Beyond the Front Page: Measuring Third Party Dynamics in the Field

Tobias Urban
urban@internet-sicherheit.de
Institute for Internet Security
Ruhr University Bochum

Martin Degeling
martin.degeling@ruhr-uni-bochum.de
Ruhr University Bochum

Thorsten Holz
thorsten.holz@ruhr-uni-bochum.de
Ruhr University Bochum

Norbert Pohlmann
pohlmann@internet-sicherheit.de
Institute for Internet Security

ABSTRACT

In the modern Web, service providers often rely heavily on third parties to run their services. For example, they make use of ad networks to finance their services, externally hosted libraries to develop features quickly, and analytics providers to gain insights into visitor behavior.

For security and privacy, website owners need to be aware of the content they provide their users. However, in reality, they often do not know which third parties are embedded, for example, when these third parties request additional content as it is common in real-time ad auctions.

In this paper, we present a large-scale measurement study to analyze the magnitude of these new challenges. To better reflect the connectedness of third parties, we measured their relations in a model we call *third party trees*, which reflects an approximation of the loading dependencies of all third parties embedded into a given website. Using this concept, we show that including a single third party can lead to subsequent requests from up to eight additional services. Furthermore, our findings indicate that the third parties embedded on a page load are not always deterministic, as 50 % of the branches in the third party trees change between repeated visits. In addition, we found that 93 % of the analyzed websites embedded third parties that are located in regions that might not be in line with the current legal framework. Our study also replicates previous work that mostly focused on landing pages of websites. We show that this method is only able to measure a lower bound as subsites show a significant increase of privacy-invasive techniques. For example, our results show an increase of used cookies by about 36 % when crawling websites more deeply.

KEYWORDS

third parties, cookies, privacy, web measurement

ACM Reference Format:
Tobias Urban, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. 2020. *Beyond the Front Page:* Measuring Third Party Dynamics in the Field. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3366423.3380203

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380203

1 INTRODUCTION

A majority of today's online services are a combination of original content and—to a non-negligible extent—third party resources [45]. Most notably, online advertising is embedded using external resources that display ads to finance these services and to provide them to users free of charge. Other third parties are included for various means, e. g., libraries are used to develop services quickly, to decrease loading times, and for analytical purposes. Consequently, this leads to a highly dynamic Web with complicated dependencies among all participants. This trend comes with the drawback that some service providers might not be aware of which third parties are delivered to customers in their name when users interact with their website. Ultimately, third parties can pose risks to users, which is obviously unintended by the service provider. For example, third parties can create security problems (e. g., malvertising [28, 43, 44]), might have negative privacy implications (e. g., trackers [1, 9, 10]), or they can include content that might impact users in other negative ways (e. g., crypto miners [26, 41]). Services themselves reinforce these dynamics as they make use of different sets of third parties in different sections and webpages. For example, news websites often insert scripts to connect with social media below articles, but not on the actual landing page. This raises the question of whether This raises the question of whether previous studies that exclusively measured the landing pages (e. g., [6, 9, 21, 34, 45, 50]) captured a complete and comprehensive view of the analyzed phenomenon.

We perform a measurement study on 10,000 websites on the Web and analyze relations between third parties. We use the notion of *third party trees* (TPT) as a metric for loading dependencies of all third parties embedded into a given website. More specifically, a TPT contains information on all third parties (TP) observed when visiting a given website and accounts for the loading sequence of each TP. Consider the following example: *adidas.com* embeds a script which loads content from *Adobe* (3rd party). The script again loads a script from *Tealium* (4th party), which also loads a script from *Akamai* (5th party). As a result, a TPT captures the hierarchical structures of third parties on a given website and enables us to study the typical characteristics and dynamic nature of the modern Web. Furthermore, we show that embedding a single TP might result in embedding a non-deterministic amount of additional TPs,

which might pose privacy or security risks. Previous work in this area has analyzed implications of the presence of multiple third parties on websites. Recently, Ikram et al. [21] raised awareness for the problem of implicit trust created by decency chains in website embeddings. Earlier work focused on the extent of tracking (e. g., [6, 10, 45]), or on the used mechanisms (e. g., [1, 9, 29]), and again other works on defense mechanisms (e. g., [36, 39]), or the effectiveness of such (e. g., [14, 32, 34]). In this work, we want to asses in more detail by whom third parties are embedded into websites and study the extent of control service providers have on the embedded third parties. Most importantly, we show that previous studies did not measure the extent of the phenomenon extensively enough and only measured a (not necessarily generalizable) lower bound of included TP content. Our results show a significant increase in used cookies (36 %) and tracking techniques (6 %) on subsites.

In summary, we make the following key contributions:

(1) We introduce the concept of *third party trees* (TPTs) that reflects all third parties and dependencies when loading a website. Utilizing TPTs, we show that some TPs load several further partners and that those are not always deterministic and possibly in conflict with current legislation.
(2) We show that only measuring the traffic generated by landing pages of a website or only a few subsites leads to the risk of only capturing a (potentially limited) subset of the loaded third parties. This implies that the obtained results might be biased and not generalizable. For example, our study indicates that subsites use substantial more cookies (over 45 %) than the site's landing pages.
(3) Using our data, we try to replicate previous work to test if they only measured an incomplete view of their studied phenomenon and show that most privacy-invasive technologies occur more often on subsites.

2 BACKGROUND

Before introducing our approach, we briefly describe third party usage and outline the privacy implications of those.

2.1 Third Party Usage

Web services make use of resources hosted by third parties for various means. Everyday use cases for third-party usage are libraries used for web development, the integration of social media content (e. g., *Facebook* Like button), to display ads on websites, or to increase the service's performance (e. g., using cached fonts). Often these third parties are embedded by adding JavaScript code or an iframe element into the website. After injection, these objects perform the desired tasks independently and might even load further resources. For example, an embedded ad might load additional third-party code that is designed to counter ad fraud, to measure the effectiveness of the ad, or to load additional fonts used by the ad. As a result, embedding a single third party can lead to a long tail of additionally embedded partners.

2.2 Online Tracking

Tracking users online is a widespread phenomenon on the Web [9]. It is used to re-identify users navigating the Web and a crucial part of the modern online advertisement ecosystem as it allows them to provide targeted ads. Techniques to track users can be divided into *stateless* and *stateful* approaches. *Stateless* approaches use specific attributes of the users' device to identify it [1, 9, 13, 37, 53] (often called "device fingerprinting"). In contrast, *stateful* approaches use the machine's state to identify users. Typically an ID is assigned to each user and is stored in a cookie on the users' device. The upside of stateless approaches is that they cannot be prevented by deleting third-party cookies. However, they are more error-prone as device-specific attributes tend to change over time [16, 52].

3 RELATED WORK

Previous work analyzed tracking mechanisms and the effects of privacy legislation through measurement studies.

Privacy & Tracking Measurements. Englehardt et al. introduce *OpenWPM* and use it to crawl the top 1 million websites and analyze their tracking capabilities [9]. They find that many websites use highly sophisticated fingerprinting methods (e. g., based on image rendering) and that most companies participate in cookie syncing. Degeling et al. analyze different cookie banner notifications and effects of the GDPR on privacy policies [7]. They find that more than half of websites provide a cookie consent notice, but only very few offer users a real choice regarding cookie usage. The effects of the GDPR have been studied extensively in the past. For example, Utz et al. [51] analyzed implementations of cookie consent banners, Urban et al. [48, 49] analyzed usability of the GDPR right to access and the effect of the GDPR on cookie syncing activities [50]. Dabrowski et al. test if the GDPR has an impact on cookie settings when users access the same websites from different countries [6]. They find that websites (around 50 %) do not set cookies when a user from the EU visits the website while they set a cookie when the user visits from a non-EU country. Most recently, Sørensen et al. analyzed the effect of the GDPR regarding third parties embedded into websites [45]. The authors measure several prominent websites and test whether the GDPR affects their third party usage. They conclude that the overall usage of cookies declined but that the GDPR was not necessarily the driver for that change.

