Towards Understanding First-Party Cookie Tracking in the Field

Nurullah Demir[1], Daniel Theis[2], Tobias Urban[3], Norbert Pohlmann[4]

Abstract: Third-party tracking is a common and broadly used technique on the Web. Different defense mechanisms have emerged to counter these practices (e.g. browser vendors that ban all third-party cookies). However, these countermeasures only target third-party trackers and ignore the first party because the narrative is that such monitoring is mostly used to improve the utilized service (e.g. analytical services). In this paper, we present a large-scale measurement study that analyzes tracking performed by the first party but utilized by a third party to circumvent standard tracking preventing techniques. We visit the top 15,000 websites to analyze first-party cookies used to track users and a technique called "DNS CNAME cloaking", which can be used by a third party to place first-party cookies. Using this data, we show that 76% of sites effectively utilize such tracking techniques. In a long-running analysis, we show that the usage of such cookies increased by more than 50% over 2021.

Keywords: first-party tracking, cookies, privacy, CNAME cloaking, tracking methods

1 Introduction

The business model of many modern (Web) applications relies on the revenue generated by "renting" space on their services to advertisement companies. These ad-tech companies try to place advertisements on the sites that meet the users' interests motivating that they will interact with the ad, and ultimately buy the advertised product or service. To place such targeted ads, ad-tech companies track users across the Web, by assigning a unique identifier to each of them, and try to understand their interests by building so-called *behavioral profiles* [MC10]. The unique user identifiers are often stored in the third-party context (e.g. in an HTTP cookie). Some consider this large-scale tracking as privacy-invasive because it often happens without users' explicit consent or knowledge [TH11], nor is the tracking made transparent to the user. The desire for more privacy and the need for more user data led to an arms race between anti-tracking tools, and novel techniques to track users. One recent (technical) step in this race was the announcement of major browser vendors to ban

[1] Institute for Internet Security – if(is), Westphalian University of Applied Sciences and KASTEL Security Research Labs, Karlsruhe Institute of Technology demir@internet-sicherheit.de
[2] Institute for Internet Security – if(is), Westphalian University of Applied Sciences, Neidenburger Straße 43, 45897, Gelsenkrichen, Germany, theis@internet-sicherheit.de
[3] Institute for Internet Security – if(is), Westphalian University of Applied Sciences and secunet Security Networks AG, urban@internet-sicherheit.de
[4] Institute for Internet Security – if(is), Westphalian University of Applied Sciences, Neidenburger Straße 43, 45897, Gelsenkrichen, Germany, pohlmann@internet-sicherheit.de

third-party cookies within the next years [Go20a; Mo20a]. While there is no immediate problem with third-party cookies, previous work showed that they are overwhelmingly used for advertisement purposes [Ur20a]. Hence, trackers need to find different ways to persist their identifiers on the users' devices. One known way to do so is the computation of *browser fingerprints*, which are distinct identifiers that are computed based on properties of the user's device or browser [EN16; La20]. However, these fingerprints change over time and, therefore, one cannot simply rely on them for tracking purposes [GLB18].

One way to cache such identifiers is to store them in a first-party context, and send them to a third party if needed. More specifically, a tracking script embedded in the first-party context of a site could store the identifier (cookie) in the first-party context and send it to the tracker in a dedicated request. Hence, deleting, or banning third-party cookies does not affect them. In this paper, we perform a large-scale measurement study on the top 15,000 sites to analyze the presence of such first-party tracking techniques in the field and use the HTTP Archive [HT21] to analyze the development of such tracking techniques. In our measurement, we find that roughly 90% of all sites include a first-party object that tracks users, and leaks the identifier to a third party. Furthermore, our results show that 10,354 (69%) of the sites in our analysis corpus hide the presence of a tracker by redirection of request on DNS level ("DNS CNAME cloaking"). Using the HTTPArchive [HT21], we show that first-party tracking, in cooperation with a third party, has been a common phenomenon in the past with a growing popularity. Our results show that the leakage of such cookies to a third party increased by nearly 50%. Previous work that analyzed the tracking ecosystem almost exclusively focused on third-party trackers from various perspectives (e.g. [Ac14; En15; EN16; Fo20; GLB18]). Other works focused on analyzed tracking attempts performed in the first-party context using CNAME cloaking [Di21; DMF20; Re21] (see Section A). CNAME cloaking is one way to pace for a third party to place an identifier in the first-party context of a site (see Section 3.2). In this work, we focus on first-party tracking via cookies and do not aim to analyze tracking enabled by CNAME cloaking, but we see it as one way to place such cookies. To be more precises, we analyze the combined usage of CNAME cloaking and first-party cookie tracking in the field – and not each technology in isolation. One important insight from our work is that tracking is no longer a phenomenon only present in the third-party context but has arrived in the first-party context at large scale. Consequently, many of the current anti-tracking tools might have to be revised to match this new tracking scope, that future studies might have to reconsider their measurement setups, and that banning third-party cookies might have less impact than it is currently expected [Lt20]. We make the following key contributions:

1. We perform a large-scale web measurement of the top 15,000 sites on the Web, and analyze if they utilize first-party identifiers. Our results suggest that first-party tracking – at the behest of a third party – arrived at scale on the Web.
2. In our experiment, over 85% of the analyzed sites store potential tracking identifiers in a first-party cookie, and send them to a third party. Utilizing the *HTTPArchive*, we show that this technique has already been used in 2019 and has ever grown since.

3. Finally, we analyze the companies participating in the first-party tracking ecosystem. We find that trackers who dominate the third-party tracking market also dominate the first-party tracking market.

