The Devil is in the Details: Detection, Measurement and Lawfulness of Server-Side Tracking on the Web

Imane Fouad, Cristiana Santos, Pierre Laperdrix

The Devil is in the Details: Detection, Measurement and Lawfulness of Server-Side Tracking on the Web

Imane Fouad
Univ. Lille, Inria, CNRS, UMR CRIStAL
imane.fouad@inria.fr

Cristiana Santos
Utrecht University
c.teixeirasantos@uu.nl

Pierre Laperdrix
Univ. Lille, Inria, CNRS, UMR CRIStAL
pierre.laperdrix@univ-lille.fr

ABSTRACT

As online privacy is cementing itself as one of the core pillars of the Internet, major changes are happening across many industries. On the technological side, users are pushing for more privacy-preserving technologies and rely on browsers and extensions that limit online tracking as much as possible. On the legal front, regulations like GDPR and the ePrivacy Directive in Europe have forced companies to change their practices and be more transparent about how they handle user data. For the ad industry, the end of third-party cookies planned for 2025 is having severe ramifications as the main source of data on which this industry is built on will be gone. In this tumultuous context, companies have come up with innovative ways to overcome current and future restrictions. A novel technique which has not received much attention called Server-side tracking (SST) moves its tracking logic away from the user's device onto an external server. In this work, our aim is to detect SST on the web and understand its lawfulness with respect to current legislation. We developed a methodology that relies on crawls spaced 2 years apart performed before and after the introduction of SST to identify trackers that moved behind SST domains and that are now hidden from view. Our results show that 389, out of 7,367 visited websites, track users behind a cloaked domain and that 28 websites perform Server-side tracking in a first-party capacity. We demonstrate that such a tracking technique can overcome the Same-Origin Policy and introduce security vulnerabilities. Together with a legal scholar, we also show that SST entails non-compliant practices and infringes the GDPR and the ePrivacy Directive.

KEYWORDS

Server side tracking, CNAME tracking, GDPR, ePD

1 INTRODUCTION

Digital advertising is the lifeblood of the Internet. In 2021, $521 billion were spent on digital ads [9], and the revenues that come from it fund the smallest sites to the biggest Internet companies. While Google, Meta, and Amazon account for about two-thirds of these revenues [8], the rest is distributed to smaller actors which, in turn, helps keep a very large part of the Internet free. To make it all work, the entire ad ecosystem relies on a wealth of tracking technologies that keep evolving. Over the past decade, many studies have looked at how users can be tracked on the Internet with cookies [40, 76], tracking pixels [48], browser fingerprinting [20, 65] and IP addresses [71]. As the Internet is cementing its central role in our everyday lives, there is currently a big push to make online privacy a priority. On the legal side, several efforts have been made like the GDPR and the ePrivacy Directive in Europe, the CCPA in California, the LGPD in Brazil, or the PDP Bill in India. They all aim to push companies to better protect user data and be transparent in how they handle and share it. On the technological side, a major shift is happening with the end of third-party cookies. Initially announced by Google for 2022 [64], this change has faced multiple delays [15, 92] and is now scheduled for 2025 [55]. As these cookies are the primary vectors for tracking on the open Web, their disappearance will have a tremendous impact on the Web economy. In this fast-evolving environment, some companies are already testing alternatives to prepare themselves for this change through techniques like contextual advertising [28], identity graphs [60, 61], or interest-based advertising [63, 69, 93]. But others are trying to be more sneaky by performing the same type of tracking as they did before and hiding it as much as possible through other means. A novel approach called ***Server-side tracking (SST)*** has been growing in popularity since its introduction by Google in 2020 but has yet to be studied by the scientific community. Instead of the client directly reporting its tracking data to different companies around the world, it is the role of the SST server to do it. This approach is currently hampering online privacy as it transforms what used to be very clear tracking requests performed by the browser on the client side into masked and hidden ones on the server side. For users, it becomes complicated to protect themselves as their favorite blocking tools are not adapted to this hidden form of tracking. For regulators, it is simply a nightmare as they cannot see clearly which companies are performing tracking in the background.

With this study, we strive to raise awareness about how covert SST can be, and motivate the scientific community to actively engage in the efforts to tackle SST and forestall potential issues in the future. Because the goal of SST is to hide trackers on the server side, known methods to identify more traditional tracking are not well suited to detect SST. The reception of a user ID does not necessarily imply that the server will share it directly with different partners. In order for us to understand if SST on the Web is used at all and by whom, we devised a unique methodology that relies on crawls made before and after the introduction of SST. This temporality is key as it gives us an insight into the websites that transitioned from regular tracking to SST, enabling their identification. We also shine a light on the legality of this practice to understand how it fits into the current legal rules of the GDPR and ePrivacy Directive. Our contributions in this paper are as follows:

Proceedings on Privacy Enhancing Technologies YYYY(X), 1–16

https://doi.org/XXXXXXX.XXXXXXX

- **We provide a custom methodology to detect SST.** In this study, we designed a detection methodology that helps detect Server-side tracking performed in a first-party capacity. We show that Server-side tracking hidden behind first-party subdomains is starting to be adopted, and we highlight that such practice can introduce security and privacy vulnerabilities. We detected that 28 websites include an SST subdomain.
- **We assess the legal compliance of SST.** We show that i) while consent is needed to collect user data in SST, it was absent on the websites we visited rendering their processing illegal; ii) it is difficult with SST to discern the purposes of cookies, which makes it impossible for an auditor to determine what is the appropriate legal basis and whether the processing is compliant or not with the GDPR and ePD; iii) regulators and auditors have a complex task of auditing legal compliance of websites when SST domains appear as first party and set first-party cookies. They might consider such first-party cookie as "strictly necessary" and thus exempted from consent, and would not reasonably expect such hidden tracking; iv) As SST subdomains are likely sharing data with other third parties on the server side without consent, the involved SST subdomain can potentially participate in an unlawful data sharing scheme.
- **We perform an in-depth technical and legal analysis on two websites.** We detected that SST is performed on a major telecommunication website and a health website handling sensitive user information, which would require additional protection for users. After analyzing their privacy policy, we concluded that these two websites do provide information regarding the data collected and included trackers. However, they do not mention the use of SST in their privacy policy which triggers a lack of transparency and non-compliance issues: the studied websites failed to inform users about their tracking behaviors and they do not get their mandatory consent.

2 BACKGROUND

2.1 Web technologies and terminology

DNS resolution. When a browser is asked to open a webpage, it first performs a DNS resolution of the *domain* of the page to know where to fetch it. For example, if a browser wants to open *site.com*, it will ask a DNS server for an IP address of a server that can serve the homepage of *site.com*. After contacting the right server for this page, the browser will then parse it and interpret its content so that it is rendered for the user to see. Because the Web is rich and dynamic, a modern webpage is full of resources from images to scripts that need to be fetched separately. A DNS resolution is performed for each domain encountered on a webpage.

CNAME cloaking. When resolving a domain name, a DNS server often returns an "A" record which contains an IP address to contact. The server can also return a "CNAME" record which points to a different domain name. This CNAME mechanism gained a lot of popularity recently because it can be abused to pass third-party content as first-party. While several studies have looked at the use of CNAME redirection for tracking [14, 31, 34, 79], our paper looks at IP cloaking more broadly and uses CNAME as one record among other DNS records to detect SST.

First-party vs third-party. Depending on where a resource is hosted, we can differentiate first and third-party resources. First-party resources are fetched from the same domain (e.g. *site.com*) or a subdomain (e.g. *images.site.com*) of the visited site. Third-party resources are fetched from a different domain than the visited one. This distinction is important for this study as the browser handles these resources and the security around them differently.

2.2 Web tracking

There are essentially two types of cookies: first-party and third-party cookies. They both function the same way but they differ in how they are created and used.

Same-site tracking. If a cookie is stored with a first-party resource as host, it enables *Same-Site* tracking as the domain of the cookie will refer to the website the user is visiting. First-party cookies are practical for tracking repeat visits to the same site. Such cookies can also help collect analytics and understand user engagement. They can be set directly by the visited website, or by a third-party service running in the context of the visited website. First-party cookies are commonly known to be deployed for analytics purposes. However, in the last few years, trackers are increasingly relying on first-party cookies for tracking purposes [25, 33, 82].