Third Party Inclusion. Closely related to our approach is the work of Kumar et al. [28] and Ikram et al. [21]. Both works use a concept of the implicit trust of the embedded third and further parties. Kumar et al. show that websites heavily rely on third parties, that almost one-third of websites embed a third party that loads further parties, and that these dependencies are a problem if one wants to serve a website fully via HTTPs. Ikramet al. also show that many websites (approx. 40 %) implicitly trust parties loaded by directly embedded third parties and see an increase in embedded malicious or at least suspicious site or script files in these chains.

Our work differs from previous work, as most tried to measure effects on a horizontal scale (i. e., visiting a lot of distinct domains) while we instead analyze websites on a vertical scale (i. e., , we visit several subsites of the same domain). Furthermore, we focus on privacy-invasive technologies and the determinism of third party dependencies. By this vertical approach and dependency identification, we can (1) analyze if subsites show different behavior compared to landing pages, (2) study effects of embedding different

third parties to websites, and (3) understand who is responsible for embedding specific third parties.

4 MEASUREMENT APPROACH

In this work, we conduct a large-scale measurement study of the dynamics of the Web on application level (i. e., the browser) to gain insights into the usage of third parties and to illuminate reasons for how they are embedded into websites. In this section, we describe our approach and highlight how we estimate the relations between specific third parties that are embedded into websites. Our study consists of a multi-stage process in which we (1) build a corpus of websites to visit, (2) use *OpenWPM* [9] to crawl these websites and gather first-party links on these websites, and finally (3) visit the crawled links and log all HTTP traffic, cookie usage, the embedded iframes, and JavaScript calls of interest.

4.1 Terminology

Before describing our approach, we define two terms we use throughout this work. By *TLD+1* we mean the last part of the hostname following the last dot in it. For example, the URL *https://tools.ietf.org* has TLD=org, hostname=tools.ietf, and TLD+1=ietf. In most cases, TLD+1 is a "second-level domain". However, some domain name registries use a second-level hierarchy. For example, New Zealand uses various second level domains for different purposes: .co.nz for organisations or .school.nz for schools. We identified the TLDs using Python's *tldextract* [40] package, which accurately splits generic or country code top-level domains (ccTLD). Furthermore, we distinguish between *landing page*s and *subsite*s. A website is a *subsite* (SB) of a *landing page* (LP) if both share the same TLD+1 but have distinct URLs. Hence, first-party links on landing pages, the page that is usually visited first, lead to subsites. We chose to use the term SB rather than "webpage" to explicitly highlight the hierarchical relation between SBs and LPs.

4.2 Website Corpus

In our analysis, we use the top 1M *Tranco* list et al. [30], which is an aggregation of four other domain top lists. We used the list generated on 03/26/2019 (ID: W9L9). First, we removed all websites with the same TLD+1 and only kept the one with the higher rank. We did so because we wanted to remove URLs of services that offer users the (almost) same functionality. For example if the list contains *google.com* (rank 1) and *google.co.uk* (rank 4) we would drop *google.co.uk* because both domains share the same TLD+1. In total, we removed 607 websites in this step. From the remaining domains, we used the top 10,000 domains and grouped them by the category of their content and also sort them into four different buckets based on their ranking.

We used the *McAfee SmartFilter Internet Database* service to retrieve a list of content categories for the websites [33]. We cluster the websites by categories because we want to check if the category of a website has an impact on the usage of cookies and other privacy-invasive technologies. Previous work has shown that, for example, *News* websites utilize more third parties (e. g., ad services) than other categories [45]. In total, 85 different categories are assigned to the websites of the dataset. An overview of the 15 most prominent categories is given in Figure 1. In the remainder, we limit the analyzed categories to the top eight categories and combine all remaining categories in "Other". Additionally, we group the websites by the following buckets based on the website's rank in the used list: (1) $1 \leq \text{rank} \leq 100$, (2) $100 < \text{rank} \leq 1,000$, (3) $1,000 < \text{rank} \leq 10,000$, and (4) $10,000 < \text{rank} \leq 100,000$. Due to the removal of duplicate domains, bucket (4) holds these 607 domains, 6.1 % of all visited domains. We use the buckets to test whether the popularity of websites has an impact on the usage of specific technologies.

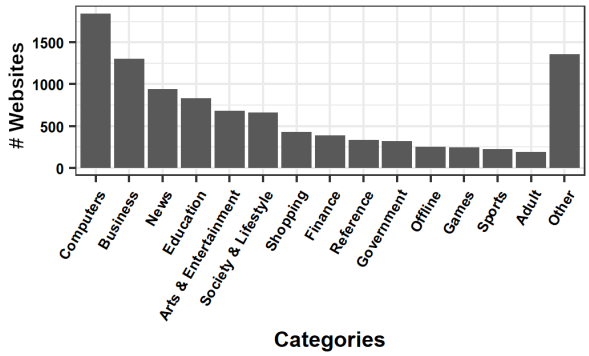

Figure 1: Overview of prevalent website topics in our dataset.

If not stated otherwise, we use the *one-way analysis of variance* (one-way ANOVA) statistical model to find differences between the analyzed groups. In all tests, we use a 95 % confidence interval.

4.3 Measurement Framework

To measure the dynamic of websites, we utilize the *OpenWPM* platform [9]. For each visit, we use the same user agent (Mozilla/5.0 (X11; Linux x86_64; rv:52.0) Gecko/20100101 Firefox/52.0) and desktop resolution (1366x768), allow all third party cookies, do not set the "Do Not Track" HTTP header or other privacy-preserving techniques (e. g., anti-tracking extensions), and use standard bot mitigation techniques to disguise our crawler (i. e., random scrolling and mouse jiggling). Furthermore, the browser adopts other properties from the operating system (Ubuntu 18.04). Aside from our bot mitigation techniques, we do not interact with the visited websites in any way, limitations of this approach are discussed in Section 6. While a website might detect our crawler, it is not detected by current mechanisms seen in the wild, as presented by Jonker et al. [24].

OpenWPM is configured to store all third party cookies set or accessed via JavaScript and HTTP headers. To capture these events, we instrumented specific JavaScript functions that access the local storage or HTTP cookies, by adjusting the .prototype of the respective functions and applying a wrapper to them that logs each call and access to these functions. Furthermore, we inspect all HTTP headers if a cookie is accessed (Cookie) or set (Set-Cookie). For our measurement study, we disabled Flash because, on the one hand, the technology will be deprecated by 2020 [2] and on the other hand, we did not find a considerable usage (< 0.01 %) of Flash cookies in a pre-study we conducted (see Section 4.3.1). We passively log all DNS responses to test if IP addresses are used, which are associated

with countries that do not automatically offer a GDPR adequate privacy protection level. We define all countries that are part of the Privacy Shield [47] and countries part of the European Economic Area (EEA) [19] to be adequate. We use *MaxMind*'s GeoIP database [31] to create this association.

4.3.1 Pre-Study. As the Web is highly dynamic, any attempt to measure it is quite challenging. To get a comprehensive view of cookie and third party usage, we conducted a pre-study to get an approximation of which measuring parameters to use (e. g., amount of subsites to visit) while limiting the crawling time and generated traffic to a reasonable amount. In the following, we limit our pre-study to TP cookies as prior work extensively analyzed those [6, 7, 10, 15, 17, 27, 42], and we want to test whether they might have missed cookies due to their measurement setup. However, in our primary analysis, we also analyze various tracking mechanisms (see Section 5.2). To find the optimal amount of subsites to visit, we randomly selected 100 websites (TLD+1) from the top 1,000 websites and visited 25, 50, 75, 100, 250, 500, and 1,000 subsites of these websites. The websites were visited in a separate measurement but using the same TLDs+1. We conducted these measurements using a browser with a profile that already has some cookies present in the local cookie store and once with a vanilla browser to see if active cookies influence cookie usage. We filled the local cookie store by randomly visiting 100 websites from the top 1,000 websites and used the resulting cookie store. In a separate measurement, we visited the landing page of the selected websites 1,000 times and recorded the used cookies to test if there is a difference if users visit the landing pages or subsites.