2 Method

As basis for our analysis, we chose to use the *Tranco* top 15,000 list generated on 07/16/2020 (ID: PX8J)[5], which is an aggregation of other popular top lists [Le19], in our analysis. We visited each site on the list and collected 30 distinct first-party hyperlinks to pages on the same site, which we used in our measurement crawls. If possible, we repeated the process to recursively collect up to 30 subsites. We chose to visit 30 pages instead of only analyzing the landing page of a site because recent work has shown that "subsites" show increased usage of privacy-invasive technologies and that visiting 30 sites is a good approximation to attest the privacy impact of a website [Aq20; Ur20a]. Due to this need for a vertical measurement setup, the number of distinct websites we analyze is limited compared to other measurement studies (e.g. [En15]).

To measure the usage of first-party tracking and the leakage of such cookies, we utilize the popular *OpenWPM* framework [EN16], which uses the Firefox browser to visit websites. We configured the framework to log all (1) HTTP requests/responses and (2) data stored in the local storage or cookie jar. Furthermore, we instrumented Firefox's DNS API [Mo20b] to log the DNS resolution of all hostnames, which is necessary to identify CNAME cloaking (Section 2.1). A detailed description of our framework can be found in Appendix B and a detailed description of CNAME cloaking is given in Appendix A. We use four different browser configurations ("profiles") for our measurement: (1) No particular configuration, (2) disabled 3rd party cookies, (3) active anti-tracking tool (i. e. *uBlock Origin* [Hi20]; henceforth, "*uBlock*"), and (4) disabled 3rd party cookies, and an active anti-tracking tool (i. e. (2) and (3) combined). We chose to use the option to block third-party cookies to test if trackers use other techniques (e.g. first-party tracking techniques) if they cannot store their identifiers as they usually do. We used *uBlock* because – to the best of our knowledge – it is the only tool at the time of writing this paper that claims to identify and block CNAME cloaking. Hence, the tool provides protection against trackers that utilize this new technique, and we can assess, concerning the other profiles, how effective it is. While the named defence mechanisms do not actively aim to protect against tracking performed by a third party they might block requests to a third party that contains the first-party identifier, which would effectively limit the tracking impact. For each profile, we performed a separate measurement crawl. Hence, we visited each page on our website corpus four times, once with each profile.

[5] Available at https://tranco-list.eu/list/PX8J.

2.1 Measuring CNAME Cloaking

CNAME cloaking is increasingly used to track users [Di21; DMF20]. We analyze this practice's presence by inspecting the DNS resolutions on the client by instrumenting the DNS resolution of the Firefox browser. More specifically, we log the URL handled by the browser and the responding DNS resolution observed by the browser, which we instrumented. For example, if we observe a request to *tracking.foo.com/search=foo* whose domain is then cloaked to *cloaked-party.com*, we use the URL *cloaked-party.com/search=foo* in our analysis but treat it a first-party request. In contrast to previous work in the field of CNAME cloaking, this allows us to analyze the real DNS resolution of each request. Previous work often performed the DNS resolutions isolated from the web measurement on different machines and at different times [Di21; DMF20]. This comes with the non-negligible assumption that DNS resolutions will not change over time and/or based on the location of users, which could change the outcome of a study. However, we want to highlight that our contribution is not the measurement of tracking that is enabled by third parties but that we are interested in tracking identifiers that are stored in a third-party context. One way to store such cookies is via CNAME cloaking. In order to compare our results to previous work, we match the resolved URL against two standard tracking filter lists, as of 04/16/2021 (i. e. the *EasyList* and *EasyPrivacy* lists [Ea20]). Such lists always come with the downside that they might not contain all tracking URLs [Fo20; Me17] and, therefore, our results can be seen as a lower bound. However, since we used lists generated after the crawl, we think they should contain most trackers. As we describe in Appendix A, CNAME cloaking is not a problem per se, but it could be used to hide the real recipient of a request. For example, a service provider who wants to host a domain on a content delivery network but wants to keep his domain (brand) name often uses CNAME cloaking to do so [DMF20].

2.2 First-Party Cookie Tracking

Since browser vendors limit the use of third-party cookies [Go20a; Mo20a], trackers need to find different means to store them. One way to store them is to use first-party cookies because these are currently not restricted. One challenge in this approach is that we have to distinguish between session cookies, which are usually not stored in third-party cookies, and user identifiers. If the same identifier reoccurs on different visits to pages of the same site, we assume that it is a user identifier and not a session ID. This assumption is valid because we perform a stateless crawl, and session IDs will reset between two visits since the local storage and cookie jar are reset. We assess a cookie to be a first-party cookie if it was set by a request or script that was loaded in a first-party context. Hence, if we visit *foo.com* and the site sets a cookie, we count it as a first-party cookie. However, let us assume *bar.com* embeds an object from *foo.com* that sets a cookie; we do not count this cookie in our first-party analyses – because it was set in a third-party environment. In theory, cookies are simple `name=value` pairs, yet Gonzalez et al. show that a single cookie often contains multiple values in proprietary formats (e.g. `key=[v1=foo;v2=bar]`) [Go17].

