Tracking the Trackers: A Large-Scale Analysis of Embedded Web Trackers

Sebastian Schelter
Technische Universität Berlin
sebastian.schelter@tu-berlin.de

Jérôme Kunegis
University of Koblenz–Landau
kunegis@uni-koblenz.de

Abstract

We perform a large-scale analysis of third-party trackers on the World Wide Web. We extract third-party embeddings from more than 3.5 billion web pages of the CommonCrawl 2012 corpus, and aggregate those to a dataset representing more than 41 million domains. With that, we study global online tracking on two levels: (1) On a global level, we give a precise figure for the extent of tracking, and analyse which trackers (and subsequently, which companies) are used by how many websites. (2) On a country-specific level, we analyse which trackers are used by websites in different countries, and identify the countries in which websites choose significantly different trackers than in the rest of the world. We find that trackers are widespread (as expected), and that very few trackers dominate the web (Google, Facebook and Twitter), except for a few countries such as China and Russia.

Introduction

The ability of a website to track which pages its visitors read has been present since the beginnings of the World Wide Web. With the advent of social media and *Web 2.0* however, another tracking mechanism has appeared: that of third-party websites embedded into the visited website by mechanisms such as JavaScript and images. Despite the widespread deployment of such tracking technologies, only qualitative small analyses about online tracking have been performed to date. We bridge this gap by performing a large-scale study of the distribution of third-party trackers on the web. The majority of websites contain third-party content, i.e., content from another domain that a visitor's browser loads and renders upon displaying the website. Such an embedding of third-party content has always been possible, but was relatively rare, since most embedded images were located on the same server as the page itself. In any case, the embedding of content was not intended for tracking. With the rise of social media and *Web 2.0*, websites increasingly began to embed links (in various forms) to third-party content, allowing the providers of such content to track users on a wide scale. The inclusion of third-party content occurs for a variety of reasons, e.g., advertising, conversion tracking, acceleration of content loading or provision of widgets.

Regardless of their primary purpose, third-party components can (and in many cases do) track web users across many sites and record their browsing behavior, and thereby constitute a privacy hazard. In order to understand and control this hazard, it is desirable to gain a deeper understanding of the 'online tracking sphere' as a whole. Previous research however has only studied small samples of this sphere due to the lack of comprehensive datasets. Recent developments allow us to study this online tracking sphere at a large scale: the availability of enormous web crawls comprised of hundreds of terabytes of web data, such as CommonCrawl[1].

Online Tracking Fundamentals

Technical foundations In its basic form, online tracking involves three types of actors: a *user* browsing the web, the *website* that she intentionally visits, and services called *third-parties*, which record her browsing to the website. Specifically, the user visits a website, whose HTML code typically contains references to external resources, such as style sheets, JavaScript code and images that are required to render the page in the client browser. These external resources allow online trackers to enter the communication, when they reside on servers controlled by the third-party. A typical example for such an external resource is a piece of JavaScript code, which the user's browser will automatically load from the third-party server, and execute. This external loading enables the third-party to record a wide variety of information about the user, such as the browser version, operating system, or approximate geolocation. This information can be used to compute a 'fingerprint' of the browser, that works suprisingly well at recognizing individual users (Eckersley 2010). Furthermore, the third-party has access to the URI of the page the user is visiting, the HTTP referrer, and, potentially, to previously set cookies (e.g., for a persistent login).

Privacy implications There is a considerable variety of third-parties, like advertisers, analytics services, social widgets, content delivery networks and image hosters, all of which have legitimate uses. However, the ability of many third-parties to record large portions of the browsing behavior of many users across a huge number of sites on the web

[1] https://commoncrawl.org/

poses a privacy risk, and is the subject of ongoing legal disputes (The Guardian 2015). The data recorded by this tracking infrastructure has been reported to contain large portions of the online news consumption (Trackography 2014), as well as intimate, health-related personal data (EFFHealth 2015). The ability to consume news and form a political opinion in an independent and unwatched manner, as well as the privacy of personal health-related data are vital for an open society, and should not be subject to commercially motivated data collection. Furthermore, recent frightening reports suggest that intelligence agencies piggyback on online tracking identifiers to build databases of the surfing behavior of millions of people (The Intercept 2015).