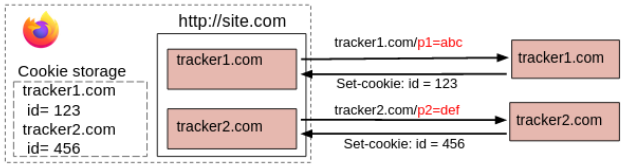
Cross-site tracking If a cookie is stored with a third-party resource as host, it enables *Cross-Site* tracking as the domain and path of the cookie will refer to an external domain that the user may not be aware of. Third-party cookies are used to track the browsing activity of a user on all the sites where this resource is present. At the time of writing, it is one of the main driving forces behind the ad ecosystem even though its days are counted due to the planned deprecation of third-party cookies in 2025 [55].

Distinguishing first-party cookies from third-party ones is straightforward and it enables users to understand the role of the cookies that are present in their browsers. In turn, this helps setup protection strategies for users who wish not to be tracked. Brave and Safari disabled third-party cookies by default. Firefox has a feature called Total Cookie Protection that creates a sandbox around each of these cookies so they cannot be used for cross-site tracking. Chrome can be configured to disable third-party cookies. This clear separation between first and third parties helps browser vendors to better protect users online. However, as can be seen below, new techniques that are gaining in popularity are undermining this security by blurring the line between the two.

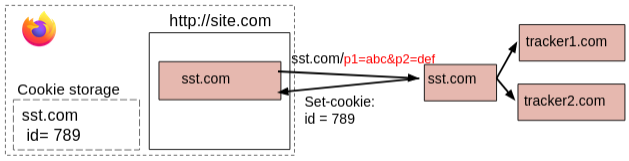
3 THREAT MODEL

SST has emerged as a significant paradigm shift from the traditional client-side tracking approach. While the latter has been widely practiced, the former offers a less detectable and more covert means of conducting tracking operations. This section presents a detailed threat model analysis of SST, with a focus on its potential risks and vulnerabilities. SST can be conducted through a third-party service or operated via an included first-party subdomain. In our analysis, we identify two key actors involved in the SST ecosystem:

- **SST Domain:** This actor refers to the domain where the SST service is hosted and operated. It serves as the centralized



(a) Client-side tracking: Direct inclusion of third party services



(b) Server-side tracking: Tracking moved to the server side

Figure 1: Overview of the SST. *(a) Before SST, the website* `site.com` *includes 2 trackers: tracker1.com and tracker2.com. A request is sent to each of the trackers to fetch the content, as part of the URL the trackers respectively receive the parameters p1=abc, and p2=def. (b) After SST, tracker1.com and tracker2.com are moved to the server side behind the SST subdomain* `sst.com`*. Both parameters p1 and p2 are sent to* `sst.com` *as part of the URL.*

location for handling tracking operations shifted from the client side to the server side.

- **Trackers:** These components collect user data and monitor online behavior. We assume that a domain is a tracker if it has the technical ability to track user activity regardless of the domain owner's intention. With the adoption of SST, the transmission of data to third parties is moved from the user's device to the server side.

In this section, we first outline the functioning of traditional client-side tracking. Next, we introduce SST, covering its implementation with default third-party services and its utilization through cloaked third-party domains.

3.1 Client-side vs Server-side tracking.

Client-side tracking. Today, web tracking is commonly known to be deployed on the client side (eg. on the end-user device). When visiting a website, the browser will load third-party resources and store third-party cookies for each of these requests. Figure 1(a) presents client-side tracking. It provides an example where the user visits `site.com` which includes content from `tracker1.com`, and `tracker2.com`. The browser performs requests to both these servers and stores a unique cookie for each of them.

Server-side tracking. Google introduced SST in 2020 [21]. As the name implies, the goal is to perform the tracking on the server side. The browser makes a single request to the SST server and reports back the tracking data to this server only. It is then the SST server's role to dispatch the tracking data to the right third party in the right format. In Figure 1(b), `tracker1.com` and `tracker2.com` are located behind the `sst.com` domain where SST is deployed. The information received by the SST server can take multiple forms: unique identifiers for each of the tracking servers, analytics data, or even old identifiers that have been synchronized to be sent to this new SST domain.

As described by Google in [21], the advantages of SST are twofold: a gain in performance for the client as only a single tracking server is contacted, and a gain in security as the developers have control over what is actually being collected on their website. However, the main problem with SST is the lack of transparency as it hides on the server side the different third parties that are contacted and what is being sent to them.

3.2 First-party SST.

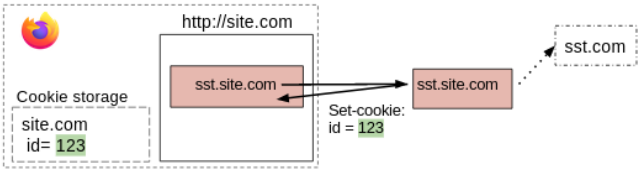


Figure 2: First-party SST. *The visited website* `site.com` *includes a first-party subdomain* `sst.site.com`*. This subdomain has an IP record that points to* `sst.com`*. The resulting cookie set by* `sst.com` *in that example appears as a first-party cookie.*

Our paper utilizes IP address comparisons as the initial step in our detection methodology to identify SST servers. By analyzing the IP addresses associated with all the domains contacted by the browser, we can effectively differentiate domains related to the website the user is visiting from the ones that are not, including those hiding behind first-party subdomains. We rely on A/AAAA/CNAME DNS records to make that distinction. In other words, some trackers use a cloaking technique that hides tracking activities under an alias domain so that they appear as if the tracking originates from the visited website itself. This way, a hidden tracker can present its own cookies as if they were first-party ones but they belong to a different organization. As can be seen in Figure 2, the `sst.com` tracking server is hiding behind the `sst.site.com` first-party subdomain and the cookies set by the tracker in that case will appear as first-party cookies. As a result of this IP address-based detection, SST servers can be exposed, shedding light on mechanisms that blur the distinction between same-site and cross-site tracking.

4 METHODOLOGY

In this section, we describe an overview of the goal of our study, the data we collected, and the steps we followed in our methodology to identify potential SST servers.

4.1 Overview

This study aims to detect SST as defined in Section 3. We consider a server or domain to participate in users' tracking if it has the technical ability to perform the tracking independently of their intention, and whether the domain is using or not the received information such as a cookie with an identifier. Figure 3 summarizes our detection methodology. In total, we performed 3 crawls from Europe: one crawl called *Pre-SST crawl* was conducted in March 2020 and predates the introduction of SST by Google. The other

two called *Post-SST crawl* and *User specific crawl* were performed in May 2022.

Step n°1: IP cloaking. Our first step is to identify the domains using IP cloaking. As explained by Google in their tutorial on SST [85], "one of the key features of server-side tagging is that it can be run in a subdomain of the websites that send data to it". Cookies then become first-party cookies which will "greatly improve the quality of [the] data collection" according to Google, as they are not facing the restrictions imposed by browsers on third-party cookies. Identifying DNS records helps us here detect third-party servers that hide behind a subdomain of the visited website.

Step n°2: Tracking. Our second step consists of identifying servers that receive information that could be used to uniquely identify users. We performed two crawls in parallel in May 2022 to assess for each request if they contain data that is akin to an identifier.

Step n°3: Server-side tracking. The goal of the final step is to identify domains that aggregate in a single request identifiers that used to be sent to two or multiple third parties. To that end, we rely on one crawl called "Pre-SST" made before the deployment of SST servers, and another called "Post-SST" made after. The temporal gap of almost two years we had between the two crawls is important so that we can detect websites that have shifted their methods of tracking from multiple third parties to a single SST server.

It is important to emphasize that our paper primarily centers around first-party SST. Nonetheless, it should be noted that domains still retain the option to employ the default third-party domain for SST services until third-party cookies are phased out.

4.2 Detection of cloaked domains.

Typically, a website will include a Server-side tracking service with the default setting, which will then generate HTTP(S) traffic by sending a third-party request to the SST domain. Such behavior can be easily detected and blocked. Therefore, SST services recommend using a custom DNS configuration so that a first-party subdomain can hide the tracking server behind it. For instance, the website `site.com` would include the SST service under the first-party subdomain `sst.site.com`. The request to `sst.site.com` will be then redirected to the SST provider.