To identify a tracking cookie, we use the following commonly accepted definition [Ac14; En15; KTK20; Ur20a; Ur20b] : (1) it is not a session cookie and has a lifetime of more than 90 days; (2) has a length of at least eight bytes (to hold enough entropy); (3) is unique in each measurement run, and the length of each value only differs by up to 25%; and (4) The values of the cookie are similar according to the Ratcliff/Obershelp [RM88] string comparison algorithm (≤60%). We resort to these definitions to allow a basic comparison of our and previous work. It is essential to mention that first-party cookies might hold an identifier to implement analytical or similar services, which are essentially a form of local user tracking. If services send the (local) identifier to a third party, which offers the service, this third party could track the user. For example, if we find the first-party cookie `OptanonConsent=ABC-123`, which is probably used by *OneTrust*, and observe a third-party request that holds this ID (e.g. *tracking.foo.com?**consentId=ABC-123***), we assume that *tracking.foo.com* could potentially track the user with ID *ABC-123*. Hence, to find potential tracking that utilized first-party cookies, one has to find instances where they leak it to any third party. In our analysis, we identify such cookie leakage by inspecting the HTTP GET and POST parameters of all third-party requests. We compare them with all cookie values, that could be used for tracking (see the definition above). If a first-party cookie is leaked in such a way, we assume that it can be potentially used to track users. One challenge in this approach is that we have to distinguish between session cookies, which are usually not stored in third-party cookies, and user identifiers. If the same identifier reoccurs on different visits to pages of the same site, we assume that it is a user identifier and not a session ID. This assumption is valid because we perform a stateless crawl, and session IDs will reset between two visits since the local storage and cookie jar are reset.

2.2.1 Cookie Tracking Over Time

The introduced methods potentially used to perform first-party tracking do not rely on any new or recently introduced techniques. Hence, it is reasonable to assume that if such practices are used that they were used in the past. To understand whether such practices were used in the past we utilize the *HTTPArchive* [HT21]. HTTPArchive is an initiative to measure and analyze how the modern Web is built. To achieve this, HTTPArchive crawls the landing pages of the most popular origins, based on the *Chrome User Experience Report* (CrUX) [Go20b], on a monthly basis since 01/2019. Thus, we can analyze the first-party tracking capabilities of all websites in the archive over time. Due to the size of the archive, which includes millions of websites in each measurement, we decided to perform the analysis on a quarterly basis. Hence, the first measurement point in our dataset is 01/19, the second one 04/19, and the last one 07/21. In the analysis, we extract all cookie values from the raw data provided in the archive. Using the raw data, we can access all necessary data that we need in our analysis to identify tracking cookies (e.g. lifetime). Similar to our approach in our active measurement, we first identify first-party cookies that could be used for tracking purposes and then test if such cookies are leaked to a third party in an HTTP GET or POST request. An important limitation of this approach is that it excludes all cookies set via JavaScript (i. e. we only see cookies set via the HTTP protocol).

#	Configuration	Date	# Sites	# Pages	1st-party c.	3rd-party c.	CNAME Cl.
1	Plain browser	07/20	11,471	272,659	408,509	155,193	14,493,533
2	No 3rd-party cookies	07/26	10,368	246,493	355,208	—*	11,017,643
3	*uBlock Origin* active	08/01	10,203	230,327	139,115	89,876	5,829,786
4	#2 and #3 combined	08/06	10,153	225,315	142,608	—*	5,831,884

Tab. 1: Overview of all measurements. ∗ Firefox deletes these cookies after creation.

3 Results

First, we want to provide an overview of the measured dataset. We conducted all four measurement runs in three consecutive weeks (each measurement took approx. five days). We conducted the first measurement (#1—no particular configuration) on 07/20/2020, and the last measurement (#4) ended on 08/12/2020. Across all measurement, we visited 272,659 distinct pages on 11,471 sites. In total, we visited 974,794 URLs across our four measurements. In total, 1,290,509 cookies were set during these visits (1,045,440 first-party and 245,069 third-party cookies). Across all measurements, we observed 37,172,846 instances of CNAME cloaking of which 783,917 (2%) ultimately resulted in tracking attempts, based on the classification of the resolved URL. A summary of the measurement runs is given in Table 1. The drop in analyzed sites and, consequently, pages can be attested to sites no longer being available (e.g. HTTP 404 errors). According to the numbers, it seems that blocking third-party cookies has no effect on the usage of first-party cookies, but using an anti-tracking tool, *uBlock* in our case, leads to a decrease in the usage of both types of cookies. Furthermore, the tool seemingly helps to block CNAME cloaking attempts.

3.1 First-Party Cookie Leakage

In the following, we discuss the leakage of first-party tracking cookies. Furthermore, we provide a general analysis of CNAME-based first-party tracking cookies in Appendix C. As we have shown 20% of the first party cookies could be used to track different users across a site (see Appendix C). In this section, we evaluate to what extent these cookies are leaked to third parties. A straightforward way to send such cookies is to include them in HTTP GET or POST parameters. Overall, we found that 68% (143,216) of the cookies that we identified as first-party trackers are sent to a single third party. A common practice in the tracking ecosystem is the so-called *cookie-syncing*, which means that two or more third parties exchange a user identifier. In our experiment, we find that first-party tracking cookies are shared, on average, with 1.3 third parties, across all profiles. 1,182 (0.8%) are shared with more than one third party. Hence, real cross-domain tracking and "cookie-syncing" has not arrived in the first-party context at scale. We identified 2,253 distinct domains that received such cookies, in 5,931,727 requests. Our results show that first-party tracking cookies are often leaked to a third party and they could be used to track users. A more detailed analysis of the companies receiving cookies this way can be found in Section 3.3.