Data Acquisition

Collection methodology and limitations We represent both websites and third-parties by their pay-level domains (the pay-level domain is a sub-domain of a public top-level domain, for which users usually pay for[2]). We develop an extractor that takes a HTML document as input and retrieves all pay-level domains of third-party services that are embedded in the HTML code. In a first pass through the document, we investigate the `src` attribute of `script`, `iframe`, `link` and `image` tags. In order to also find third-parties that are dynamically embedded via JavaScript code, we parse all JavaScript and collect string variables that match a URI pattern. We run this extractor on CommonCrawl 2012, which has a datasize of approximately 210 terabytes in uncompressed form (Spiegler 2013). We choose CommonCrawl 2012 over latter corpora, as it has been created using a breadth-first search crawling strategy, which produces a much better representation of the underlying link structure of the web than the crawling of predefined lists used in latter corpora (Lehmberg, Meusel, and Bizer 2014). We run our extraction via Hadoop in the Amazon Elastic MapReduce service. We aggregate our data by pay-level domain, as we focus on high-level patterns of the tracker distribution. A limitation of our collection methodology is that our extractor cannot find transient trackers (trackers which are created from dynamically fetched external JavaScript code), as we only parse the JavaScript in an HTML page, but cannot execute it for performance reasons.

Dataset statistics and characteristics We extract the pay-level domains of third-parties contained in 3,536,611,510 web pages of the CommonCrawl corpus, and select all third-parties that occur in at least 0.01% of the domains contained in our corpus. We enrich the data for the resulting 1,375 third-parties obtained. We determine the registration countries and registering organizations for the third-party domains. Next, we manually check the websites of the domains to determine their owning companies, and label the third-parties according to their purpose and business model. We find 355 pay-level domains to belong to potentially privacy threatening tracking services, as their purpose or business model aims at recording the behavior of users visiting the websites into which these services are embedded. Thereby, we construct the *bipartite tracking network*[3], which represents the 36,982,655 embeddings of the 355 potentially privacy threatening third-parties in the 41,192,060 website pay-level domains of our corpus.

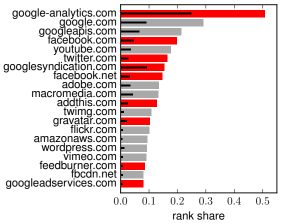

Figure 1: The twenty most common third-parties by rank share. Tracking third-parties are highlighted.

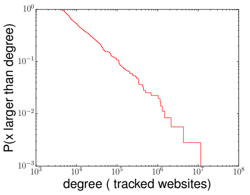

Figure 2: Cumulative distribution of the number of domains visible to tracking services.

Analysis of Tracking Services

Ranking tracking services We study the extracted tracking network to gain insights into the distribution of online tracking services on the web. We use Apache Flink (Alexandrov et al. 2014) for conducting the analysis. Our main question is to what proportion the browsing behavior of users on the web is visible to particular tracking services. Furthermore, we are interested in how strongly the tracking capabilities differ among various services. Unfortunately, we are not aware of any data source available to scientists that would allow us to quantify the number of visitors over time for our 41 million pay-level domains in 2012. We therefore employ PageRank (Page et al. 1999), a well-known measure of the relevance of websites, as proxy for ranking them by the traffic which they attract. We obtain the network of 623,056,313 hyperlinks between the pay-level domains in CommonCrawl[4], and compute its PageRank distribution to get an importance ranking for all the pay-level domains in our corpus. We derive a ranking measure for third-parties from the PageRank distribution p of the pay-level domains in the hyperlink network as follows. Let D denote a set of pay-level domains to inspect (e.g., all pay-level domains belonging to a certain top-level domain) and let $t(D)$ denote the subset of pay-level domains contained in D having third-party t embedded. We define the *rank share* $r_{D,t} = \sum_{j \in t(D)} p_j / \sum_{i \in D} p_i$ of a third-party t in domain set D as the sum of the PageRanks of domains from D that have t embedded, normalized by the overall sum of the pageranks of domains in D.