We consider that a first-party subdomain is a potential SST subdomain if it points to a different organization than the one from the visited website. We refer to such subdomains as *cloaked domains*.

We extracted the DNS records of all first-party subdomains on all visited websites, and then checked the registered organization behind the IPs of both A/AAAA and CNAME records using the whois library [95]. Similarly, we extracted the registered organization of the visited website. We filter out subdomains that are registered with the same organization as the visited website to ensure that only cloaked domains are further analyzed.

4.3 Detection of ID sharing.

A domain can set a cookie either via an HTTP(S) request or systematically via JavaScript. A cookie is defined by the triplet (host, key, value), where the host refers to the domain that has access to the cookie. Next, when the browser sends a request to the same domain or its subdomains, the browser will automatically attach the cookies in the *Cookie* header of the outgoing HTTP request. We study the cookies stored both via HTTP(S) and JavaScript, and we consider that a domain is accessing the cookie whether it is setting it or receiving it either through HTTP(S) or JavaScript. To detect cookies potentially used to identify users, we used two machines that appear as different users, as done by previous works [4, 41, 42, 49, 50, 52]. We followed the ID detection algorithm designed by Englehardt et al. [40]. Following this approach, we executed the following steps.

- Cookies are typically set in a "key = value" format, where the "value" could be a single value or structured in a subkey, subvalue format as follows:

  $$subkey_1 = subvalue_1 \& ... \& subkey_n = subvalue_n$$

  Therefore, we consider first both cases and parse the cookie values according to the structured format whenever possible using as delimiters any character not in [a-zA-Z0-9,'-','_',','].
- Next, we eliminate subvalues with a length smaller than 8.
- We compare the cookies that appear in both Post-SST, and *User specific crawl* with the same host and key, and we eliminate subvalues that have more than 66% similarity across the two crawls according to the Ratcliff-Obershelp algorithm [77]. Identifiers should be user-specific, enabling user identification. Therefore, we exclude cookies that do not exhibit diversity across the two crawls. We further exclude cookies that do not reappear on the *User specific crawl* given that we do not have proof that such cookies are user-specific.
- We do not impose any restrictions on the cookie lifetime. Previous works [4, 41] filtered out cookies that expire less than a month after being placed in the browser. In our study, we removed this limit because domains can continuously update cookies with a short lifetime and do the mapping of these cookies on the server side which will allow long-term tracking as shown by previous works [52]. We conducted an analysis where we present our findings on the studied cookies' lifetime in Appendix A.

We consider that a cookie (host, key, value) is an identifier if the cookie value or at least one of its subvalues satisfies the specified criteria. In summary, we consider that a cloaked domain is performing tracking if it is additionally sharing at least one ID value. We refer to such domains as *Cloaked trackers*.

4.4 Detection of SST

The public beta of the Server-side Google Tag Manager was launched in August 2020. Before, SST was not proposed by major web companies and it was not known to be deployed. Therefore, we believe that subdomains serving SST will appear on recent crawls but not on crawls performed before August 2020. The first step of our detection of SST consists in detecting subdomains added to the websites after the SST service was launched. To build the set of SST subdomains, we performed two crawls: the first crawl called *Pre-SST crawl* was performed in March 2020 before the appearance of the Server-side Google Tag Manager, and the second called *Post-SST crawl* in May 2022 after the launch of the SST service. For each visited website, we extracted the set of subdomains appearing in the *Post-SST crawl* but not included in the *Pre-SST crawl*. In the following, we will refer to these subdomains as *emerging subdomains*.

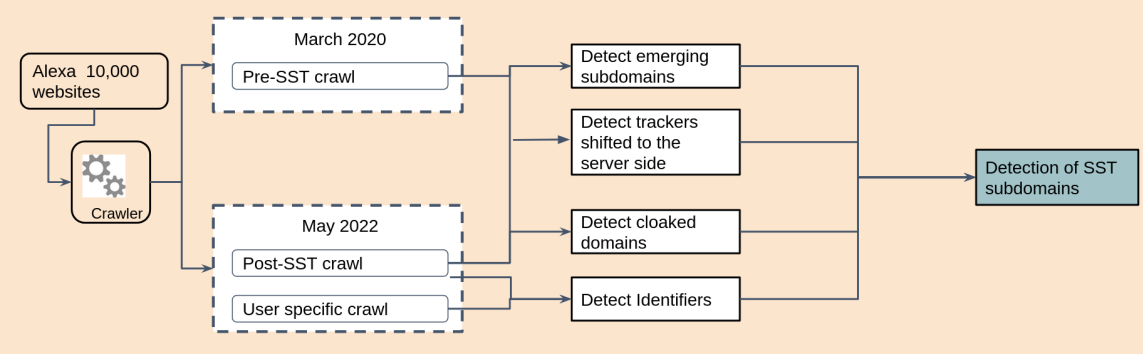


Figure 3: Methodology overview: Detection of SST on Alexa top 10,000 websites. *For each website, we performed a Pre-SST crawl on March 2020 , we later perform a Post-SST crawl, and User specific crawl on May 2022. Using the Post-SST crawl, we detect emerging subdomains that are included in the websites Post-SST, next, we detect the cloaked domains that are hosted by a third-party organization. We then extract the set of IDs using the Post-SST crawl, and User specific crawl. We consider that a subdomain is a cloaked tracker if it is (1) a cloaked domain, and (2) is sharing at least one ID. We further analyze the cloaked trackers, and we classify them as SST subdomains if they are additionally exchanging with trackers removed from the website, and shifted to the server side. All these steps are discussed in Section 4.*

Typically, SST helps to reduce the traffic on a website. This strategy involves the migration of trackers initially integrated into a website to the server side, as outlined in Figure 1

On a given website where we detect cloaked trackers, we consider that the corresponding domain is performing server-side tracking and that trackers are hidden behind an SST domain `sst.site.com` if the SST domain is receiving parameters and/or cookies previously sent to at least two trackers, `tracker1.com` and `tracker2.com`, that are no longer appearing on the website. In such case, we suppose that `tracker1.com` and `tracker2.com` moved to the server side behind the SST domain `sst.site.com`.

Summary. We consider that a domain is performing SST if (1) it is an emerging subdomain, (2) it is registered with a distinct organization compared to the visited website and it uses cloaked tracking to hide trackers behind it, and (3) it contains tracking data that used to be sent to different domains. We will refer to these subdomains as *SST domains*.

4.5 Limitations

Identifying SST is complex because it operates on the server side and not in the user's browser where requests can be analyzed. This makes it tricky to distinguish domains engaged in SST from regular tracking ones since there is no definitive signal to differentiate them. In this study, we designed a unique method that aims to identify websites that transitioned from regular tracking to SST. By utilizing data from crawls made before and after the introduction of SST, we are in a unique position to detect such a shift and see which trackers were moved to an SST server. The main limitation is that we only detect a subset of the servers partaking in SST. Websites that decided to use SST and change their entire tracking pipeline with seemingly no direct link with their previous tracking architecture would not be detected by our approach. Moreover, the switch from third-party tracking to first-party SST may have also changed the structure of the requests and their content in a way that our methodology cannot detect the shift. Despite this limitation, our paper aims to shed light on the elusive nature of SST and encourage the scientific community to join efforts in addressing this issue because this technique has the potential to create a lot of damage with its clear lack of transparency.

Additionally, we consider a domain to be an SST one if it has the technical ability to perform the tracking on the server side independently of their intention, and whether the domain is using or not the received information. This information leakage is still a privacy concern that could be exploited by the domain anytime.

Finally, another limitation is that stateless crawling might have a small influence on the website behavior. Zeber et al. [96] showed that stateless crawls of websites include more requests to third-party trackers compared to stateful crawls. This is expected as in stateless crawls we do clean the browser storage between every website visit. As a result, the third-party service will recreate the content on the user's browser at every visit, and thus we will have more interactions with the third-party services. For our study, this is not a problem as we focus our work on first-party-based SST within visited websites.

5 SST IN THE WILD

In this section, we present the results of our crawls on 10,000 visited websites. First, we describe our measurement setup. Then, we present the prevalence of tracking under cloaked domains and its security implications. Finally, we analyze the prevalence of SST and the data sent to the SST domains.