Defense Mechanisms In profile #3 and #4, we enabled the *uBlock* extension to test its effectiveness in protecting users. The tool does not primarily aim to protect users' from cookie value leakage. However, since it blocks some URLs that are knowingly used for tracking purposes it might also prevent first-party cookie leakage. In our analysis, we observed 43% fewer requests that contained a cookie value. This decrease consequently leads to a decrease of 98,819 (31%) tracking cookies leaked and 456 (20%) fewer third parties getting access to the cookie. Similar to our findings on first-party tracking cookies, the leakage of them is also positively impacted by the used extension. This is probably also explained by the fact that the tool blocks requests to different third parties

3.1.1 First-Party Cookie Tracking Over Time

Our active measurement shows that first-party cookie tracking and leakage are severe and joint problems on the Web. In the following, we want to analyze the development of first-party cookie-based tracking usage. To achieve this, we use data from 2019, 2020, and 2021 that we extracted from the *HTTPArchive* (see Section 2.2.1). Across all 11 measurement points (four in 2019 and 2020 each and three in 2021), we analyzed 15,805,385 websites, which issued 5,592,914,767 requests. Our dataset contains 119,676,680 first-party cookies. Of those cookies 9,489,765 (8%) could potentially be used to track users, according to our definition. In total, 42,068 (0.3%) of the websites actively sent a cookie to a third party, and 321,644 (0.3% of the first-party cookies, and 3.4% of the tracking cookies) were leaked in such away. In the following, we only analyse sites that appeared in all measurement points (1,777,206 sites). In total, 13,116 (0.7%) of the websites actively sent a cookie to a third party, and 63,235 (0.2%) of the first-party cookies, and 2.4% of tracking cookies were leaked in such away. We reason that this high discrepancy to our active measurement is because of the large amount of less popular sites in the *HTTPArchive* and our finding that popular sites tend to use first-party tracking more often than less popular ones.

Figure 1 provides an overview of the temporal development of first-party cookie based tracking on the Web. As one can see, the number of leaking parties and leaked cookies are directly related to each other. This relation shows that a number of websites seem to use a more or less fixed set of services (e.g. a consent management system) that use the first-party context to track users. Overall, the data point at 07/21 contains 118% more first parties that use the analyzed tracking cookies than the first measurement point (07/19). The number of third-party trackers that receive such cookies remains relatively stable, with an increase from 456 (01/19) to 621 (07/21), which accounts for 36%. This development indicates that there is a steadily growing set of players in the ecosystem that utilize this technique (see Section 3.3), but more and more websites adopt these services of these players.

3.2 First-Party Cookie Tracking via CNAME Cloaking

Previously, we discussed straightforward attempts of first-party cookie tracking, which presumably happens in coordination with a third party. Another option for a third party to

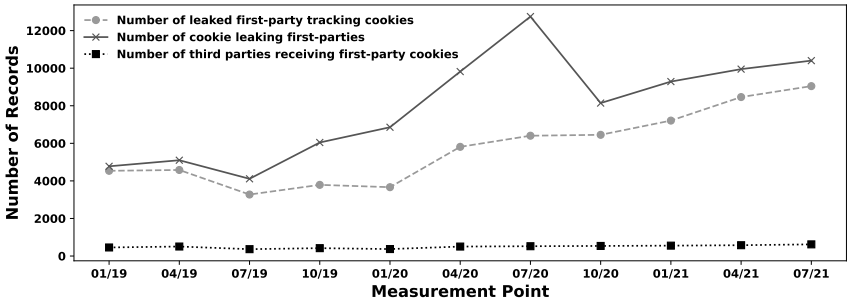

Fig. 1: Number of distinct sites and third parties that engage in first-party tracking and number of leaked first-party tracking cookies, over time.

set and access cookies in a first-party context is that the first party redirects requests on DNS level to another domain ("CNAME cloaking" – see Section 2.1). Utilizing CNAME cloaking, HTML objects (e.g. JavaScript) operate in the first-party context (e.g. *t.foo.com*) on browser level but are actually loaded from another domain (e.g. *tracking.org*) on DNS level. Across all four measurements in our dataset, we found over 37M requests for which the DNS name was cloaked (45% of all requests). Of those requests 2.1% (783,917) are sent to an URL that is marked to be used for tracking purposes, by the lists mentioned in the Section 2.1 *after* the cloaking of the CNAME.

Cloaked First-Party Cookies: In this work, we focus on the usage of first-party cookies to track users. Aside from cloaking the tracking URL, one way to abuse CNAME cloaking for tracking is to place tracking cookies in the first-party context. Therefore, we are interested in analysing if cloaked requests use the first-party context to store (tracking) cookies. Overall, we found 69,961 (6.7%) of all first-party cookies that are set by such requests. According to our definition of tracking cookies (see Section 2.2), 28% (19,676) of those cookies can be used to track users. Overall, the identified tracking cookies are utilized by 8,508 (74%) sites in our dataset. Hence, it seems that the third parties actively try to circumvent anti-tracking methods. In our dataset, more popular sites (lower rank) and sites of a specific category (e.g. "News") utilized CNAME based cookie tracking.

Defense Mechanisms *uBlock* provides tracking protection for CNAME cloaking based tracking as it actively inspects the DNS resolution of requests and makes decisions based on the result of it. Other tools operate on the objects (e.g. URLs) accessible on browser level and, therefore, cannot detect DNS level tracking approaches. Therefore, we are interested in the effect of the 'lower level' DNS blocking and the impact on users' privacy. In our measurement, we used the *uBlock* browser extension in profile #3 and #4. In both profiles, we see a decrease of CNAME cloaking of around 46% what results in a decrease of 15% in identified trackers, which use this technology. However, we still found trackers that were not detected by *uBlock*, which is in line with previous work that showed that list based blockers

miss some trackers [Me17]. The blocking of the identified requests resulted in a decrease of 101,387 (21%) first-party tracking cookies. Our results show that modern anti-tracking tools need to adapt to modern first-party tracking techniques – especially CNAME tracking.

3.3 First-Party Cookie Tracking Ecosystem

In the previous section, we have analyzed the extent of first-party tracking – emphasizing first-party tracking cookies – which are set through various means. In the following, we shed some light on the ecosystem behind this first-party tracking and the companies that are active in it. Concerning the websites that utilize first-party cookies for tracking, we found that 10,719 (91%) sites use them. To get a better understanding of who uses such techniques, we use, on the one hand, the ranking (according to the *Tranco* list [Le19]) and, on the other hand, the category of the website's content (see Section 2) to test for usage differences. We used the X^2 test ($\alpha = 0.5$) to evaluate statistical significance in these categories. Both categories (rank and category) show a strong correlation (p-value < 0.0001) in the usage of first-party tracking cookies. More specifically, we found that popular websites (rank $\leq 3,000$) and websites of categories like "Business", "News", or "Education" tend to use more of such tracking techniques. This finding hints that websites are more aware of upcoming changes respectively current protection mechanisms and try to circumvent them.