We compared the number of TP cookies set in each measurement of the pre-study and found that subsites of websites typically set significantly more cookies than the respective landing page does. In our measurement, the mean amount of cookies used increased by approx. 20 (41 %), when visiting subsites rather than only the landing page. This shows that if one wants to perform cookie/third party measurements, one should always include subsites to the measurement setup rather than only measuring landing pages. Furthermore, we measured a mean increase of 12 cookies (27 %) per website visit if a browser is used that already has cookies in the local cookie storage. When it comes to the change of cookie usage based on the number of visited subsites, we found that the mean amount of accessed/set cookies stabilizes around 50 (SD: 100; median at 12) after visiting 100 subsites (see Figure 2). In conclusion, to magnify the number of cookies set, we use a browser profile that has cookies set and visit 100 subsites and the landing page of each website.

4.3.2 Measurement Sequence. We used the same method to create the browser profile for our experiment crawls that we utilized in the pre-study. This profile is loaded before each website visit but is not altered. Hence, each website visit uses the same profile and the order of visited websites does not impact the results. In total, we conduct the measurements from three different locations (Europe (DE), North America (US), and Asia (JP)) to account for possible geographical differences [6]. For all measurements, we used two computers located at a European university. For each of our regional measurement runs, we created a new distinct browser profile. We used a commercial VPN service (*NordVPN*) to obtain an IP address from the locations outside the EU. Using a VPN service comes with the risk that it might inject content into the communication stream [25]. However, we did not find any hints of this practice for the used service, neither in the Terms of Service nor publicly on the Internet.

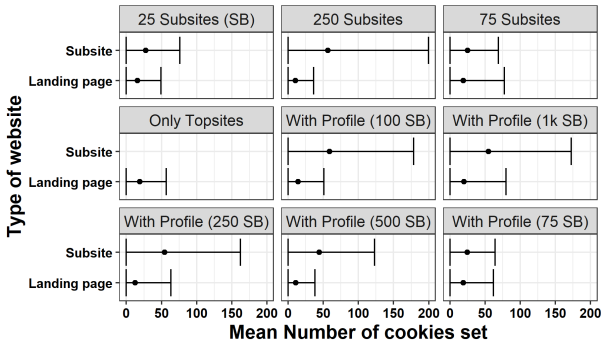

Figure 2: Mean number of cookies set in our pre-study with the corresponding standard derivation

We configured *OpenWPM* to visit the landing page of each website and to gather all first-party hyperlinks on that site (subsites) one day before the first measurement. Therefore, some of these links might not be present on the front page anymore at the time we perform the measurements from different regions or might not exit anymore after all. We did so to increase comparability between our measurements since we visited the same landing pages and subsites in each measurement. Additionally, we collect all first-party hyperlinks on the subsites but only use them (in random order) if there are not enough subsites linked on the landing page. Afterward, we choose 100 random subsites that we used during the experiment crawls. In each measurement, we visited 549,715 (SD 16,851) distinct URLs on average.

4.4 Cookies

A cookie is a key-value pair set on a client by a visited website or third-party present on that website. In this work, we count every single key-value pair as one cookie, and not all textual data stored on the client, because each pair can be used for different purposes. We heuristically group cookies in different categories based on their lifetime. As for HTTP cookies, we compute the lifetime of a cookie-based on the `expires` attribute and the timestamp when the request/response was sent/received, or JavaScript command was executed. If we cannot determine the lifetime of a cookie or if it is negative, we consider a cookie as a "Session" cookie, which is deleted by the browser when the HTTP session ends. In total, we use four lifetime categories: (1) "Session", (2) "Short" ($\leq$ 1 week), (3) "Persistent" ($\leq$ 1 year), and (4) "Permanent" ($>$ 1 year). We used the evolution of maximum cookie lifetimes in the Safari browser, enforced through the *Intelligent Tracking Prevention* [3], as an orientation to determine them.

4.4.1 Cookie classification. Cookies can be used for various means. We want to asses the specific purposes why third parties set cookies and which purposes are most dominant to get a better understanding of real-world cookie usage. We use the following cookie

type classes defined by the *International Chamber of Commerce UK* [22]: (1) "Strictly Necessary Cookies" are needed to provide basic functionality of a website, (2) "Performance Cookies" aggregate (anonymously) user's usage of the website, (3) "Functionality Cookies" personalize the website's usage, and (4) "Targeting/Advertising Cookies" are used to track users or to display them personalized ads. For our analysis, we used *Cookiepedia*, a platform that provides public classifications of cookie classes [38]. This process might be error-prone as cookie classes are assigned by hand but are—from our point of view—the best approximation of online cookie usage today. In total, we can classify 45.3 % of all observed cookies.

4.5 Third Party Trees

In this work, we evaluate the number of partners loaded by an embedded third-party object. To do so, we model *third party trees* (TPTs) for each visited URL (for each landing page and all subpages, respectively), which include all third parties loaded on the visited page. A similar concept was used by Ikram et al. [21] and Kumar et al. [28] to analyze resource loading dependencies (termed "inclusion chains"). We extend this concept as we visit several sites of a single domain, which enables us to construct a more comprehensive and realistic view of a website's dependencies, and we do not limit ourselves to JavaScript inclusions. We use the term *tree* rather than *chain* as our concept describes a more complete view of a website's TP relations and not a single instance of TP inclusion.

We build the trees based on the analysis of JavaScript, iframes, and Cascading Style Sheets (CSS) that can be used to load third-party code dynamically. Other HTML objects (e. g., images) can also be requested from third parties, but these objects cannot load additional code dynamically and would not spawn any children in the tree. In our analysis, we omit these objects if they are located right below the root ($depth = 0$) but consider them if they occur as leaves in longer branches. We omit them because they would make the results harder to interpret as one cannot decide if these parties do not load further third parties or simply cannot do so. However, we consider these objects in our general analysis (see Section 5.1). A third party tree is designed to show which party is responsible for loading another party. To account for HTTP redirects, we substitute the respective TLD+1 with the redirects TLD+1 in the trees and delete all edges that create a redirection loop. Therefore, we add each loaded script and inserted iframe as a child of the respective ancestor (script/frame) in the tree, if needed. For example, if a script, which is loaded from *foo.com*, loads another script from *bar.com* we add *bar.com* as a child of *foo.com* in the tree. Thus, we can measure the number of third parties loaded due to each embedded object. Regarding iframes, we use *openWPM*'s feature to save the nested iframe structure of a website. Based on this structure, we insert each frame (i. e., the source TLD+1) at the corresponding position in the tree. For JavaScript code, we inspect the call stack of each script, test if code from another party is executed (e. g., a function in an external library), and include this party at the respective position in the TPT (based on the call stack entries). To find CSS dependencies introduced through the @import command, we analyze the content type of HTTP requests and test if the origin and target of the request URL both load CSS. Eventually, each TPT consists of all scripts, style sheets, and iframes loaded by a website. Each branch of a tree represents the sequence in which different third parties (domains) were embedded.

```
Node(adidas.com)                (Visited Website)
  |- Node(MediaMath [C])        depth=0; breadth=4
  |  |- Node(Improve Digital)    depth=1; breadth=0
  |  |- Node(PubMatic [C])
  |  |- Node(OpenX [C])
  |  |- Node(Index Exchange)
  |- Node(TrustArc)
  |- Node(Adobe)
  |  |- Node(Tealium [C])
  |  |  |- Node(Akamai [C])      depth=2
  |  |  |  |- Node(Instana)      depth=3
  |  |  |  |  |- Node(Adobe)     depth=4
```

Figure 3: Example of an observed third party tree. The listed companies represent the companies operating the observed URLs. [C] illustrates the cookie setting parties.

If not stated otherwise, we use the TLD+1 of a third party domain as the node identifier; otherwise, we use the companies associated with the TLD+1. We use the *WhoTracks.me* database [5] to link domains to the respective companies owning them. Thus, a branch in the tree could consist of multiple domains operated by the same company (e. g., *foo.com* → *googletagmanager.com* → *googleapis.com* → *youtube.com*). However, we collapsed requests stemming from one company into one leaf. In the previous example, we would not add *googleapis.com* even if *youtube.com* would load a script form that domain. We did so because otherwise, the resulting trees would result in a much deeper length if several resources were loaded from the same TLD+1. For example, if *foo.com* was embedded and would than load *metric.foo.com*, subsequently *ad.foo.com* and finally *foo.com/?ad_loaded=1* the resulting branch would be much deeper. Overall, the maximum depth using this more lax approach would increase by magnitudes from eight to 52. Thus, a branch consists of all TLD+1/companies that could perform a task on the client.