5.1 Measurement setup

We used the OpenWPM platform [75] with the Firefox browser to perform three passive web measurement crawls of the Alexa top 10,000 websites [13]: *Pre-SST crawl* in March 2020, and the other two crawls *Post-SST crawl* and *User specific crawl* in May 2022 (see Appendix C for more details). We used two distinct machines with different characteristics so that they appear as different users, as done by previous works [4, 41, 42, 50]. We then excluded all the websites that were not successfully visited in at least one of the

	Domains	Websites
First-party	14,731	6,725
Emerging	7,522	4,302
Cloaked	996	767
Cloaked trackers	474	389
SST	**32**	**28**

Table 1: Prevalence of the SST: Following the methodology described in Figure 3, we found that 28 websites are deploying SST.

performed crawls and ended up with a total list of 7,367 visited websites. The goal of our work is to study the presence of SST on each of the visited websites in isolation. Therefore, for each crawl, we used *stateless crawling instances*. We defined a stateless crawling instance of a website X as follows: (1) we visit the home page of the website X and keep the page open until all content is loaded to capture all cookies stored (we set the timeout for loading the page to 90s), (2) we clear the profile by removing the Firefox profile directory. The rationale behind the stateless crawling is to capture tracking behaviors on all websites from a fresh user profile. Starting without any cookies or cached resources forces the browser to contact all trackers again and recreate all cookies, thus allowing the detection of tracking practices on each of these sites. In all our crawls, we did not interact with cookie banners on the visited websites, that is we never accepted nor rejected cookies during our visit. Therefore, all detected tracking behaviors in this work are performed prior to the user's consent. For each crawling instance, we extracted the following information: HTTP(S) requests, HTTP(S) responses, cookies, and script calls collected by OpenWPM. Table 1 presents an overview of the prevalence of first-party SST on the 7,367 successfully visited websites.

5.2 Cloaked first-party domains

To detect potential SST domains, we focus our study on first-party subdomains included in the website after the introduction of server side-tracking which are additionally hidden behind a third-party organization. We refer to such domains as Cloaked domains. Using DNS records, we collected the organizations located behind the first-party emerging subdomains. We then extracted the set of subdomains belonging to an organization that is different from the one associated with the visited websites. Table 2 summarizes the prevalence of first-party subdomains, emerging, and cloaked domains detected on the 7,367 visited websites.

Alexa rank	**0-100**	**100-1000**	**1,000-10,000**
First-party	97.78%	93.88%	90.87%
Emerging	62.22%	50.31%	59.37%
Cloaked	8.89%	9.06%	10.6%

Table 2: Percentage of websites including first-party, emerging, and cloaked domains. Cloaked domains are uniformly distributed across diverse website ranks.

Prevalence of first-party domains. We categorize a domain as a first-party domain if it meets two conditions: (1) it shares the same 2nd-level top-level domain (TLD) as the visited website, and (2) its domain name differs from that of the visited website. For instance, when visiting *site.com*, *sst.site.com* is considered as a first-party subdomain. Among the 7,367 websites we examined, we observed that 6,725 (91.28%) of them incorporate at least one first-party subdomain. Our analysis led us to identify a total of 14,731 first-party domains, with an average of 2 subdomains per website. These first-party subdomains demonstrate a consistent distribution across various websites, regardless of their ranking, with a slight increase seen among the top 100 websites.

Prevalence of emerging domains. We focus our study on the subset of identified first-party domains that have appeared after the introduction of SST. Specifically, we detect domains appearing in the *Post-SST crawl* dataset but absent in the *Pre-SST crawl* dataset. We found that out of 14,731 first-party subdomains, 7,522 (51.06%) are additionally *emerging subdomains*. These domains appear on 4,302 distinct websites (58.39% of the visited websites). Next, we will focus our study on this set of domains.

Prevalence of cloaked domains. We analyzed a set of 7,522 emerging first-party subdomains and extracted the corresponding organizations. We successfully identified the organization behind 6,273 (83.39%) of the first-party subdomains. we found that 996 first-party subdomains belong to a different organization compared to the visited website. We refer to these subdomains as *cloaked* domains. We detect that *cloaked domains* appear on 767 websites (10.41% of the visited websites).

Summary. We applied the cloaked detection methodology described in Section 4 to the studied 7,367 websites. We detected 767 websites that include at least one cloaked domain. These websites represent 10.41% of the visited websites.

5.3 Tracking under cloaked domains

We consider that a domain is a tracking cloaked domain if it is a third-party service hidden behind a *cloaked first-party domain*, and is additionally receiving or setting at least one ID cookie either through HTTP(S), or via JavaScript. In this study, we classify a domain as a tracker based on its technical capability to engage in tracking, irrespective of its intention or stated purpose. The configuration enabling the receipt of identifier cookies empowers domains to track user activities. It is important to note that, in this study, we did not provide consent for any cookies. Consequently, such practices could potentially intrude upon user privacy.

Tracking under cloaked domains. Using the *User specific crawl*, we detected ID cookies that are set or sent to the potential SSTs. We collected a total of 133,874 cookies on the 7,367 visited websites in the *Post-SST crawl*. We analyzed the 133,874 cookies, and we extracted the set of cookies containing IDs using the two crawls *User specific crawl*, and *Post-SST crawl* as described in Section 4.3. In total, we detected 38,313 cookies with IDs in the *Post-SST crawl*. In the following, we only consider this set of 38,313 ID cookies.

We studied the sharing of the detected ID cookies with the studied 996 cloaked domains, and we detected that 474 (47.59%) domains are either receiving or setting an identifier on the user's browser. We refer to this set of domains as tracking cloaked domains. We found that these trackers appear on 397 websites (5.28% of the visited websites), and we focus our analysis on them in the following.

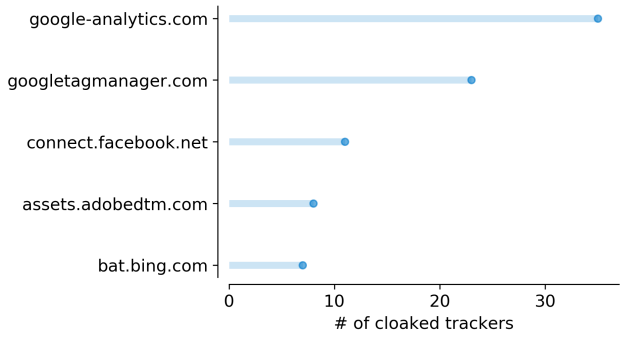


Figure 4: Top third-party domains setting cookies sent to cloaked trackers with a different origin.

In total, we detect that the 474 cloaked trackers are operated by 44 organizations. We identified Akamai as the primary entity responsible for these tracking mechanisms [11]. Akamai self-identifies as a prominent provider of cloud computing, security, and content delivery services. Followed by Amazon, and Google comes in the third rank. In the following, we analyze the IDs shared with the detected 474 cloaked trackers.

Security implications. To improve security and prevent abuse on the web, the Same-Origin Policy (SOP) [81] ensures that access to a cookie is limited to the domains with the same origin as the owner of the cookie. However, cloaked domains are not affected by this policy and can bypass it. When configured as a first-party subdomain, a third-party service hiding behind a cloaked tracker provider can access the cookies set by domains on the first-party website origin. We analyzed the cookies actively accessed by the cloaked trackers through JavaScript or received through HTTP(S), and set by a distinct domain.

We found that 119 cloaked trackers receive at least one ID cookie set by a different third-party domain. These domains are gaining access to cookies set by domains with a different origin, a scenario that contravenes the same origin policy, which would typically restrict such access. We detect a total of 91 distinct third-party domains setting cookies accessed by these cloaked trackers. Figure 4 presents the top 5 domains setting cookies that are accessed by a cloaked tracker. We detect that the top domain is `google-analytics.com` which uses cookies to distinguish users for analytics purposes. As reported by Englehardt and Narayanan [40], Google Analytics was present in more than 60% of the top 1-Million websites in 2016. If these websites decide in the coming months to transition to cloaked trackers to hide their Google Analytics activity, these cloaked trackers will not only be able to collect standard Google Analytics data but they will also be able to collect data from cookies set by other domains in the same first-party context. Contrary to the old ways of performing tracking, SST gives even more power to third parties like Google as the bypass of the Same-Origin Policy enables it to collect even more data than ever before by acting as a relay for other actors in the ad industry. In the end, even if third-party cookies are being phased out in 2025, a transition towards user tracking in a first-party context brings a lot of new problems that this phasing out wants to fix.