Companies Active in the Ecosystem We use the *Cookiepedia* (see Section B) to map identified cookies to companies using them. Using this approach, we could classify 95,236 (36%) of the observed cookies. These can be attributed to 283 distinct companies. As previously briefly noted, the aggregated view of active companies, that use first-party tracking cookies, shows that *Google* (54%) and *Facebook* (14%) are the most common ones, followed by *Adobe* (8%) and *OneTrust* (3%). However, we see a different picture when we look at the companies to which the cookies are leaked. *Dynamic Yield* (17%) and *Adobe* (15%) are the top companies that receive them, these results are in line with previous work [DMF20]. *Google* only receives 7% of all leaked cookies.

4 Related Work

Online tracking has received a lot of attention from the academic community and the industry alike and, therefore, a huge body of work exists that analyzes tracking methods along different dimensions. Quantifying Web tracking though measurement is a popular way to analyze different tracking techniques like device fingerprinting (e.g. [EN16; Va18]), tracking cookies (e.g. [Ac14; En15]), connections between tracking companies (e.g. [PKM18]), other ways to track users (e.g. [Fo20]), and the efficiency of privacy-preserving tools to prevent racking (e.g. [Me17; RKW12]). The impact of newly introduced legislation on online tracking (like the CCPA or GDPR) was also analyzed in detail by different works (e.g. [Sa19; SK19; Ur20b]). However, all of these works focus on tracking performed by third parties that are

embedded into a website. Our work analyzes a new trend in the tracking market that moves the tracking code to the first party.

Recently, different works focused on CNAME cloaking. In mid 2020, Dao et al. were the first ones to analyze first-party tracking by performing a large scale Web measurement to identify the usage of CNAME cloaking [DMF20]. In their paper, they show that this technique has already been utilized for several years and they highlight the limitation of current privacy preserving technologies to counter such tracking. Dao et al. also provide a first mitigation strategy for CNAME based cloaking, using machine learning [DF20]. Complementary to our work, Ren et al. take a look at the effect of CNAME cloaking on browser cookie policies and how they propagate in the ecosystem [Re21]. They find that CNAME cloaking is often used to circumvent browser policies and that sensitive data is leaked by this bypass. Dimova et al. perform a rigorous analysis of CNAME cloaking-based tracking and discuss privacy and security issues that arise by the usage of this technique [Di21]. Finally, the dangers of CNAME cloaking is also discussed in several non-academic articles(e.g. [Co19; La21]) and on browser vendor pages (e.g. [Wi20]), which hints the piratical relevance of the problem. Most recently, Quan et al. also took a look at cookie-based first-party tracking [Ch21]. They use an elaborated JavaScript taint analysis framework to understand the process how first-party cookies are set and by whom. They show that nearly all of the popular websites (97%) utilize first-party cookies for user tracking. Our approach differs from the named work because, on the one hand, we do not limit our analysis to one technique only but analyse first-party tracking along different axes and take a look at the implications of first-party cookie tracking, sometimes enabled by CNAME cloaking (e.g. by analysing defence mechanisms, the ecosystem, or temporal development). Hence, we do not analyze novel tracking techniques in isolation but aim to understand how they are used together.

5 Discussion and Conclusion

This paper analyzed the scale of (potential) first-party tracking techniques on the top 15,000 websites. In contrast to previous work, we have shown that tracking is not exclusively implemented through third-party means but that first-party content is used for such purposes as well. Hence, limiting or blocking third-party content on websites will not prevent that users will be tracked by third parties. Our work shows that first-party cookies and requests that are redirected on DNS level are used for these purposes. One takeaway from these findings is that privacy-preserving technologies need to extend their attempts to the first-party context while still covering third-party contexts. First-party cookies that hold a personal identifier constitute a privacy risk since users might assume that those are exclusively used by the first party and are needed to run the service. However, we have shown that they are often leaked to a third party, which could abuse them for tracking purposes. While the extend of tracking and used methods (e.g. first-party cookie syncing) are not yet present at scale we argue that they will gain in importance once first-party tracking is more commonly used.

Acknowledgements This work was partially supported by the German Federal Ministry for Economic Affairs and Energy (grant 01MK20008E "Service-Meister" and grant 01MN21002H "IDunion"), the Ministry of Culture and Science of North Rhine-Westphalia (MKW grant 005-1703-0021 "MEwM"). Furthermore, we would like to thank Norman Schmidt and Katharina Meyer for their efforts in developing the analysis pipeline.

References

[Ac14] Acar, G.; Eubank, C.; Englehardt, S.; Juarez, M.; Narayanan, A.; Diaz, C.: The Web Never Forgets: Persistent Tracking Mechanisms in the Wild. In: ACM Conference on Computer and Communications Security. CCS '14, 2014.

[Aq20] Aqeel, W.; Chandrasekaran, B.; Feldmann, A.; Maggs, B. M.: On Landing and Internal Web Pages: The Strange Case of Jekyll and Hyde in Web Performance Measurement. In: ACM SIGCOMM Internet Measurement Conference. IMC'20, 2020.

[Ch21] Chen, Q.; Ilia, P.; Polychronakis, M.; Kapravelos, A.: Cookie Swap Party: Abusing First-Party Cookies for Web Tracking. In: International Conference on World Wide Web. WWW, 2021.