An example of a third party tree is given in Figure 3, including the companies' names, not TLD+1s. The tree shows the visited website (*adidias.com*), the directly embedded third parties (*MediaMath*, *TrustArc*, and *Adobe*—$depth = 0$), the partner of the third partners (fourth parties at $depth = 1$—e. g., *Improve Digital*), and further embedded services e. g., *Akamai* ($depth = 2$) or *Instana* ($depth = 3$). The services that actively set cookies are marked with a [C]. The example illustrates that by embedding a single service, many other direct partners of that third parties might be embedded into a website (e. g., *MediaMath* embeds four partners). Furthermore, embedding a single third party might implicitly lead to a long branch of direct and indirect partners of the used third party (e. g., *Adobe* that creates a branch of $depth = 4$). Note that at depth four, a service from *Adobe* is embedded. This is not a loop, but simply, the previously loaded party utilizes a different service of *Adobe*.

5 RESULTS

We conducted our measurements in the second quarter of 2019 and found around 93 % of the landing pages in our dataset to be accessible. The remaining websites provided services that seem not

Table 1: General overview of our three measurement crawls. The number of visited websites and subsites with the corresponding number of observed TPs, cookie setting TPs (C TPs), and used cookies is shown.

Region		Websites	Subsites	TPs	C TPs	Cookies
Europe	(EU)	9,267	561,087	12,076	5,393	20.6M
Asia	(AS)	9,266	530,356	12,926	5,815	18.2M
N. America	(US)	9,333	557,702	13,687	6,115	20.4M

to be intended for rendering in a web browser (e. g., APIs) or did not exist anymore. In total, we visited over 1.5 million websites that embedded over 37,000 third parties producing over 4.5 TB of data. More than 17,000 third parties access/set over 59 million cookies across all website visits in our experiment. An overview of our measurements is given in Table 1.

5.1 General Overview

First, we tested how many cookies are set/accessed when visiting subsites in contrast to the respective landing pages to test the potential bias in previous studies that focused on the landing page only. In our measurements, as shown in Figure 4, subsites set considerably more (36 %) cookies than the respective landing pages. On average, 55 cookies were set when loading a landing page while 78 were set when a subsite was accessed. The difference between the number of cookies used by third parties is statistically significant when comparing (1) different categories (ANOVA test p-value < 0.001) and (2) when comparing landing pages to subsites (p-value < 0.001) However, we did not find a statistically significant effect of the originating region of the visit and the rank of the website on the cookie setting behavior. Our results show that landing pages of websites show a different cookie usage behavior than the respective subsites as those make more usage of third parties. To get a better understanding of the implications of increased cookie usage, we analyze the primary purposes of why cookies are set.

5.1.1 Lifetime and Cookie Types. Aside from the number of cookies set, it is interesting to analyze why they are set and how long they stay active in the browser. Overall, we could classify 45.3 % of all observed cookies in terms of distinct used keys. Regarding absolute numbers, we could classify 74 % of all observed cookies. Most of the observed cookies are used to track website visitors or to provide targeted ads (99 %). The "type" of the cookie shows a strong correlation with the amount of cookie set for this type (p-value < 0.0001). This means that specific types of cookies are set more often than others. Furthermore, the purpose of a cookie is not related to its lifetime, a X^2 test does not show a correlation between "type" and "lifetime". Furthermore, third parties use similar types and lifetimes for their cookies, no matter on which website they are embedded in. We did not find a correlation between the "type" or "lifetime" of a cookie and the website's category. Our results show that cookies are overwhelmingly used to track users or to provide them with targeted ads. Furthermore, cookies in all categories use various lifetimes. Given the primary purpose of cookies ("Targeting/Advertising") and the measured increased usage of cookies on subsites, we see that subsites show different behavior in that regard

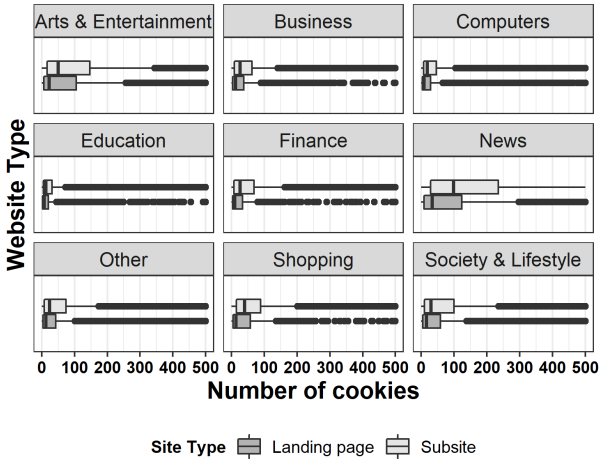

Figure 4: Mean number of cookies used by each visited landing page and each respective subsites, by category of the visited website. To increase the readability, we capped the bars at 500. 1.8 % of sites had a higher number of cookies; this doesn't impact the computed values.

(see also Section 5.2). Tracking users on subsites provides a more comprehensive view of their online activities. For example, visiting the landing page of an online shop does not necessarily indicate which products a user is interested in, but this information can be extracted on subsites.

5.1.2 Legal Compliance. With the introduction of the General Data Protection Regulation (GDPR) [46] and the California Consumer Privacy Act (CCPA) [4], service providers have to be more aware of business partners they work with. If a business partner tracks users or uses personal information in other ways and is not located in a GDPR adequate member state [19] or not a member of the Privacy Shield [47], they need to agree on a data processing contract (Article 28 §3 GDPR) that "appropriate safeguards" (Article 46 §1 GDPR) are taken which enforce privacy rights of EU citizens. Based on the IP addresses observed in our measurements (see Section 4.3), we analyzed if connections were established to IP addresses that are associated with countries that are not a member of the EEA or part of the Privacy Shield. In the remainder of the paper, we call these parties "non-adequate" or "possibly problematic" to improve the reading flow of this work. Note that every business can agree by contract that the data of EU citizens are processed according to EU legislation and, therefore, these parties might pose no problem at all (Article 28 §3 GDPR). However, the current legal debate only focuses on TPs as "joint controllers" [11, 20] and does not cover fourth or further parties. We want to highlight that a binary classification of what is compliant with legal regulation and what is not is impossible to make without looking at the specific service agreements between websites and third parties.

Figure 5 shows the origins and targets of all requests for which service providers need to make sure that they have taken appropriate safeguards. These numbers only refer to our EU measurement, and the results are not violations of the legislation, but provide

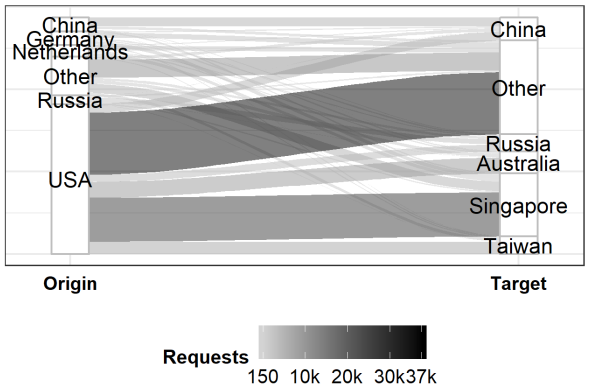

Figure 5: Origins (left) and targets (right) of requests to services whose IP address is not mapped to an IP in an adequate country.

insights to potential data flows that might conflict with the legal requirements. The origins/targets are based on the observed IP addresses in our measurements. Overall, 4.7 % of all cookies were set by services outside adequate geolocations and only 7.1 % of the visited domains (TLD+1) exclusively used TPs that are located at adequate geolocations. Domains using only adequate locations are located in the US (59 %), followed by Germany (7 %), and the United Kingdom (3 %). In our dataset, Singapore is the most prevalent target of non-adequate requests (26 %), followed by China (5 %) and Australia (5 %). The US is the most common origin of such requests (63 %), followed by China (6 %) and Germany (5 %). We did not find a statistically significant impact of the region on the question of whether or not a third party from non-adequate geolocation is used. When looking at the services located in possibly non-adequate geolocations, we found that almost half only used sometimes (53 %), and the other half always used possibly non-adequate geolocations (47 %). Overall, roughly 10 % of all observed TPs used IP addresses in possibly problematic geolocations.

