05, which means the differences between different countries are not significant. Besides, we did not find the difference of the ratio of first and third party requests.

There are two possible reasons for the result. First, the standard deviation between different countries are not identical. So the assumption of ANOVA test is violated. Second, we have only 25 samples in each country, so the results are easily affected by some outliers. We determined the existence of outliers from the large standard deviation value in these tables. In order to correctly analyze the data, more data is required.

C. Cookies

Table IV, V, and VI show mean and standard deviation of number of third party cookies, ratio of third and first party cookies, and number of first party cookies in different countries. We found that the number of third party cookies in Germany and Australia are more than the

United States and Japan. However, the standard deviations are still large, so the difference is not significant.

D. AdBlock rules

Using our current sample of the top 25 websites from each country, a significant difference was found in the number of hits, both by country and by privacy regulation model. Aggregating across requests and responses, we found a significant difference ($p < 0.03$) between countries, with German sites reporting a slightly higher number of tracking/advertisement hits compared to other countries.

This same trend remained evident after re-running the analysis using a series of dummy variables to represent the presence of different regulatory models. Germany, our representative of the comprehensive model, had a significantly higher ($M = 7.89$) number of hits per site, as compared to our mixed ($M = 5.16$) or sectoral ($M = 5.07$) model countries.

While these results may seem surprising initially, the small sample of sites does not too unlikely that we are getting a skewed view of the landscape. If we are able to obtain data for a larger sample of sites (as mentioned below), significant changes may occur.

V. Discussion

Going into this experiment, we assume that there will be a significant difference in tracking between countries employing the different privacy models. We expect 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 are 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 have a slight expectation that it will, but we really don't know going into the experiment. Whether or not these assumptions have been confirmed is yet to be determined.

At the time of writing this draft, the results are incomplete. We have a script that allows us to collect the top n sites from a specified region utilizing the Alexa.com rankings. We have run tests on all regions except China and Russia for the top 25 websites in each country. After analyzing these results we did not notice anything significant (although there is some debate here) in the results. We are unsure of why this would be the case but we suspect it may have something to do with the small sample size. We will be keeping these results as a benchmark

of comparison however. The top 25 websites in each region may not be representative of the region as a whole. In response, we decided to collect the top 250 websites for each region. We are currently analyzing the results for this collection of websites but it is taking a bit longer as the set is much larger.

In addition to the above, we are also going to compare the results for the top 250 websites from each region with the top 500 websites globally. It will be interesting to see how the results compare for these two data sets.

The results of HTTP requests and cookies show that although there are some differences of the number of third party requests and cookies between these countries, they are not significant because their p-values are more than 0.05. In order to obtain further results, we will analyze those data after collecting the data from top 250 websites of each country. Because the number of samples increases, we expect that the equal variance assumption of ANOVA test will be satisfied. If the assumption is then satisfied, we can determine whether the tracking behavior of top sites in these countries are similar.

VI. 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. This directly affects our ability to measure tracking when connecting from these countries. We also cannot analyze China which may be an incredibly interesting case due to their "Great Firewall" and general views on privacy. Extending our study to incorporate China will be possible soon as AWS EC2 is currently testing Beijing as a regional offering.

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 ads 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.

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

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

[9] 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.

[10] 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.

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

[12] 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.

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

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

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

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

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

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

TABLE I

NUMBER OF THIRD-PARTY HTTP REQUESTS

3rd-party requests	Mean	Standard Deviation
United States	20.46	28.85
Japan	20.96	21.92
Germany	36.29	44.86
Australia	26.26	36.90

TABLE II

RATIO OF THIRD-PARTY AND FIRST-PARTY HTTP REQUESTS

ratio of requests	Mean	Standard Deviation
United States	1.806	3.331
Japan	1.426	2.618
Germany	1.654	3.513
Australia	1.231	2.236

TABLE III

NUMBER OF FIRST-PARTY HTTP REQUESTS

1st-party requests	Mean	Standard Deviation
United States	55.63	71.77
Japan	67.63	154.78
Germany	81.25	81.89
Australia	54.30	57.13

TABLE IV

NUMBER OF THIRD-PARTY HTTP COOKIES

3rd-party cookies	Mean	Standard Deviation
United States	22.82	44.165
Japan	30.70	53.726
Germany	79.26	128.493
Australia	88.966	234.177

TABLE V

RATIO OF THIRD-PARTY AND FIRST-PARTY HTTP COOKIES

ratio of cookies	Mean	Standard Deviation
United States	0.651	1.476
Japan	3.003	7.726
Germany	2.327	7.678
Australia	1.258	2.330

TABLE VI

NUMBER OF FIRST-PARTY HTTP COOKIES

1st-party requests	Mean	Standard Deviation
United States	106.14	157.091
Japan	135.00	434.499
Germany	209.78	378.571
Australia	85.35	79.622