Safe cloaked domains. Among the 996 cloaked domains, we demonstrated that 474 are associated with an identifier, while the remaining 522 domains do not employ tracking techniques, meaning they don't send or receive any identifier cookies. We refer to these domains as "safe cloaked domains." Such domains can be used on the web for functionality purposes, like serving as CDNs, for example. We examined the 522 safe cloaked domains. First, we checked the names of their subdomains. We found that the most common top subdomains associated with these domains are "img", "cdn", and "api". They respectively appear in 4.6%, 4.79%, and 4.02% of the safe cloaked domains, indicating that they are predominantly used for CDN services.

Summary. We show that 47.59% of the cloaked domains receive at least one ID cookie. We also show that cookies shared with 119 cloaked trackers are set by a distinct third-party tracker and that such behavior breaches the Same-Origin Policy, and introduces serious security and privacy concerns.

5.4 Detection of first-party SST

We define an *SST domain* as a cloaked tracker that is additionally exchanging with trackers on the server side (see Section 4). In the following, we will analyze the set of 474 detected cloaked trackers.

Prevalence of first-party SST. When implementing SST, the trackers are shifted from the client to the server side, residing behind the SST domain. This shift creates transparency challenges because it becomes impossible to identify the presence of these third-party services through the traditional request/response method, as there are no more direct requests to the tracker. To address this concern, we conduct an in-depth analysis of tracker behavior. This involves a comprehensive examination of the parameters utilized in both the URL and POST data, as well as a thorough assessment of the cookies established by these trackers. Our primary objective is to ascertain whether a domain is actively interacting with two or more trackers on the server side, which would make it an SST domain. This determination is made by evaluating whether the domain exhibits behavior consistent with that of the hidden trackers themselves. Furthermore, we strengthen our analysis by considering the predefined filters—namely, emerging, cloaked, and tracking (as elaborated in Section 4)—to provide a comprehensive assessment. When the observed behavior aligns with these established criteria, we classify the domain as an SST domain exchanging with the corresponding trackers on the server side. We considered the 389 websites where we detected tracking cloaked domains, and looked at the shifting of URL, POST parameters, and cookies to them. To detect this shifting, we extracted all the parameter names and cookies that are present in both the *Pre-SST crawl* and the *Post-SST crawl*. Then, we identified the ones that were sent to a domain present in the *Pre-SST crawl* that has shifted to a different domain in the *Post-SST crawl* with the old domain disappearing between the two crawls. We qualify as SST domains the ones that are the recipients of this shifting of parameters and cookies. The intuition behind this selection is that the adoption of SST on a website would prompt a developer to shift existing trackers behind the SST server so that the user's browser does not send identifiers directly to trackers but to a single intermediary server which is the SST one.

Trackers shifted to Server side	Prevalence	Tracker category	# of SST websites	Disconnect	Top SST organization
doubleclick.net	65.07%	Cross-site	21	Blocked	Google
google.com	61.84%	Cross-site	15	Blocked	Google
google.fr	47.81%	Cross-site	17	Blocked	Google
facebook.net	29.65%	Same-site	13	Blocked	Google
google-analytics.com	69.54%	Same-site	11	Blocked	Google

Table 3: Top 5 trackers shifted to the server side. Prevalence presents the percentage of websites in *Post-SST crawl* including the tracker. SST websites present the number of websites where the tracker is moved to the server side, Disconnect indicates whether the tracker is detected by the Disconnect filter list, and the top SST organization is the top Server side tracking organization operating with the tracker.

We detect 32 SST domains sharing URL parameters names and cookies with trackers no longer appearing on the website. We found that the 32 SST domains are managed by 7 organizations. The top organization behind the SST domains is Google followed by Amazon, and Akamai in the third rank. In fact, Amazon and Google alone are responsible for 81.25% of SST domains. We detected a total of 69 unique tracking subdomains from 59 distinct 2nd-level TLD domains shifted to the server side behind at least one of the 32 SST domains. Table 3 presents the top 5 trackers that moved to the server side behind the SST domains. We found that the top tracker was removed from the *Post-SST crawl*, and shifting parameters keys and cookies to the SST subdomains is `doubleclick.net`. We detect that 21 SST domains are reusing `doubleclick.net` behavior in *Pre-SST crawl* on the same website in *Post-SST crawl*. Notably, Google, being the predominant SST entity, hosts `doubleclick.net` on the server side across 18 websites.

Within Table 3, we provide the prevalence of trackers that have shifted to the server side indicating their representation within the studied dataset of 7,367 websites. The top tracker *doubleclick.net* has a prevalence of 65.07%. The significant prevalence of these trackers across websites underscores the potential magnitude of impact should they be universally migrated to the server side.

Trackers classification. We categorize trackers that have migrated to the server side based on their behavioral attributes, distinguishing them as Same-site and Cross-site trackers. The definitions for both Same-site and cross-site behaviors are detailed in Section 2. We designate a domain as engaging in cross-site tracking if, during interactions with a visited website, at least one of its subdomains initiates the setting or receipt of third-party identifying cookies. Furthermore, we classify a domain as being a Same-site tracker according to two criteria: (1) it is never performing cross-site tracking within our dataset of 7,367 visited websites, and (2) it is performing the behavior characteristic of an analytics entity, as elaborated in section 2.2. Out of the top 5 trackers shifted to the server side, 3 are classified as Cross-site trackers.

In total, we detected 32 cross-site trackers shifted to the server side. Shifting cross-site tracking to the first-party server side adds complexity. Cookies set by the first-party SSTs are considered first-party cookies, limiting immediate cross-site tracking unless synced with third-party cookies or combined with methods like fingerprinting, which brings more complexity to the cross-site tracking. Paradoxically, this complexity can enhance privacy. It creates obstacles for cross-site tracking while acting as a safeguard against extensive tracking. We perform a preliminary analysis on the data shared with the SST domains that can potentially be used for fingerprinting in Section 5.5. This aspect deserves more investigation, which could be detailed in future works, to grasp its impact on user privacy and cross-site tracking practices.

Did the protection mechanisms detect the shifted trackers? To detect whether a tracker is blocked by protection mechanisms, we check whether it is blocked by the Disconnect filter list. Disconnect [35] is a popular list for detecting domains known for tracking. It is used in the Disconnect browser extension, and in the tracking protection feature of the Firefox browser. Given that we are using crawls in this study from March 2020 and May 2022, we used the combination of both Disconnect lists from the two dates. The adoption of SST brings significant privacy concerns, as trackers initially blocked by current protective measures can evade these blocking mechanisms by migrating to the server side. Consequently, ensuring control over tracking in the web, and protecting our privacy through domain blocking becomes a more complex task. We found that out of the detected 69 trackers shifted to the server side, 40 (57.97%) are included in the Disconnect filter list and therefore would be blocked if included directly on the website.

Summary. Out of the detected 389 websites where at least one cloaked tracker is detected, we analyzed the shifting of trackers to the server side and detected 32 SST domains. We detected a total of 69 unique tracker subdomains shifted to the server side. If included directly in the website, 57.97% of these trackers would have been blocked by privacy-preserving filter lists.

5.5 Data shared with SST subdomains

We analyzed the request URLs and Post data sent to the SST domains, and the script calls performed by these domains. We searched for the presence of browser and machine features that can help link to the user's identity, namely *UserAgent, AppVersion, Language, Platform, CookieEnabled, appName, DoNotTrack, Location, Screen resolution, and Canvas*. These features are the most common ones found in the field of browser fingerprinting and can help identify a user or, at the very least, provide statistics for analytics [4, 12, 23, 24, 32, 41, 54, 66, 72].

We detected that the SST subdomains accessed the *UserAgent, CookieEnabled, AppVersion, Language, appName, and Screen resolution*. Table 4 summarizes the top features accessed by the SST subdomains. We found that 6 SST domains access at least one of the studied browser and machine information, and that the top feature accessed by the SST subdomains is *UserAgent*.

features	# SSTs	features	# SSTs
UserAgent	6	CookieEnabled	6
appVersion	3	screen	3

Table 4: Top features accessed by the SST subdomains.