In the following, we analyze the services that use sometimes adequate and sometimes non-adequate geolocations. This is an interesting subset as service providers might not be aware of the possibility that these TPs change their geolocations over time. In contrast, third parties that always send data to possibly problematic geolocations are more easy to identify and, therefore, the transfer of data to these non-adequate countries are likely part of the data processing contracts. Requests to TPs that only sometimes used adequate geolocations were most of the time resolved to an EU IP address but sometimes (< 1 %) to addresses outside the EU. For example, sometimes a similar resource of a third party was requested from different locations in the same measurement. Meaning, the URL *csm.ad-network.foo* was resolved to sgp.*csm.ad-network.foo* in Singapore and nl.*csm.ad-network.foo* in the Netherlands. This is challenging as service providers cannot ensure that only EU endpoints of the used third party are used. In our measurement, *gstatic.com* (a service operated by *Google*) with 20 % of all inclusions of possibly non-adequate services and *upravel.com* (a Russian advertising service) with 15 % are the top services that might pose a problem to service providers. The next service only accounts for 1 % of these possibly conflicting services (i. e., there is a long tail distribution). One likely explanation is that these are effects of load balancing or similar techniques and that the servers belonging to these IP addresses are controlled by the same third party. However, service providers need to account for this behavior in the data processing contracts with the TP, and the TP must assure that GDPR adequate data processing rules are in place no matter where their servers are located.

Summary. Our results show that measuring only landing pages of websites might only reveal a fraction of the websites' real use of third parties. Furthermore, we found that websites make extensive use of cookies, primarily to serve ads or to track users, and we observed that some embedded TPs might be conflicting with current legislation. To further investigate the effects of visiting subsites and not only landing pages, it is interesting to look at further areas that might be implicated by our findings.

5.2 Replication and Comparison

To provide a more comprehensive overview of our measurements in comparison with previous work, we tried to replicate the main findings of previous work using our data set. We differentiate between studies we could replicate using our data (●—see column "*Rep.*" in Table 2) and studies we would partly replicate (◑). Furthermore, we indicate ("*Res.*") if we could produce similar results (✓). To reproduce the results, we analyzed the landing pages of each website (if the paper did so) or used the same amount of subsites. If we could replicate the results, we measure them on all visited subsites to test if these studies measured a comprehensive generalizable view or as shown in our study, subsites show a different behavior ("*Scales*"). We differentiate if visiting subsites makes a measurable difference in contrast to only visiting landing pages (✗). The results are given in Table 2. Our replication studies do not aim to replicate *all* results of previous work, but we only focus on the main takeaways and results closely related to our work. We do not claim that our replications are sound or complete, but we tried to faithfully replicate previous work as good as possible using our data set.

In contrast to Dabrowski et al. [6], and as previously stated, we could not find statistical evidence that the originating region of a request influences cookie setting practices in general. On the one hand, this could be a result of different experimental setups as we tried to maximize the "cookie setting behavior" of each website to achieve more generalizable results. Dabrowski et al. used a headless browser that can be easily detected by websites and, therefore, might affect the loaded TPs (e. g., ads might not be loaded to counter ad fraud). On the other hand, we performed our experiment on a larger scale and interacted (e. g., scrolling) with the websites, which could fundamentally affect the results.

Furthermore, we found that subsites set significantly more cookies than the respective landing pages. As for the results of Sørensen et al. [45], we could verify that the GDPR has no immediate effect on third party usage. Sanchez-Rola et al. [42] show that opting-out of cookies often has no measurable effect on cookie setting practices in the field. We could only partly reproduce this work as we never interacted with any cookie banners, but our results show

Table 2: Overview of previous work we tried to replicate (Rep.), the scale of the work ("LP" := landing page, "SB" := subsite), the results (Res.) of our replication, and if these experiments show different behavior in a vertical setup (Scales).

1st Author	Ref.	Year	Venue	Scale	Main finding	Rep.	Res.	Scales
Dabrowski	[6]	2019	PAM	LP	Websites set 49 % less cookies if user located in the EU visit them.	●	✗	✓
Sørensen	[45]	2019	WWW	LP + ⌀9 SB	Effects of the GDPR to third-party usage is not definite.	●	✓	✗
Sanchez-Rola	[42]	2019	AsiaCCS	LP	Tracking is often still present even if opted-out.	◑	✓	✗
Urban	[50]	2020	AsiaCCS	LP + 3–5 SB	Cookie syncing reduced by around 40 %.	◑	✓	✓
Merzdovnik	[34]	2017	EuroS&P	LP + 2 SB	State of the art tracking blocking tools can limit user tracking but still have blind spots.	◑	✓	✓
Englehardt	[9]	2016	CCS	LP	Websites use various fingerprintig methods.	○	–	✓
Kumar	[28]	2017	WWW	LP	Implicitly included TPs pose a challenge when upgrading to HTTPs.	●	✓	✗
Ikram	[21]	2019	WWW	LP	Implicitly included parties might pose a security threat.	●	✓	✓
Iordanou	[23]	2018	IMC	user browsing behaviour	In the EU, tracking data is transferred across countries but rarely leaves the EU.	●	✓	✓

that cookies are still widely used and that there are no regional differences, while in the EU users should opt-in before cookies are being used. We used data of our prior work collected before the GDPR became effective [50]. Using this data and comparing the regional data in our experiments, we could verify that cookie syncing seems to be influenced by different legislation. Scaled to our collected data, we found an increase of cookie syncing activities on subsites in contrast to landing pages. This replication cannot be seen as representative as our measurement misses essential features, especially to identify IDs, to assess cookie syncing since we only used one profile in each region.

To test whether our results of increased cookie usage on subsites also applies to user tracking, we use the numbers presented by Merzdovnik et al. [34] on the presence of trackers on websites as a baseline. To test if a tracker is active on a website, we use the *EasyPrivacy* List [8], which is a list combining URLs of known trackers. However, we do not test whether anti-tracking tools are useful or not. In our measurement, we found that trackers mostly occur on subsites in comparison to their respective landing pages (an increase of approx. 6 %). 2.5 % of the measured websites do not embed any trackers on the landing page but use trackers on subsites. Overall, we could show that tracking on subsites increases and that future work concerning this area should include subsites into their measurement. In terms of overall tracking occurrence, we produced results comparable to the "plain" profile used by Merzdovnik et al. Finally, we tested the prevalence of device fingerprinting scripts in our data set, as previously studied by Englehardt et al. [9]. As the scripts identified by Englehardt et al. are probably outdated, we only found four of them in our total dataset, we used the popular "*Fingerprint2*" library [12] to test for the presence of such trackers. Hence, our results can be seen as a lower bound as we only test for the presence of one script. We identified a mean increase of device fingerprinting of 25 % on subsites in contrast to the respective landing pages. In all three measurements, we found 13 domains (0.14 %), which did not use the script on the front page but on subsites. Overall, we found the tracking script on 0.15 % of the landing pages while Englehardt et al. identified device fingerprinting on 1.8 %, and the most common script on 0.45 % of the analyzed websites.

Summary. In this section, we demonstrated that only measuring landing pages hides the scale of different phenomena observable on the Web. Furthermore, the behavior of TPs differs on different subsites, which raises the question to what extent service providers are in control of TPs embedded into their services. To tackle this challenge, one needs to understand relations between TPs and the determinism of which third parties will be loaded into a service.

5.3 Third Party Trees

As described above, we are interested in understanding dependencies between third parties and possibly resulting in challenges for service providers and users. Therefore, we created *third party trees* (see Section 4.5) to better understand the implications of embedding a single third party into a website. Figure 6 shows the depth of the measured third party trees by category of the visited websites. Remember that each visited website (i. e., distinct URL) produced its own TPT, and the directly embedded third parties are of depth zero. The average third party branch has a depth of one (median also one), and the deepest branch of a tree we found has a depth of eight. In total, 43.0 % of the observed branches have a depth of one or more, which means that these trees include parties that are not necessarily known to the service provider. Therefore, several third parties (in terms of TLDs+1, not distinct companies) load at least one additional partner. Each node in the trees has, on average, 0.9 (SD 37) direct children (breadth) with a maximum of 361, and each branch compromises on average 0.9 (SD 6.4) different companies (max 127). In total, 2,901 TPs (10 %) are embedded that never included any child. The depth of a tree is impacted by the category of a website (p-value < 0.0001). Similar to the results of previous work, "News" websites tend to use more cookies and third parties [45]. As over 40 % of all TPs at least load one additional partner, it is interesting to look if these use cookies, for example, to track users or to serve them targeted ads.