Predominant third-parties We start our analysis by computing the third-parties with the highest rank share in our corpus. Figure 1 shows the top twenty third-parties by rank share (and also illustrates the fraction of websites in which they are embedded as black bars). The by far most common third-party is `googleanalytics.com` with a rank share

[2] http://webdatacommons.org/structureddata/vocabulary-usage-analysis/

[3] available at https://ssc.io/trackingthetrackers

[4] http://webdatacommons.org/hyperlinkgraph/

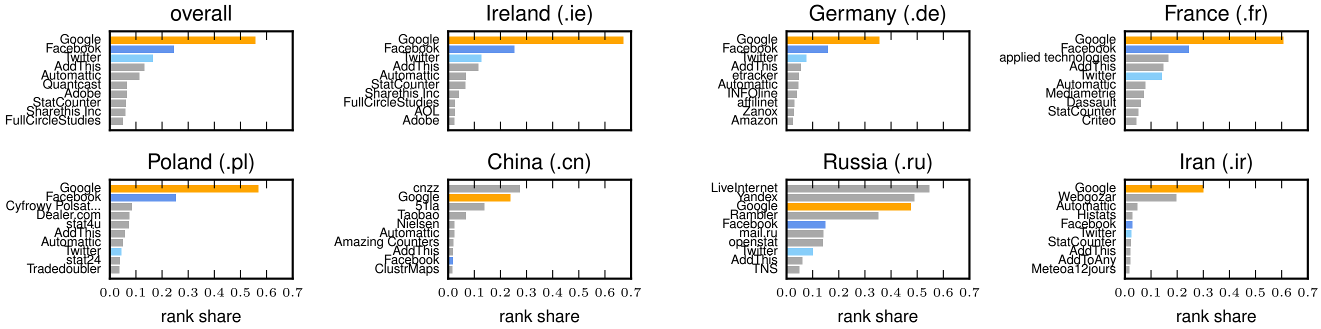

Figure 3: The ten companies per top-level domain with the highest rank share. Google, Facebook and Twitter are highlighted.

of 0.507 (which means that the pay-level domains embedding `googleanalytics.com` amount to more than half of the mass of the PageRank distribution in our web corpus). It is embedded on 24.8% of all pay-level domains in our corpus. We find that five out of the ten most dominant third-parties belong to Google. The next dominant family are social media related third-parties such as `facebook.com`, `twitter.com` and `addthis.com`. On the lower end, we find content delivery services, e.g., `twimg.com`, the image hosting platform of Twitter, the cloud platform Amazon Webservices `amazonaws.com` and Facebook's content delivery platform `fbcdn.net`. We highlight tracking third-parties in Figure 1, and find that the dominant third-party `googleanalytics.com` as well as eight additional out of the twenty predominant third-parties are known to intensively record user browsing behavior.

Differences in tracking capability Next, we study the differences in tracking capability between third-parties. Therefore, we investigate the distribution of the number of pay-level domains visible to an individual tracking service. This distribution corresponds to the degree distribution of the left vertex set (the trackers) in the bipartite tracking network. Figure 2 shows a cumulative plot of this distribution. We encounter a highly disproportionate distribution: 90% of the tracking services are embedded on less than ten thousand pay-level domains, while tracking services in the top 1% of the distribution are integrated into more than a million pay-level domains. Visually, this distribution appears to follow a power law, a well-known property of many networks which represent real-world phenomena. We fit a power-law distribution according to the method presented in (Clauset, Shalizi, and Newman 2009). We find that starting from degree 6,848, the distribution follows a power law with coefficient 1.725, which is very close to the coefficient observed in hyperlink networks[5].

