Variations in Tracking in Relation to Geographic Location (Project Proposal)

Nathaniel Fruchter, Hsin Miao, Scott Stevenson

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

I. Project Description

In this project, we will be comparing the amount of trackers on websites that operate in various countries with different privacy models. We have chosen France and Germany to represent the comprehensive model, the United States and Japan to represent the sectoral model, Australia and New Zealand to represent the co-regulatory model, and finally Russia and China to represent the mixed/no-policy model. The sites that we are interested in are Alexa Top 500 sites that have domains in multiple countries. For example, we can compare Google.com (US version) with Google.com (German version with no country subdomain) with Google.com.de (German version with country subdomain) and determine if there is a difference in terms of trackers. We will also examine how connecting to these sites from different countries affects tracking. For example, if connecting to Google.com.de from the United States and Germany yields different results in terms of trackers.

We plan to locate and identify these trackers using 3rd party HTTP requests and cookies, examining other criteria if necessary. Automation of the process will be handled using the OpenWPM [1] tool which allows for synchronization across browsers and virtual machines ensuring that requests will occur at the same time. A Tor proxy [2] will be used to allow us to spoof the location we connect from.

II. Background and Motivation

Privacy laws have been enacted worldwide with the purpose of protecting internet users' private information. Privacy laws can be divided into four main models [3] 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.

III. 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 [4] 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 [5] 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 [6], 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 [6], with advances only occurring in the comprehensive regulatory environment provided by the European Union (see Background).

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. [7] 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 [8] to empirical analyses of search engine personalization [9]. In this vein, numerous comparison-style studies have also been run, touching on diverse subjects such as discrimination in online advertising [10] and the effectiveness of online privacy tools [11].

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 [12] 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 [7]. 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 [1], 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 [13] and has been validated in several studies [1][7]. We hope to build upon this framework and avoid further problems, especially those surrounding replication of effort and methodological inconsistencies.

IV. TIMELINE

1) 10/24 - 11/4:
    a) Set up the Tor network. (Scott)
    b) Set up the OpenWPM platform. (Hsin)
        i) Run test cases and fully test effectiveness of definitions.
    c) Obtain the website list from Alexa. (Nathaniel)
        i) Compile into sets of websites, broken down by country and analysis type.
    d) Look into Parallel Data Lab and/or set up appropriate hardware for virtual machines. (Scott)
2) 11/4 - 11/25:
    a) Write and test OpenWPM script for experiment automation.
    b) Run experiment.
    c) Data analysis.
        i) Extract data from OpenWPM databases.
        ii) Statistical analysis of data sets.
3) 11/25 - 12/4:
    a) Finalize the result.

b) Design the poster.

c) Write the final paper.

d) Extend literature review.

4) 12/4: Poster fair.

5) 12/12: Final paper.

The next step is to collect the number of cookies and HTTP requests in the database file. They are stored in two different tables: cookies and http_requests. We extracted the hyperlinks of cookies and HTTP requests from these two tables by using 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 extracted from www.amazon.de/ with the hyperlink fls-eu.amazon.de, it is a first-party cookie because the hyperlink contains the string ?amazon?. In contrast, if a cookie also extracted from amazon with the hyperlink zanox.com, it is a third-party cookie because it does not contain ?amazon?. By implementing these procedures, we can use statistical tools to analyze the collected data.

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.

REFERENCES

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

[2] Tor, "Tor Project. [online]. available: https://www.torproject.org/index.html.en."

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

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

[5] 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."

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

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

[8] 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.

[9] 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.

[10] L. Sweeney, "Discrimination in Online Ad Delivery. [online]. available: http://queue.acm.org/detail.cfm?id=2460278."

[11] 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.

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

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