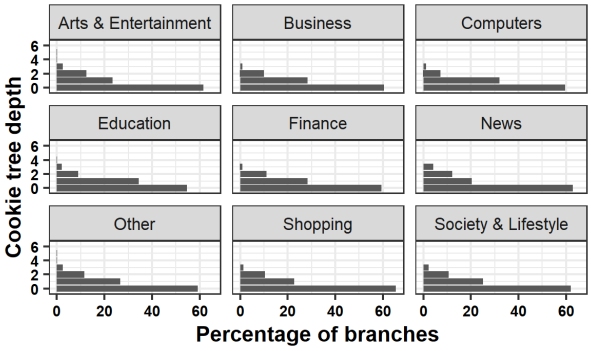

Figure 6: Relative distribution of the measured third party tree depth split by the websites' categories.

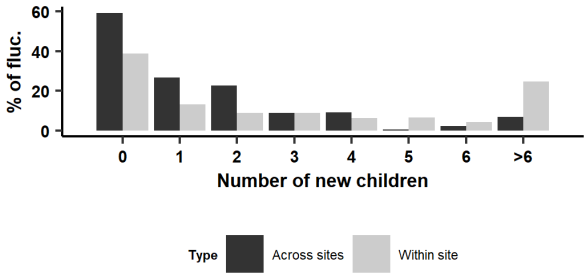

Figure 7: Children included in only some of the branches (fluctuation) created by a specific TP within each visited site (grey) and across all sites (black).

5.3.1 Cookies Set in Trees. Not every party in each TPT, more specifically in each branch, will necessarily set a cookie. Therefore, we analyzed the depth of the cookie setting parties and the overall amount of cookies set in each branch. We limit ourselves to cookies but expect, based on our results presented in Section 5.2, that other privacy-invasive techniques would likely produce similar results. Starting with the depth of set cookies, on average, 1.5 parties in each branch do *not* set a cookie. In 48 % of all branches no party and only in 125 branches (approx. 0.002 %) all parties set a cookie. The website's category and its rank both show statistical significance in d the number of cookies set in each branch (both p-values < 0.001). Furthermore, we found that deeper branches do not necessarily, in relative numbers, lead to more cookies being set. As for the depth on which cookies are being set, we found that most cookies (72 %) are set by the fourth party ($depth = 1$). The main reason why most cookies are set on depth one is likely because most trees are of depth one. Hence, deeper trees occur less often and, consequently, in absolute numbers, set fewer cookies.

Overall, slightly more than 18 % of cookies are set on a depth larger one (fifth party or higher). If service providers want to choose services that do not use cookies, for example, because they want to protect their customers from tracking, they face the problem that often the fourth party sets a cookie. Therefore, service providers have to carefully monitor the behavior of all embedded third parties for such behavior. Since one-fifth of cookies are set in depth one, it is worth investigating how much control or knowledge service providers have about these parties. TPs that always include the same third parties can be seen as more predictable because the third parties do not change, and service providers know which third parties will be included in their websites. Furthermore, TPs that do not create deep branches are better to assess for service providers since hierarchies and dependencies are easier to understand. Therefore, we analyze the deterministic of branches generated by directly embedded TPs.

5.3.2 Determinism of Third Party Trees. The determinism of each branch that is generated by an embedded TP is import if service providers want to understand which TPs are loaded and who is responsible for loading them. If it is known, before loading the third party object, which other third parties might be embedded, service providers can evaluate the potential risks of a TP for their users. Therefore, we tested the fluctuation of embedded companies for each TP in the measured trees. First, we tested the fluctuation within each visited website (TLD+1) and its subsites. Meaning that we test which third parties are embedded into the visited website by each observed third party on a specific subsite in a specific region. Secondly, we tested the fluctuation across all websites and all regions, meaning that we test if a wider spread view of a third party provides more insight of the further loaded parties or if they show different behavior on different websites.

Half of the branches (50.4 %) have at least one fluctuating partner in them. Figure 7 shows the measured fluctuation of a TP within (gray) and across (black) the visited sites. The x-axis shows the relative amount of fluctuating companies in all branches of an embedded third party. Zero means branches of this TP always include the same third parties, and six means that six distinct TPs only accrued in some of the branches. These numbers *exclude* third parties that never had any children because these would naturally be zero and might lead to a false conclusion about the deterministic of TPs. The results show that almost a third (62 %) of third parties that embed other third parties use fluctuating partners (e. g., due to real-time ad bidding) when loaded on different subsites. Across all regions, we see that there is a long tail distribution of companies that only occur in some of the branches, note the increase in more than six new children. Regarding the impact of the originating region, in which the measurement was performed, we found no statistical significance on its impact on the fluctuation. However, the weighted mean (local) fluctuation was the highest in the US (5.78) and lowest in the EU (5.49).

On a global scale, we find a different picture. We see that the global fluctuation in the EU is more distributed than it is in other regions. We found no statistical evidence that the region affects the local or global fluctuation of children. In conclusion, we see that measuring TPs on a global scale does not necessarily provide a generalizable view as some TPs behave differently on different sites (e. g., due to the advertised products or partners in different regions). Our results show that the list of third parties embedded in a website is not deterministic, which makes it challenging for service providers to account for all TPs that might be present on their websites. Embedding some third parties leads to an often

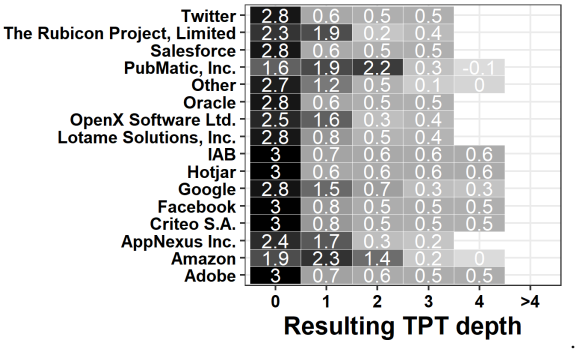

Figure 8: Resulting branch depth of objects embedded by different companies (scaled for each individual company).

changing set of embedded third parties (e. g., different TPs providing ads). However, service providers only have little control over these processes as they often depend on third parties to provide their service. As the (non-)deterministic of these trees is related to the embedded TP, it is interesting to analyze the depth of trees generated by different TPs (companies).

5.3.3 Companies. Figure 8 shows the average, scaled branch depth that is created by embedding a single object of different companies. All values are scaled for each company, not overall, and include all TLDs+1 operated by the company. Thus, Figure 8 presents the resulting depth of each company and does account for the overall occurrence of each company. Furthermore, the figure only lists the top 15 companies, regarding absolute amounts of embeddings of these companies. All remaining companies are combined in the category "Other". The top companies account for over 98 % of absolute third party embeddings. In general, embedding most TPs results in short trees of depth zero. However, ad-tech companies—the primary source to finance many websites—offer a more widespread resulting TPT depth (e. g., *PubMatic* or *Rubicon Project*) which reduces the options to choose partners that do not load many other partners. We found statistical significance that the embedded company impacts the depth of the generated tree (p-value ≈ 0.008). Regarding the position of companies in the trees, we found that larger companies (e. g., *Google* or *Facebook*) occur mostly at depth zero (absolute numbers) while service providers of TPs (e. g., companies that counter ad fraud) occur deeper in the trees.

Summary. Our results indicate that it is quite challenging for service providers to keep track of all third parties that might be embedded into their services. Furthermore, before loading the directly embedded TP, it is often not definite which other parties might be loaded—especially ad networks load various fluctuating partners.