Summary. SST domains are receiving a number of information using the HTTP(S) requests: URL and post data. In our dataset, we detect that 18.75% of the SST domains are receiving at least one of the user's browser and machine features.

5.6 Websites hosting SST subdomains

In this section, we study the websites where SST occurs. First, we analyze the impact of Alexa ranking on the adoption of the SST, and then we report on the category of websites hosting SST.

Popularity of websites including SST. We analyzed the distribution of the SST across the visited websites rank. We extracted the percentage of SST subdomains for bins of 100 websites across the 7,367 visited websites. We found that SST is consistently deployed on the visited websites. The ratio of SST varies between 0% and 7%.

In conclusion, in this study, we found that the website rank does not have an impact on the deployment of SST.

Categorization of websites including SST. The McAfee service [70] uses various technologies and artificial intelligence techniques, such as link crawlers, and customer logs to categorize websites. We used the McAfee service to categorize the visited websites. We successfully categorized 7,268 (98.66%) visited websites. The top category of visited websites is General News accounting for 9.73% of visited websites. We analyzed the category of websites including SST, and we found that SST is happening on 20 categories of websites. The top category with the highest number of websites including SST is Online Shopping. In fact, 17.86% of websites including SST are categorized as Online Shopping (see Appendix B).

Summary. In this section, we showed that SST is commonly deployed across websites with different Alexa rankings. Both the less popular and the most popular websites use SST. Moreover, we showed that SST is used in 20 distinct categories of websites, and the top category including such behavior is Online Shopping.

6 EVALUATING THE COMPLIANCE OF SST

In this section, together with a legal expert and co-author, we discuss the main legal issues of potential SST on 28 websites that include at least one SST subdomain. Our experiments assessing the practices of the crawled websites are performed from the EU. As the crawled websites monitor EU user's behavior, the GDPR [53] and the ePD [43] are applicable and establish the obligations impending on these websites.

Legal framework. The GDPR [53] applies to the processing of personal data [45] and imposes obligations on organizations (named data controllers, in our case represented by the website owners), paired with heavy fines for non-compliance. Any data controller (inside or outside of the EU) monitoring the behavior of users located in the EU must follow its rules (Article 3 GDPR). Websites are required to choose a legal basis to process personal data (Article 6(1)(a)). In case this legal basis is consent, the GDPR also defines strict requirements for valid consent (Articles 4(11) and 7). Consent must be prior to any data collection, freely given, specific, informed, unambiguous, readable, accessible, and revocable (Articles 4(11) and 7 GDPR) [83]. Additionally, data controllers will need to be accountable and show compliance with the GDPR principles: transparency, fairness, lawfulness, and security, among others (Article 5(2) GDPR). Compliance with the GDPR is enforced by Data Protection Authorities (DPAs) which monitor and supervise the application of the GDPR (Articles 55-57 GDPR). The ePrivacy Directive (ePD) [43] provides *supplementary* rules to the GDPR in particular for the use of tracking technologies. When storing cookies or other tracking technologies or when gaining access to information already stored in the user's terminal equipment, the ePD applies. The following legal analysis is based on the GDPR, ePD, as well as in its recitals (which help the interpretation of rules in a specific context, though they are not mandatory for compliance). We also consulted case decisions of the Court of Justice of the EU, guidelines of both the European Data Protection Board (an EU advisory board on data protection),and Data Protection Authorities to evaluate the compliance of SST.

6.1 Personal data is shared in SST

Findings. We classify a domain as an SST domain when it is actively receiving an Identifier cookie (Section 5.3). This Identifier cookie can be deployed for the identification of users. In Section 5.5, we further analyze the utilization of diverse browser and machine features by these SST domains, which could potentially be employed for fingerprinting, thereby enhancing user identifiability.

Legal analysis. Personal data means any information relating to an identified or identifiable natural person (data subject); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an online identifier (Article 4(1) [53]). Online identifiers, such as cookies are considered to be personal data (Recital 30 [53]. The Article 29 Working Party explicitly states that fingerprints can also constitute personal data[3]. In order to determine whether a person is *identifiable*, account should be taken of *all the means that can reasonably be used* by any entity to identify that person (Recital 26 GDPR). This approach to identifiability [30, 47] means that anyone possessing the means to identify a user renders such a user identifiable. Our results show that ID sharing occurs through cookies and HTTP requests (Figure 4 and Table 4) between a user's browser and SST servers, and thus, SST domains could contain enough data to uniquely identify particular devices or application instances and thus render a user identifiable. Given this fact, SST domains seem to have reasonable means to combine a set of information elements relating to an identifiable person, rendering that information personal data. As the GDPR applies, the processing of personal data needs to be compliant thereto.

6.2 Consent to tracking is absent in SST

Findings. As found in section 5.3, SST domains received *ID cookies*. In the case of first-party SST domains, the identifier is set/sent on the user's browser as a first-party cookie. We evaluated the purpose of the cookies set by the SST domains. We used the Cookiepedia open database [29] that contains the cookie name to get the purpose of a cookie. Cookiepedia is the largest database

of precategorized cookies with over 11 million cookies. It uses the classification system developed by "The UK International Chamber of Commerce" (ICC) and relies on four common purposes of cookies: i) Strictly Necessary, ii) Performance (also known as analytics, statistics, or measurement cookies); iii) Functionality, and iv) Targeting (known as advertising). We found that 35% of the cookies set/sent to the SST domains and classified by Cookiepedia are categorized as Targeting/Advertising.

Legal analysis. According to Article 5(3) of the ePD [43], user's consent is required for accessing or storing non-technically necessary tracking technologies, like advertising, on a user's device. A website must then ask the user's consent if third parties place cookies on the visitor's computer [44]. Accordingly, websites including SST subdomains must ask user's consent for the deposit of cookies and other tracking technologies for advertising purposes. In this work, we did not interact with the cookie banner, and thus, never consented to any form of tracking or data sharing on the visited websites to any third party. However, we observed that data was shared between a user's browser and SST domains and with the third-party trackers on the server side (as reflected in Figure 4 and Table 4) *without consent*, which is required for advertising purposes. This practice renders all the processed data within SST unlawful since it was shared without the required legal basis.

6.3 Difficulty to assess a purpose of a cookie within SST

Findings. We consider that a service is an SST domain if it is performing IP cloaking (see Section 4.2). That is, the first-party subdomain points to a different organization compared to the visited website: the SST domain is, in fact, a third-party domain *hidden* as a first party. The cookies set by the SST domain will appear in the browser as first-party cookies set by a website's domain. However, these cookies are, in practice, set by a third-party service (the SST service).

Legal analysis. When accessing a website, users must be able to access all necessary information about the different types of purposes of cookies being used by that website [18]. While analysing the cookies present on a website, an auditor or a regulator will need to capture the *purpose* of each cookie. This defined purpose can then help to determine whether the processing is legally compliant, what safeguards the GDPR imposes, and which legal basis can be used [17]. However, when cookies are used with SST, it is not possible to assess the purpose and usage of cookies, since the domain of a cookie is not defined. This is especially the case when the cookie set by the SST service is later shared with multiple domains. We might guess who the cookie setter is only if the cookie name is commonly known, for example, the _ga cookie, but we can not make this kind of assumption for all cookies whose name is not defined. Hence it is highly difficult to derive the purpose of cookies within SST. The difficulty of discerning the purposes of cookies entails an important consequence: it will make it impossible for an auditor to determine what is the appropriate legal basis and whether the processing is compliant with the GDPR and with the ePD [51].

6.4 Lack of transparency in SST

Findings. First-party SST services are included in the detected 28 websites as first-party domains. These domains are setting cookies on the user's browser. Even though these cookies are set by third-party services, such cookies are hard to block because they are set in a first-party context.