[Cl18] Cliqz GmbH: WhoTracks.me Data—Tracker database, `https://github.com/cliqz-oss/whotracks.me/tree/master/whotracksme/data`, 2018.

[Co19] Cointepas, R.: CNAME Cloaking, the Dangerous Disguise of Third-party Trackers, https://medium.com/nextdns/cname-cloaking-the-dangerous-disguise-of-third-party-trackers-195205dc522a, 2019.

[DF20] Dao, H.; Fukuda, K.: A Machine Learning Approach for Detecting CNAME Cloaking-based Tracking on the Web, 2020.

[Di21] Dimova, Y.; Acar, G.; Olejnik, L.; Joosen, W.; Van Goethem, T.: The CNAME of the Game: Large-scale Analysis of DNS-based Tracking Evasion. Privacy Enhancing Technologies Symposium, PoPETS '21 3/, 2021.

[DMF20] Dao, H.; Mazel, J.; Fukuda, K.: Characterizing CNAME Cloaking-Based Tracking on the Web. In: Network Traffic Measurement and Analysis Conference. TMA, 2020.

[Ea20] EasyList: EasyPrivacy, `https://easylist.to/`, 2020.

[En15] Englehardt, S.; Reisman, D.; Eubank, C.; Zimmerman, P.; Mayer, J.; Narayanan, A.; Felten, E. W.: Cookies That Give You Away: The Surveillance Implications of Web Tracking. In: International Conference on World Wide Web. WWW, 2015.

[EN16] Englehardt, S.; Narayanan, A.: Online Tracking: A 1-Million-Site Measurement and Analysis. In: ACM Conference on Computer and Communications Security. CCS, 2016.

[Fo20] Fouad, I.; Bielova, N.; Legout, A.; Sarafijanovic-Djukic, N.: Missed by Filter Lists: Detecting Unknown Third-Party Trackers with Invisible Pixels. Privacy Enhancing Technologies Symposium, PoPETS '20 2/, 2020.

[GLB18] Gómez-Boix, A.; Laperdrix, P.; Baudry, B.: Hiding in the Crowd: An Analysis of the Effectiveness of Browser Fingerprinting at Large Scale. In: International Conference on World Wide Web. WWW '18, 2018.

[Go17] Gonzalez, R.; Jiang, L.; Ahmed, M.; Marciel, M.; Cuevas, R.; Metwalley, H.; Niccolini, S.: The cookie recipe: Untangling the Use of Cookies in the Wild. In: Network Traffic Measurement and Analysis Conference. TMA, 2017.

[Go20a] Google Inc.: 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, 2020.

[Go20b] Google Inc.: Chrome User Experience Report | Tools for Web Developers, https://developers.google.com/web/tools/chrome-user-experience-report?hl=de, 2020.

[Hi20] Hill, Raymond: uBlock Origin, https://github.com/gorhill/uBlock/, 2020.

[HT21] HTTP Archive: The HTTP Archive Tracks How the Web is Built, https://httparchive.org, 2021.

[KTK20] Koop, M.; Tews, E.; Katzenbeisser, S.: In-Depth Evaluation of Redirect Tracking and Link Usage. Privacy Enhancing Technologies Symposium, PoPETS '20 4/, 2020.

[La20] Laperdrix, P.; Bielova, N.; Baudry, B.; Avoine, G.: Browser Fingerprinting: A Survey. ACM Transactions on the Web 14/2, 2020.

[La21] Lakshmanan, R.: Online Trackers Increasingly Switching to Invasive CNAME Cloaking Technique, https://thehackernews.com/2021/02/online-trackers-increasingly-switching.html, 2021.

[Le19] Le Pochat, V.; Van Goethem, T.; Tajalizadehkhoob, S.; Korczyński, M.; Joosen, W.: Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In: Symposium on Network and Distributed System Security. NDSS, 2019.

[Lt20] Ltd., R. M.: When the third-party cookie crumbles, https://www.raconteur.net/marketing/third-party-cookies/, 2020.

[MC10] McDonald, A. M.; Cranor, L. F.: Americans' Attitudes About Internet Behavioral Advertising Practices. In: ACM Workshop on Privacy in the Electronic Society. WPES, 2010.

[Mc20] McAfee LLC: Customer URL Ticketing System, https://trustedsource.org/, 2020.

[Me17] Merzdovnik, G.; Huber, M.; Buhov, M.; Nikiforakis, N.; Neuner, S.; Schmiedecker, M.; Weippl, E.: Block Me If You Can: A Large-Scale Study of Tracker-Blocking Tools. In: IEEE European Symposium on Security and Privacy. EuroS&P, 2017.

[Mo20a] Mozilla: 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/, 2020.

[Mo20b] Mozilla Inc.: DNS JavaScript API, https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons/WebExtensions/API/dns, 2020.

[On19] OneTrust LLC: Cookiepedia, https://cookiepedia.co.uk/, 2019.

[PKM18] Papadopoulos, P.; Kourtellis, N.; Markatos, E. P.: The Cost of Digital Advertisement: Comparing User and Advertiser Views. In: International Conference on World Wide Web. WWW, 2018.

[Re21] Ren, T.; Wittman, A.; De Carli, L.; Davidson, D.: An Analysis of First-Party Cookie Exfiltration due to CNAME Redirections. In: Workshop on Measurements, Attacks, and Defenses for the Web. MADWeb, 2021.

[RKW12] Roesner, F.; Kohno, T.; Wetherall, D.: Detecting and Defending Against Third-party Tracking on the Web. In: Conference on Networked Systems Design and Implementation. NSDI, 2012.

[RM88] Ratcliff, J. W.; Metzener, D. E.: Pattern Matching: The Gestalt Approach. Dr Dobbs Journal 13/7, 1988.