6 LIMITATIONS

In the following we discuss limitations of our work. We use the classification of *Cookiepedia*, which might be wrong to some extent and is incomplete. We could only classify slightly over 45 % of all observed cookies but show that an overwhelming majority (99 %) tracks users or serves targeted advertisements. We mapped requests from different services to a single company, if possible. If we observed multiple requests to domains owned by one company (e. g., ads.foo.com and fonts.foo.com), we collapsed them to a single request if they occurred in sequence. Our measurement platform, a customized *OpenWPM* instance, does not interact with any cookie banners that are present on the visited websites. Hence, we do not capture cookies set by third parties that honor opt-in choices of (European) users. However, previous works demonstrated that cookie consent notices often do not offer choices to opt-in [51], do not work at all [42], and that the used libraries often are not complaint to current legislation [7]. Therefore, our results are a lower bound since (1) we shortened the TPTs and (2) some cookies might only be used after affirmative action of the user.

7 DISCUSSION

We have shown the challenges service providers face when they rely on third-party code and try to account which third parties are loaded when users use their service. It is the high dynamic and previously nominal regulation of the Web that now presents challenges to service providers. As service providers might carefully select the directly embedded third parties (e. g., ad networks), they cannot control which third parties might get included when these third parties loaded their content (e. g., due to ad real-time bidding). The primary tool website providers have to solve these challenges are data processing contracts that include indirectly embedded third parties. From a research perspective, we have shown that a simple horizontal scaling of websites to visit (i. e., websites from a given toplist) is not sufficient to measure a phenomenon of interest. Meaning that future work should (1) scale their experiments vertically and (2) previous results of different Web measurement areas should be re-visited to measure the given challenges adequately. Finally, our assessment of purposes if cookies underlines the dire need of privacy protection mechanisms to limit cookie-based tracking—which is currently promoted by several browser vendors (e. g., *Firefox* [35], *Chrome* [18] and *Safari* [3]).

8 CONCLUSION

In this work, we have analyzed the cookie setting practices of the top 10k websites on the Web. We found that 99 % of all cookies we could classify were set with the intention to track users or to serve them targeted ads. Furthermore, we modeled *third party trees*, which assemble all third parties embedded into a website and loading dependencies among them. By analyzing the third party trees, we found that the median depth of such trees is one (max eight), that there is a sever fluctuation of children in different branches with the same parent node (third party), that especially ad networks result in longer tree branches, and that only 7 % of all visited websites (TLD+1) never embedded a third party that might pose possible legal problems. Moreover, we have shown that studies that only measure landing pages of websites miss a substantial amount of embedded third parties and cookies set.

ACKNOWLEDGMENTS

This work was partially supported by the Ministry of Culture and Science of the State of North Rhine-Westphalia (MKW grants 005-1703-0021 "MEwM" and Research Training Group NERD.nrw). We would like to thank *Cybot* (*Cookiebot*) for their support.

REFERENCES

[1] Gunes Acar, Marc Juarez, Nick Nikiforakis, Claudia Diaz, Seda Gürses, Frank Piessens, and Bart Preneel. 2013. FPDetective: Dusting the Web for Fingerprinters. In *Proceedings of the 2013 ACM Conference on Computer and Communications Security (CCS '13)*. ACM Press, New York, NY, USA, 1129–1140.

[2] Adobe Inc. 2017. Flash & The Future of Interactive Content. https://theblog.adobe.com/adobe-flash-update/ Accessed: 2019-10-05.

[3] Apple Inc. 2019. Intelligent Tracking Prevention 2.3. https://webkit.org/blog/9521/intelligent-tracking-prevention-2-3/ Accessed: 2019-04-24.

[4] California State Legislature. 2020. The California Consumer Privacy Act.

[5] Cliqz. 2018. WhoTracks.me Data - Tracker database. https://whotracks.me/blog/gdpr-what-happened.html Accessed: 2019-04-24.

[6] Adrian Dabrowski, Georg Merzdovnik, Johanna Ullrich, Gerald Sendera, and Edgar Weippl. 2019. Measuring Cookies and Web Privacy in a Post-GDPR World. In *Proceedings of the 2019 Conference on Passive and Active Measurement (PAM '19)*. Springer-Verlag, Cham, 14.

[7] Martin Degeling, Christine Utz, Christoper Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. 2019. We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. In *Proceedings of the 2019 Symposium on Network and Distributed System Security (NDSS '19)*. Internet Society, San Diego, California, USA, 20.

[8] EasyList. 2019. EasyPrivacy. https://easylist.to/easylist/easyprivacy.txt Accessed: 2019-04-24.

[9] Steven Englehardt and Arvind Narayanan. 2016. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM Conference on Computer and Communications Security (CCS '16)*. ACM Press, New York, NY, USA, 1388–1401.

[10] Steven Englehardt, Dillon Reisman, Christian Eubank, Peter Zimmerman, Jonathan Mayer, Arvind Narayanan, and Edward W. Felten. 2015. Cookies That Give You Away: The Surveillance Implications of Web Tracking. In *Proceedings of the 24th World Wide Web Conference (WWW '15)*. ACM Press, New York, New York, USA, 289–299.

[11] European Court of Justice. 2018. Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein vs Wirtschaftsakademie Schleswig-Holstein GmbH, - Case C-210/16. http://curia.europa.eu/juris/document/document.jsf?text=&docid=202543&doclang=EN&d&part=1&cid=341550

[12] Fingerprint.js. 2019. Fingerprint.js is the most advanced open-source fraud detection JS library . https://fingerprintjs.com/ Accessed: 2019-04-24.

[13] David Formby, Preethi Srinivasan, Andrew Leonard, Jonathan Rogers, and Raheem Beyah. 2016. Who's in Control of Your Control System? Device Fingerprinting for Cyber-Physical Systems. In *Proceedings of the 2016 Symposium on Network and Distributed System Security (NDSS '16)*. Internet Society, San Diego, California, USA, 15.

[14] Imane Fouad, Nataliia Bielova, Arnaud Legout, and Natasa Sarafijanovic-Djukic. 2020. Missed by Filter Lists: Detecting Unknown Third-Party Trackers with Invisible Pixels. In *Proceedings of the 20th Privacy Enhancing Technologies Symposium (PETS '20)*. Springer-Verlag, Berlin, Heidelberg. https://hal.inria.fr/hal-01943496

[15] Gertjan Franken, Tom Van Goethem, and Wouter Joosen. 2018. Who Left Open the Cookie Jar? A Comprehensive Evaluation of Third-party Cookie Policies. In *Proceedings of the 27th USENIX Security Symposium (SEC '18)*. USENIX Association, Berkeley, CA, USA, 151–168.

[16] Alejandro Gómez-Boix, Pierre Laperdrix, and Benoit Baudry. 2018. Hiding in the Crowd: An Analysis of the Effectiveness of Browser Fingerprinting at Large Scale. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. ACM Press, New York, New York, USA, 11.

[17] Roberto Gonzalez, Lili Jiang, Mohamed Ahmed, Miriam Marciel, Ruben Cuevas, Hassan Metwalley, and Saverio Niccolini. 2017. The cookie recipe: Untangling the use of cookies in the wild. In *Proceedings of the 2017 Network Traffic Measurement and Analysis Conference (TMA '17)*. IEEE, Piscataway, NJ, 1–9.

[18] Google Inc. 2020. Building a more private web: A path towards making third party cookies obsolete. https://blog.chromium.org/2020/01/building-more-private-web-path-towards.html Accessed: 2020-01-15.

[19] Government Digital Service. 2019. Countries in the EU and EEA. https://www.gov.uk/eu-eea Accessed: 2019-10-05.

[20] Higher Regional Court, Düsseldorf, Germany. 2018. Opinion of Advocate General Bobek on Fashion ID GmbH & Co. KG vs Verbraucherzentrale NRW eV - Case C-40/17. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=ecli:ECLI:EU:C:2018:1039

[21] Muhammad Ikram, Rahat Masood, Gareth Tyson, Mohamed Ali Kaafar, Noha Loizon, and Roya Ensafi. 2019. The Chain of Implicit Trust: An Analysis of the Web Third-Party Resources Loading. In *Proceedings of the 2019 World Wide Web Conference (WWW '19)*. ACM Press, New York, NY, USA, 2851–2857.