Predominant tracking companies per country We compute the rank share of tracking third-parties in the subset of website domains belonging to a specific country (represented by its top-level domain). We aggregate the results on company level (in order to match a third-party to the country in which its owning company is based). Figure 3 shows the results of this analysis for a selection of countries: Ireland, Germany, France, Poland, China, Russia and Iran. We highlight the bars for the three globally most dominant companies, Google, Facebook and Twitter. These three companies have a special role, as we encounter them in the majority of top ten lists, in many cases accumulating the largest amount of rank share. Even among these three, Google has an outstanding position: we find it in a dominating role in the majority of countries, often with an amount of rank share that is more than double of what the second-placed company accumulates. We find that in many cases, the top ten companies consist of the three dominant US companies, Google, Facebook and Twitter, accompanied by a set of companies resident in the country under observation. Examples are Zanox (affiliate marketing) and INFOnline (digital audience measurement) in Germany, Criteo (advertising) in France, as well as Yandex and LiveInternet in Russia. These country-resident companies hardly ever appear in the top ten list of another country. The pattern of dominance of Google, followed by Facebook and Twitter is present in the overall corpus as well as in the vast majority of countries; however there are a few notable outliers, e.g., China and Russia where country-resident companies such as Yandex or CNZZ outrank Google.

Correlation analysis of the dominance of Google, Facebook and Twitter We further investigate the country-specific role of the three dominating companies Google, Facebook and Twitter. We therefore define a simple, dichotomous measure of dominance per country. We say that these three companies have a dominating role in a country if they accumulate more than half of the sum of the rank share of the top ten companies. We compute this measure for the 50 countries under investigation in our corpus, and encounter the dominance pattern for a vast majority of 46 countries. Only four countries do not exhibit this pattern: China, Russia, Iran and Ukraine.

We compute several country-specific indicator variables from additional datasets: Our political indicators consist of the *democracy index* and the *freedom of the press index* (The Economist Intelligence Unit 2012; Freedom House 2012). For the latter, we revert the scale to make high values indicate high freedom of the press. We use the *percentage of the population which speaks English* as socio-

[5] http://konect.uni-koblenz.de/statistics/power

cultural indicator (Wikipedia 2015). Finally, we derive several economic indicators. We compute the *online ad spending per capita* as the sum of digital and mobile ad spending per country normalized by its population. The *online ad spending ratio* is the ratio of the sum of digital and mobile and spending to the overall media ad spending. Lastly, the *US trade volume* denotes the sum of imports and exports of the US with the given country, normalized by the size of the population of the country (CatchaDigital 2013; U.S. Census Bureau 2015).

We calculate the point-biserial correlation coefficient ρ of the indicators to our dichotomous dominance variable. We find a very strong and at the same time statistically significant correlation with the political indicators, democracy index ($\rho = 0.662$, p-value < 0.001) and freedom of the press ($\rho = 0.612$, p-value < 0.001). The socio-cultural indicator, amount of English speakers, is only moderately correlated and only statistically significant at the 0.05 level ($\rho = 0.343$, p-value 0.028). The economic indicators, online ad spending per capita ($\rho = 0.333$, p-value 0.152), US trade volume ($\rho = 0.167$, p-value 0.24) and online ad spending ratio ($\rho = 0.062$, p-value 0.794) show low to moderate correlation, which is not statistically significant. These findings are surprising as they indicate that a positive characteristic such as freedom of the press is accompanied by a potentially very negative characteristic: the recording of people's browsing behavior by companies outside of the legal control of their countries institutions.

Related Work

The privacy hazards of online web tracking have been studied extensively (Krishnamurthy and Wills 2006; 2009). Several works have concentrated on specific actors such as social networks (Chaabane, Kaafar, and Boreli 2012; Krishnamurthy and Wills 2010) or intelligence agencies (Englehardt et al. 2015), as well as detecting trackers (Kalavri et al. 2016). In contrast to our work, these papers represent focused studies about several thousand prominent, English-language websites only.