[Sa19] Sanchez-Rola, I.; Dell'Amico, M.; Kotzias, P.; Balzarotti, D.; Bilge, L.; Vervier, P.-A.; Santos, I.: Can I Opt Out Yet?: GDPR and the Global Illusion of Cookie Control. In: ACM Symposium on Information, Computer and Communications Security. AsiaCCS, 2019.

[SK19] Sørensen, J. K.; Kosta, S.: Before and After GDPR: The Changes in Third Party Presence at Public and Private European Websites. In: International Conference on World Wide Web. WWW, pp. 1590–1600, 2019.

[TH11] TRUSTe; Harris Interactive: Consumer Research Results—Privacy and Online Behavioral Advertising, https://www.eff.org/files/truste-2011-consumer-behavioral-advertising-survey-results.pdf, 2011.

[Ur20a] Urban, T.; Degeling, M.; Holz, T.; Pohlmann, N.: Beyond the Front Page: Measuring Third Party Dynamics in the Field. In: International Conference on World Wide Web. WWW, 2020.

[Ur20b] Urban, T.; Tatang, D.; Degeling, M.; Holz, T.; Pohlmann, N.: Measuring the Impact of the GDPR on Data Sharing. In: ACM Symposium on Information, Computer and Communications Security. AsiaCCS, 2020.

[Ut19] Utz, C.; Degeling, M.; Fahl, S.; Schaub, F.; Holz, T.: (Un)informed Consent: Studying GDPR Consent Notices in the Field. In: ACM Conference on Computer and Communications Security. CCS, 2019.

[Va18] Vastel, A.; Laperdrix, P.; Rudametkin, W.; Rouvoy, R.: FP-STALKER: Tracking Browser Fingerprint Evolutions. In: IEEE Symposium on Security and Privacy. S&P, 2018.

[Wi20] Wilander, J.: CNAME Cloaking and Bounce Tracking Defense, https://webkit.org/blog/11338/cname-cloaking-and-bounce-tracking-defense/, 2020.

A Background on First-Party Tracking

In contrast to traditional third-party tracking, in which the first party (website) does not necessarily know which trackers might be embedded into the website [Ur20a], in our tracking model, the first party needs to support the tracker actively. The high-level idea is that the first party stores the tracking information in its first-party context and then sends it to a third party. Figure 2 displays the tracking (attack) model that we assume in this work. Our model consist of three parties: (1) the tracker (`tracking.org`), (2) the visited website (`foo.com`), and (3) the users' browser. More specifically, the website places a (static) interactive object (e.g. a JavaScript code) in its first-party context that establishes a connection to a known third-party interface (e.g. by issuing an HTTP GET request). Once the user visits a website, the first-party object checks if an identifier for the user exists in the local first-party context. If it does not exists, a new identifier is set (❶ in Figure 2). This identifier can either be computed based on the device's properties (i. e. a device fingerprint) or be generated by the third-party (❷). The browser will store the identifier in the first-party context of the site since it was set by a first-party object (❸). On consecutive visits the first-party tracking object can read the previously set tracking identifier (❹) and send it to the third party (`tracking.org` – ❺). Utilizing this mechanism, the tracker can bypass defense mechanisms that aim to limit or eliminate third-party cookies.

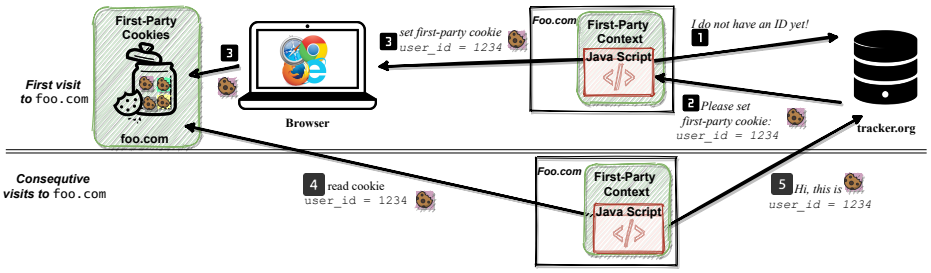

Fig. 2: Overview of first-party tracking methods.

In this work, we call this practice potential "*first-party cookie tracking*" because it works like traditional tracking only that the identifier is stored in the first-party context. This method comes with the drawback that the identifier is not accessible cross-site per-se since it has to be set for each site and might be accessed by other third parties that get access to the storage. Naturally, an identifier that is stored by the first-party cannot be used by a third party to track the user. However, once the identifier is leaked to a third party, it can potentially be used by that third party to track the user across a site. This could be the case for consent management services that assign a (first-party) identifier to a user, and, consequently, track them across the site. In this case, tracking is done with an eligible purpose (consent management), but the third party *could* also abuse it.

Canonical Name Cloaking: One popular way to link a third-party resource with a first-party element is to use so called *CNAME cloaking*. Within the DNS protocol, the *canonical name record* (CNAME) is used to map a domain name to another. This mapping is commonly used to host multiple services (e.g. *news.foo.com*, *ftp.foo.com*, and *mail.foo.com*) on the same IP. To do so, one creates an alias (CNAME) for each service that all refer to the same `DNS A` record of *foo.com*. However, the CNAME can also point to another server. For example, the CNAME of *news.foo.com* could point to *bar.com*. On the Web, this means that the browser will load the content from *bar.com*, and not *foo.com*. CNAME cloaking is commonly used on the Web. However, one can also exploit this to avoid that users block specific content (e.g. ads or tracking) as this is commonly done based on the loaded URL [Me17]. For example, if a user wants to avoid content loaded by *tracker.com* and blocks all requests to the domain on the application level (e.g. by using an ad-blocker), a web service could circumvent that by resolving all requests to *t.foo.com* to the tracking domain – on DNS level.