[22] International Chamber of Commerce UK. 2012. *Cookie guide*. International Chamber of Commerce UK.

[23] Costas Iordanou, Georgios Smaragdakis, Ingmar Poese, and Nikolaos Laoutaris. 2018. Tracing Cross Border Web Tracking. In *Proceedings of the 2018 Internet Measurement Conference (IMC '18)*. ACM Press, New York, NY, USA, 329–342.

[24] Hugo Jonker, Benjamin Krumnow, and Gabry Vlot. 2019. Fingerprint Surface-Based Detection of Web Bot Detectors. In *Proceedings of the 2019 European Symposium on Research in Computer Security (ESORICS '19)*. Springer-Verlag, Cham, 586–605.

[25] Mohammad Taha Khan, Joe DeBlasio, Geoffrey M. Voelker, Alex C. Snoeren, Chris Kanich, and Narseo Vallina-Rodriguez. 2018. An Empirical Analysis of the Commercial VPN Ecosystem. In *Proceedings of the 2018 Internet Measurement Conference (IMC '18)*. ACM Press, New York, NY, USA, 15.

[26] Radhesh Krishnan Konoth, Emanuele Vineti, Veelasha Moonsamy, Martina Lindorfer, Christopher Kruegel, Herbert Bos, and Giovanni Vigna. 2018. MineSweeper: An In-depth Look into Drive-by Cryptocurrency Mining and Its Defense. In *Proceedings of the 2018 ACM Conference on Computer and Communications Security (CCS '18)*. ACM Press, New York, NY, USA, 1714–1730.

[27] David M. Kristol. 2001. HTTP Cookies: Standards, Privacy, and Politics. *ACM Trans. Internet Technol.* 1, 2 (Nov. 2001), 151–198.

[28] S. Kumar, S. S. Rautaray, and M. Pandey. 2017. Malvertising: A case study based on analysis of possible solutions. In *Proceedings of the 2017 International Conference on Inventive Computing and Informatics (ICICI '17)*. IEEE, San Francisco, United States, 288–291.

[29] Andreas Kurtz, Hugo Gascon, Tobias Becker, Konrad Rieck, and Felix C. Freiling. 2016. Fingerprinting Mobile Devices Using Personalized Configurations. *Proceedings of the Privacy Enhancing Technologies Symposium* 2016, 1 (2016), 4–19.

[30] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS '19)*. Internet Society, San Diego, California, USA, 20.

[31] MaxMind Inc. 2019. GeoIP Databases & Services. https://www.maxmind.com/en/geoip2-services-and-databases Accessed: 2019-10-05.

[32] J. R. Mayer and J. C. Mitchell. 2012. Third-Party Web Tracking: Policy and Technology. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy (S&P '12)*. IEEE, San Francisco, United States, 413–427.

[33] McAfee LLC. 2019. Customer URL Ticketing System. https://trustedsource.org/ Accessed: 2019-10-05.

[34] G. Merzdovnik, M. Huber, D. Buhov, N. Nikiforakis, S. Neuner, M. Schmiedecker, and E. Weippl. 2017. Block Me If You Can: A Large-Scale Study of Tracker-Blocking Tools. In *Proceedings of the 2017 IEEE European Symposium on Security and Privacy (EuroS&P '17)*. IEEE, San Francisco, United States, 319–333.

[35] Mozilla Corporation. 2019. Today's Firefox Blocks Third-Party Tracking Cookies and Cryptomining by Default . https://blog.mozilla.org/blog/2019/09/03/todays-firefox-blocks-third-party-tracking-cookies-and-cryptomining-by-default/ Accessed: 2019-04-24.

[36] Nick Nikiforakis, Wouter Joosen, and Benjamin Livshits. 2015. PriVaricator: Deceiving Fingerprinters with Little White Lies. In *Proceedings of the 24th World Wide Web Conference (WWW '15)*. ACM Press, New York, New York, USA, 820–830.

[37] N. Nikiforakis, A. Kapravelos, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna. 2013. Cookieless Monster: Exploring the Ecosystem of Web-Based Device Fingerprinting. In *Proceedings of the 2013 IEEE Symposium on Security and Privacy (S&P '13)*. IEEE, San Francisco, United States, 541–555.

[38] OneTrust LLC. 2019. Cookiepedia. https://cookiepedia.co.uk/ Accessed: 2019-10-05.

[39] Xiang Pan, Yinzhi Cao, and Yan Chen. 2015. I Do Not Know What You Visited Last Summer: Protecting Users from Third-party Web Tracking with TrackingFree Browser. In *Proceedings of the 2015 Symposium on Network and Distributed System Security (NDSS '15)*. Internet Society, San Diego, California, USA, 15.

[40] Python. 2019. tldextract 2.2.1. https://pypi.org/project/tldextract/ Accessed: 2019-04-24.

[41] Jan Rüth, Torsten Zimmermann, Konrad Wolsing, and Oliver Hohlfeld. 2018. Digging into Browser-based Crypto Mining. In *Proceedings of the 2018 Internet Measurement Conference (IMC '18)*. ACM Press, New York, NY, USA, 70–76.

[42] Iskander Sanchez-Rola, Matteo Dell'Amico, Platon Kotzias, Davide Balzarotti, Leyla Bilge, Pierre-Antoine Vervier, and Igor Santos. 2019. Can I Opt Out Yet?: GDPR and the Global Illusion of Cookie Control. In *Proceedings of the 2019 ACM Symposium on Information, Computer and Communications Security (AsiaCCS '19)*. ACM Press, New York, New York, USA, 340–351. https://doi.org/10.1145/3321705.3329806

[43] Muazzam Siddiqui, Morgan C. Wang, and Joohan Lee. 2008. Data Mining Methods for Malware Detection Using Instruction Sequences. In *Proceedings of the 26th International Conference on Artificial Intelligence and Applications (AIA '08)*. ACTA Press, Anaheim, CA, USA, 358–363.

[44] Aditya K Sood and Richard J Enbody. 2011. Malvertising – exploiting web advertising. *Computer Fraud & Security* 2011, 4 (2011), 11 – 16.

[45] Jannick Kirk Sørensen and Sokol Kosta. 2019. Before and After GDPR: The Changes in Third Party Presence at Public and Private European Websites. In *Proceedings of the 2019 World Wide Web Conference (WWW '19)*. ACM Press, New York, New York, USA, 11.

[46] The European Parliament and the Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L 119/1.

[47] The International Trade Administration. 2019. The Privacy Shield. https://www.privacyshield.gov/ Accessed: 2019-10-05.

[48] Tobias Urban, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. 2019. "Your Hashed IP Address: Ubuntu.": Perspectives on Transparency Tools for Online Advertising. In *Proceedings of the 35th Anual Computer Security Applications Conference (ACSAC '19)*. ACM Press, New York, NY, USA, 702–717.

[49] Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. 2019. A Study on Subject Data Access in Online Advertising After the GDPR. In *Data Privacy Management, Cryptocurrencies and Blockchain Technology (DPM'19)*, Cristina Pérez-Solà, Guillermo Navarro-Arribas, Alex Biryukov, and Joaquin Garcia-Alfaro (Eds.). Springer International Publishing, Cham, 61–79.

[50] Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. 2020. Measuring the Impact of the GDPR on Data Sharing. In *Proceedings of the 15th ACM Symposium on Information, Computer and Communications Security (AsiaCCS '20)*. ACM Press, New York, NY, USA.

[51] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. 2019. (Un)informed Consent: Studying GDPR Consent Notices in the Field. In *Proceedings of the 2019 ACM Conference on Computer and Communications Security (CCS '19)*. ACM Press, New York, NY, USA, 973–990.

[52] Antoine Vastel, Pierre Laperdrix, Walter Rudametkin, and Romain Rouvoy. 2018. FP-STALKER: Tracking Browser Fingerprint Evolutions. In *Proceedings of the 39th IEEE Symposium on Security and Privacy (S&P '18)*. IEEE, San Francisco, United States, 728–741.

[53] Q. Xu, R. Zheng, W. Saad, and Z. Han. 2016. Device Fingerprinting in Wireless Networks: Challenges and Opportunities. *IEEE Communications Surveys Tutorials* 18, 1 (2016), 94–104.