Legal analysis. An informed consent request must include *"information on the purpose(s) of the cookies and an indication of possible cookies from third parties or third party access to data collected by the cookies on the website (...). Details of third-party cookies and other technical information should also be included to fully inform users"* [18]. DPAs still rely on the distinction between first-party cookies and third-party cookies [18, 62] to determine the entity that will potentially bear data protection obligations and to serve as an initial indicator to prioritize compliance actions [18] (alongside the purpose of cookies). Both Data Protection Officers (DPOs) – who oversee and evaluate the overall compliance of the companies' websites–, and DPAs will have a complex task to audit the legal compliance of websites when SST domains appear instead as first-party subdomains, and set first-party cookies. In this SST scenario, whenever a user or auditor verifies that only first-party cookies are stored in the browser by a first-party subdomain, they might consider such first-party cookie as *"strictly", or "technically necessary"* for the website to function, and accordingly, a functionality that could have been "explicitly requested" by the user [16] (therefore exempted of consent), and also less privacy-invasive. Website visitors and auditors would not reasonably expect such hidden tracking of their online activities to take place, and hence might not be in a position to avoid data collection. In the light of the principle of transparency [89], the GDPR imposes that users should be made aware of the risks in relation to the processing of personal data (Recital 39 [53]). This principle also requires that users should be informed *in situations where the proliferation of actors and the technological complexity of practice make it difficult for the data subject to know and understand whether, by whom and for what purpose personal data relating to him or her are being collected, such as in the case of online advertising* (Recital 58 [53]). In this line, the Court of Justice of the EU ruled [1] that the information provided by the company operating cookies *"must be clearly comprehensible and sufficiently detailed so as to enable the user to comprehend the functioning of the cookies employed."* It ruled further [2] that the disclosure *"must enable the data subject to be able to determine easily the consequences of any consent he or she might give and ensure that the consent given is well informed."*. Consequently, SST might infringe the transparency principle. Notwithstanding strict guidelines for online tracking, SST likely creates a potentially obscure processing, for both users and regulators, since it hides on the server side the third-party scripts, what data are being sent to them, and the conditions under which data is processed. The difficulty in distinguishing first-party from third-party cookies misrepresents the legitimate expectations of users and infringes the fairness principle and the data protection by default obligation that demands the most privacy-friendly default settings (Article 25 [53]).

6.5 Personal data is shared in SST

Findings. In Section 5.5, we detected that 6 (18.75%) SST domains are accessing the user's browser and machine features, while this sharing is performed without user's consent.

Legal analysis. Under the security principle, the GDPR prohibits the processing of personal data unless it is kept secure, and this includes "protection against unauthorised or unlawful processing" (Article 5(1)(f), Recitals 39, 78, and 83 [53]). To avoid it, controllers are responsible for ensuring the security of the processing through technical and organisational measures, "taking into account the state of the art, risks and severity for the users". Accordingly, the studied websites might incur in a data breach (Article 4(12) [53]).

7 USE CASES

In this section we perform an in-depth technical and legal analysis of two use cases: i) one website that relies on Google SST, and ii) another handling user-sensitive data that includes SST domains. For each website, we analysed the cookie banner and the privacy policy and identified potential legal violations.

7.1 Google's SST

In this study, we detected 13 SST subdomains belonging to Google: *sgtm.lcwaikiki.com, server.walmart.com.mx, gtm.minkabu.jp, gtm.beforward.jp, sgtm.t-mobile.com, analytics.teepublic.com, gtms.stern.de, tms.hft.hellofresh.com, sst.rocketnews24.com, analytics.bmj.com, tracking.zameen.com, data.statista.com, and sst.anibis.ch.* We detected that Google is among the top domains performing SST. Moreover, out of the top 5 SST subdomains providers, we found that Google is the only organization providing detailed descriptions of their SST service [57]. We selected the `t-mobile.com` [87] website which belongs to a telecommunication company providing services in the US and in the EU and which includes an SST subdomain operated by Google.

Technical Description. In our crawl, when visiting *t-mobile.com*, we detected that a request was sent to `sgtm.t-mobile.com`. We found that `sgtm.t-mobile.com` is an SST subdomain operated by Google. The request sent to the `sgtm.t-mobile.com` SST subdomain includes user's identifiers, machine information, and information on the visited website. In fact, a total of 21 parameters are attached with the request and one of them being an identifier named *cid*. This *cid* parameter is composed in part of the *_ga* cookie that refers to Google Analytics user's identifier.

Cookie banner and privacy policy analysis. The cookie banner of the website mentions the use of cookies for several purposes, including for advertising and analytics purposes. The privacy policy [88] mentions that the following data is collected: device identifiers (like cookies, beacons, Ad IDs, and IP addresses etc), geolocation data, cookie IDs, device IDs including mobile advertising IDs, IP address, MAC address collected through tracking technologies like web beacons, pixels, and other tracking technologies. The policy also mentions the use of analytics (Google Analytics) and advertising, though it does not mention that personal data is processed by an SST server. This website should request consent of the user to track for advertising purposes (Articles 4(11), 6(1)(a), and 7 [53]). The cookie banner of this website does not have a rejection button, only a button named "Manage cookies". Once this button is clicked, the user is redirected to a new page [86] entitled *"Do not sell my personal information Third Party Data Sharing"*. This page's information refers to the sharing of data as defined by the US California Consumer Privacy Act (CCPA) [6], whereby a user has the option to either accept or opt-out of the "selling of private information" to third parties. If a user has turned "on" the "Do not sell" preference, T-Mobile will restrict the data it shares with third parties. The cookie banner is not compliant with the ePD or GDPR rules for consent to tracking technologies and thus any tracking operated by this website will potentially be unlawful.

7.2 Sensitive websites

The GDPR [53, Recital 51] stipulates that personal data which are particularly sensitive by their nature, merit specific protection, as their processing could create significant risks to the fundamental rights of users. This data includes personal data revealing sensitive information such as data concerning a natural person's sex life or sexual orientation [53, Article 9], among others. Processing such categories of data is *forbidden*, unless allowed by the user's explicit consent [53, Article 9(2)]. We detected SST on a sensitive website: `bmj.com`. which is a health-related website intended for healthcare professionals that helps to share knowledge on healthcare. While `bmj.com` does not handle the medical data of patients per se, visiting such a website can reveal a specific medical condition or a health problem that tracking companies can infer. On this regard, the French Data Protection Authority refers that health data can be derived from crossing data allowing inferences on health status or health risk of a person [27]. An important decision of the Court of Justice of the EU [26] ruled that sensitive information can be inferred from data available online. This inference can in turn lead to targeted ads for a user on precise medical products. For example, Google allows on its platform health-related ads in some countries [56]. Another example is the alternative to the end of third party cookies detailed by the IAB called Seller Defined Audiences (SDA) [58]. It proposes in its taxonomy several health-related interests like "Smoking Cessation", "Chiropractors" or "Hair Loss Treatments" [59].

Technical description. In our crawl, when visiting `bmj.com`, we detected that it includes an SST domain: `analytics.bmj.com` operated by Google. This SST domain first receives a request with a unique parameter (the GTM identifier), and then sets a cookie on the user's browser. The resulting cookie has the same structure and name as the `google-analytics.com` cookie *_ga*. We noticed in this visit that the Google Analytics service is not included in this domain, and we strongly believe that the Google Analytics service is hidden behind the SST subdomain `analytics.bmj.com`, as depicted in Figure 1. The identifier stored in the resulting *_ga* cookie is then sent both to the SST subdomain, and to doubleclick.net as parameter value. The usage of SST behind a first-party domain can help trackers avoid blocking mechanisms, but also limits them to within-site tracking. However, in this example, we showed that `analytics.bmj.com` is deploying additional tracking mechanisms namely cookie synchronization with doubleclick.net to enable cross-site tracking.

Transparency and processing on the server side. When consulting the privacy policy of this website [22], we observed that

_ga cookie is declared therein as a *first-party cookie*. In the list of third parties in this same policy, Google is present as a third party. The privacy policy does not mention the tracking and processing of personal data on the server side. Even though the SST provider is declared in the policy (Google), the policy does not specify the fact that it is behind a first-party subdomain. When visiting the website, neither a user nor a website auditor will be able to see direct requests to GA, and hence might assume that there is no request to the third-party GA service, and that all the tracking happens in the first-party context which might be less privacy invasive [18]. However, in practice, the request to GA is made through the SST even without any interaction with the cookie banner. The guidelines of the former EDPB [16] are important for SST: it suggests that first-party analytics cookies are not likely to create a privacy risk *when* limited to first-party aggregated statistical purposes and when they are used by websites that mention them in their privacy policy. It also suggests that first-party analytics should be clearly distinguished from third-party analytics, which use a common third-party cookie to collect navigation information related to users across distinct websites, and which pose a substantially greater risk to privacy. [16].