Conclusion

The scope of our analysis allows us to make several novel observations about online tracking. We found that 9 out of the 20 predominant third-party domains belong to trackers, and confirm the extraordinary tracking capability of Google Analytics. Furthermore, we found that the distribution of the number of website domains tracked follows a power law. While there are many small trackers which are country-specific (e.g., to Germany, France, etc.), this is not true for the largest tracking services. These are Google, Facebook and Twitter, all US companies acting on a global scale, and representing the largest trackers in almost all countries. The exception to this pattern are a small number of countries such as China, Russia and Iran, which represent outliers in terms of political factors such as democracy and freedom of the press. Finally, we could not determine a statistically significant correlation of tracking with economic factors such as indicators related to ad spending.

Acknowledgements We thank Felix Neutatz, Hung Chang and Stephen Merity. This work has been supported by an AWS in Education Research Grant award, by the DFG (FOR 1306 Stratosphere), by the BMBF (BBDC 01IS14013A), and by the EC (FP7 grant 610928, REVEAL).

References

Alexandrov, A.; Bergmann, R.; Ewen, S.; Freytag, J.-C.; Hueske, F.; Heise, A.; Kao, O.; Leich, M.; Leser, U.; Markl, V.; et al. 2014. The stratosphere platform for big data analytics. *VLDB Journal* 23(6):939–964.

CatchaDigital. 2013. Worldwide ad spending forecast.

Chaabane, A.; Kaafar, M.; and Boreli, R. 2012. Big friend is watching you: Analyzing online social networks tracking capabilities. In *2012 ACM OSN Workshop*, 7–12.

Clauset, A.; Shalizi, C.; and Newman, M. 2009. Power-law distributions in empirical data. *SIAM* 51(4):661–703.

Eckersley, P. 2010. How unique is your web browser? In *Privacy Enhancing Technologies*, 1–18. Springer.

2015. Electronic Frontier Foundation - HealthCare.gov Sends Personal Data to Dozens of Tracking Websites.

Englehardt, S.; Reisman, D.; Eubank, C.; Zimmerman, P.; Mayer, J.; Narayanan, A.; and Felten, E. W. 2015. Cookies that give you away: The surveillance implications of web tracking. In *WWW*, 289–299.

Freedom House. 2012. Freedom of the press.

Kalavri, V.; Blackburn, J.; Varvello, M.; and Papaginannaki, K. 2016. Like a pack of wolves: Community structure of web trackers. In *PAMS*.

Krishnamurthy, B., and Wills, C. 2006. Generating a privacy footprint on the internet. In *ACM SIGCOMM*, 65–70.

Krishnamurthy, B., and Wills, C. 2009. Privacy diffusion on the web: a longitudinal perspective. In *WWW*, 541–550.

Krishnamurthy, B., and Wills, C. 2010. Privacy leakage in mobile online social networks. In *USENIX Conference on Online social networks*, 4–4.

Lehmberg, O.; Meusel, R.; and Bizer, C. 2014. Graph structure in the web: aggregated by pay-level domain. In *ACM Web Science*, 119–128.

Page, L.; Brin, S.; Motwani, R.; and Winograd, T. 1999. The pagerank citation ranking: bringing order to the web.

Spiegler, S. 2013. Statistics of the common crawl corpus 2012. Technical report, Technical report, SwiftKey.

The Economist Intelligence Unit. 2012. DemocracyIndex.

The Guardian. 2015. Belgian court orders Facebook to stop tracking non-members.

The Intercept. 2015. From radio to porn, British spies track web users' online identities.

Trackography. 2014. Meet the trackers; me and my shadow.

U.S. Census Bureau. 2015. U.S. trade in goods by country.

Wikipedia. 2015. List of countries by English-speaking population.