B Measurement Framework

In this section, we provide further details on our measurement Framework.

Experimental Setup For each visited page, we configured the framework to use the same user agent (`Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:77.0) Gecko/20100101`

`Firefox/77.0`) and desktop resolution (`1366x768`) to ensure that these attributes will not impact the computation of browsers fingerprints. According to Urban et al. [Ur20a], websites tend to use more cookies if the cookie jar and local storage of a browser instance is populated. Therefore, we created a base profile that is loaded by *OpenWPM* upon each page visit. To populate this profile, we successively visited each landing page in our dataset once and used the resulting Firefox profile in our measurement. The base profile is not updated after each visit (i. e. we perform a *stateless* crawl), which, on the one hand, implies that the order of visited websites has no impact on the results and, on the other hand, that trackers have to re-set their identifiers. Hence, this setup allows us to distinguish between user identifiers and session IDs.

We conducted the measurement from a university network within the EU. To perform the measurement, we profit of two machines which in total have 95 GB of RAM, 56 CPUs (*Intel Xeon CPU @ 2.10GHz*), and more than 5 TB of hard disk space. When visiting a website we used a 30 second timeout interval to compensate for e.g. site crashes or slowly or not loading websites. We did not retry a page visit in case it failed in the first attempt. In our crawl, we used simple faked user interactions (i. e. random scrolling and mouse jiggling) because previous work found that this increases the number of observed tracking attempts [Ur20a].

Identifying Third Parties In our analysis, we use the *WhoTracks.Me* database to map domains (eTLD+1) to the organizations running them [Cl18]. This clustering allows us to discuss the impact of a company rather than the impact of a domain because one company can run multiple tracking domains. Previous works have shown that some types of websites tend to use more (third-party) user tracking techniques than others (e.g. “News” websites) [SK19]. Therefore, we analyze if similar observations can be made for first-party tracking techniques as well. We use the *McAfee SmartFilter Internet Database* service to identify the primary purpose (“category”) of a domain [Mc20]. This mapping allows us to understand which categories of sites utilize the technique of interest and gives us a brief impression of who mainly uses it. The *Cookiepedia* [On19] is a crowd sourced index for cookies, their purposes, and “owning” companies. In this work, we use the service to map identified cookies to the respective companies who “own” and use them.

Limitations: Like most large-scale web measurement studies, our work comes with the limitation that our crawler is not logged in to any website, nor do we perform any meaningful interaction with the websites. Hence, we will miss some pages and features of sites that are only triggered after a user interaction (e.g. payment processes) or after successful login (e.g. loading of personal data). Regarding tracking technologies, this means that some sites that wait until the user opts-in, might not set any cookies. Since the literature suggests that these opt-in choices (“cookie banners”) are often not honored [Sa19; Ut19], our results can only be seen as a lower bound to the extend of the problem. Furthermore, our automated crawler might be detected by a page, which might then show different content than for real user visits.

C First-Party Tracking Cookies

Using our approach to classify tracking cookies, we found that, on average, each site uses 6 potential first-party tracking cookies (max: 598; min: 0) and over all profiles 19% of all first-party cookies are used to track users. Overall, we found that 210,721 (20%) of all first-party cookies could be used to track users. In total, we found such tracking cookies on 10,700 (93%) of the analyzed sites. Hence, a vast majority in our dataset of websites offer third parties their first-party storage to place identifiers. We used the name of a cookie as an indicator to distinguish which third-party utilizes the cookie. We observed a long tail distribution that shows that some prominent non-trivial cookie names are used by several websites, which indicates that they all originate from commonly used services or libraries. According to the classification of *Cookiepedia* [On19], the top cookies, of the aforementioned classified cookies, are used by *Google* (`_ga` 28%), *Facebook* (`_fbp`, 14%), and again *Google* (`__gads`, 7% and `_gcl_au` 6%). The results show that the large, known tracking companies are also utilizing first-party tracking tools, in addition to third-party tools. In absolute numbers, popular websites (i. e. sites with a low rank on the *Tranco* list) tend to use more first-party cookies and also utilize more of them to track users. In our experiment, popular websites (rank $\leq 3,000$) use on average 45 cookies (SD: 65) of which 21% (SD: 21) are used to track users while less popular websites (rank $> 3,000$) use, on average, 34 (SD: 44) cookies and 20% (SD: 61) of them were classified as trackers. We chose the top 3,000 sites because less popular websites show a different behavior, number wise (see the standard deviations). The X^2 test ($\alpha = 0.5$) also found a strong correlation between the absolute number of present trackers and the rank of the website (p-value < 0.001). Hence, such cookies are more likely to appear on popular sites. However, if we look at the relative number of first-party tracking cookies in relation to other cookies, we find that popular websites use fewer trackers than other sites.

Defense Mechanisms One simple defense mechanism to avoid third-party cookies all together is to tell the browser never to store them. While third-party cookies are seemingly unrelated to first-party cookie sit is still interesting to analyses whether websites switch their tracking techniques if the preferred one is actively blocked (i. e. do they store identifier in the first-party context if the third-party context is disabled). In our measurement, this applies to profiles two and four (i. e. we used Firefox's option to "never" store third-party cookies). It is worth noting that the browser accepts and sets a third-party cookie but deletes it right away. Hence, the script or request that sets the cookie can do this 'normally' but cannot access the cookie afterward. Overall, we observed a change of roughly 13% in the use of first-party cookies if third-party cookies are disabled. This decrease can probably be attributed to the slight reduction of visited websites and the usual noise of large-scale Web measurements. Hence, blocking third-party cookies has a negligible effect on the use of first-party tracking cookies, which means that trackers do not use this way to evade the deletion of third-party cookies.