Cookie banner and explicit consent. The cookie banner refers that cookies are used for targeting and advertising purposes, thus requiring user consent. Furthermore, for the purposes of online tracking in a sensitive-related website such as www.bmj.com, only the *explicit consent* exception seems to be the applicable legal basis to process this special category of data [53, Article 9(2)(a)]. An *explicit consent* request should abide to the following requirements [10, 36, 91, 94]: i) include double confirmation or verification from the user, ii) consist of a separated request from any other consents [46] (Recital 43 [53]), iii) specify the nature of the special category of data. Without this explicit consent from users, tracking on sensitive websites may therefore be found to infringe the lawfulness principle (Article 9 (2)(a) [53]), rendering any subsequent processing *unlawful*. The cookie banner depicts the options: "I Accept" and "Show Purposes", and a link to the "Cookie policy". There is no button for the rejection of cookies in the first layer of the banner, only on the second layer once we clicked on "Show Purposes". This configuration conflicts with requirements of "configurable banner" and "balanced choice" (Articles 4 (11), 7(3) [53]) [1], which are compulsory for an unambiguous consent of a user.

8 DISCUSSION AND RECOMMENDATIONS

This study aimed to look at Server-side tracking and highlight the dangers it can pose to users' privacy. We summarize the key lessons and observations we made during our crawls and analyses and we provide some recommendations for policy-makers.

SST is hard to block and will become even harder to block. At the time of writing, there are no existent solutions to block Server-side tracking specifically. We *recommend* alternative approaches to block trackers hidden behind cloaked DNS records and we refer to 2 ways to protect users against them. The first is using a filter list maintained by AdGuard that is updated weekly [7]. However, the introduction of Manifest V3 for browser extensions is hampering the efficiency of ad blockers by imposing a limit on the number of rules that can be used. In August 2022, AdGuard was the first to deploy an ad blocker based on Manifest v3 and they met a lot of barriers that limit the capacity of their extension [68]. The second way is to rely on the advanced filtering system of the uBlock Origin extension [90]. We also propose that some blocking rules can be defined to detect a first-party request being sent to a third-party server but this extension is currently the only one with such capabilities.

SST is hard to audit and makes it almost impossible to verify its lawfulness. When legal frameworks like GDPR came into effect, it was possible to verify if what happens on a webpage corresponds to what was declared in a privacy policy. An auditor just needed to use her own browser with some tooling to monitor the different requests made to third parties and everything could be checked remotely. For SST, the story is different. Despite having the same legal basis as regular tracking servers, this additional SST server completely changes the way monitoring from regulators' audits can be performed and how its lawfulness is verified:

- It becomes impossible to verify that the **transparency obligations** are respected on the server side. An auditor would need access to the SST server to see which trackers are actually present and if they all relate to the right legal basis.
- In conventional tracking scenarios, we recognize three primary actors: end-user, website owner, and third-party tracker, each with well-defined responsibilities and liabilities. However, SST introduces an additional element: the intermediate service, represented by the SST domain which plays a crucial role in data distribution to third-party trackers on the server side. Therefore, it carries its own set of **responsibilities and liabilities**. Assessing such responsibilities and liabilities of the SST domain can pose challenges. This complexity arises due to the nuanced interactions and dependencies among different entities, making it hard to pinpoint distinct boundaries of accountability.
- The GDPR grants users **data subject rights**, such as the right to object, deletion, and in particular, the right to access data (Article 15 GDPR). In traditional tracking methods, exercising this right to access data is relatively straightforward, as users typically contact the website owner or the third-party tracker directly to obtain their collected data. However, in the case of SST, determining *whom* to contact can be less clear for the user. Since third-party trackers are not visible to the user, data subjects are not able to know that third parties are collecting their data and are not able to exercise their right to access (among all the other GDPR rights) against such third parties.

With this study, we *alert* regulators about the need to consider the potential risks and violations entailed by Server-side tracking. Since everything happens in the background, alternatives must be found to properly verify the lawfulness of SST servers and provide sanctions in case the right obligations are not met.

A warning sign regarding the future of tracking. The use of server-side tracking coupled with the reliance on IP cloaking were prompted by the tracking community because of a change that is currently being operated in the ad ecosystem. They may be followed by new techniques in the future that are using existing standards or even completely new ones. Our worry as academics and also as web

users is that there are currently no technical changes planned in the future that would enable anyone to get more insight into these practices and future ones going forward. While progress is being made to phase out third-party cookies, a strong push is happening at the same time to develop these sneakier ways of tracking that would bypass some restrictions put in place. Because of the high potential for abuse, we *recommend* that discussions between browser vendors, standardization bodies, regulators, and major web actors should identify where progress can be made so that users are protected and can retain control over their data throughout its complete lifetime. We commend that the way this data is shared must also be mirrored in privacy policies, no matter if it is shared via direct connection to a third party or through Server-side tracking.

9 RELATED WORK

In the last decade, several studies analyzed web tracking technologies and showed that web tracking is constantly evolving. With new protection techniques, new tracking technologies are introduced. In this section, we first provide an overview of previous works on detection of web tracking techniques, and then present the literature on CNAME-based tracking.

Cookie-based tracking. Cookie-based tracking is the most commonly known tracking technique, and it has been widely studied in the last decade [5, 19, 39, 41, 67, 73, 74, 78, 80, 84]. We distinguish two main categories of cookie-based tracking: *within-site tracking*, and *cross-site tracking*. Previous works showed that Google is the top organization performing these tracking behaviors. Englehardt et al. [41] showed that Google is tracking users on over 70% out of the 1 million visited websites.

CNAME tracking. To help protect their privacy, users are deploying privacy-preserving extensions. A number of these extensions rely on filter lists such as Disconnect [35], EasyList [37], and EasyPrivacy [38]. Filter lists are built based on regular expressions, and know trackers hostnames. These lists are continuously updated to include and block known trackers. To evade these protection tools, trackers deploy the CNAME-based tracking. They rely on the CNAME records to be included in a same-site context, thus avoiding being blocked. Several papers studied CNAME tracking [14, 31, 34, 79]. Ren et al. [79] presented security breaches of CNAME tracking. They showed that CNAME tracking can break the browser cookie policy. In fact, sensitive cookies can be transferred to CNAME trackers. Dao et al. [31] has proven that browsers and privacy protection extensions are ineffective against CNAME tracking techniques. They presented an alternative supervised learning-based approach to detect CNAME tracking. Yan et al. [34] performed a large-scale analysis of CNAME-based tracking of 5.6M web pages. They detected 10,474 domains pointing to a CNAME-based tracker and found that 9.98% of the top 10,000 websites employ at least one CNAME-based tracker. In our study, we detect CNAME tracking on 5.28% of the visited websites. While Yan et al. employed manual analysis to filter out non-tracking cookies, primarily by checking the purpose of cookies if available to detect whether they are used for tracking, we employ a more in-depth automatic detection technique to filter out such cookies. The variance in our findings could be attributed to the difference in the detection of ID cookies.

While several related works have looked at the use of CNAME redirection for tracking, our paper uses it as a first step in our detection methodology to identify Server-side tracking servers. In this work, we analyze the IP cloaking more globally and use it as a first filter to detect Server-side tracking. The focus of our study is to detect SST servers hidden behind first-party subdomains. To the best of our knowledge, our work is the first to study Server-side tracking and uncover the trackers hidden behind these SST domains.

10 CONCLUSION AND FUTURE WORK

This work sheds light on the emerging ecosystem of SST, a tracking system that helps evade tracking countermeasures. Thanks to crawl data collected before and after the introduction of SST on the web, we are able to identify that 28 websites out of 7,367 have shifted a part of their tracking from third-party domains to a first-party SST server. While this number is low, our methodology only looks at a specific subset of websites partaking in SST activities and the actual number of SST servers may be much higher in reality. Moreover, we anticipate this number to increase with time. With the upcoming deprecation of third-party cookies, websites are still in a transitional phase and are testing alternatives with more and more companies choosing SST to continue tracking users.

By collaborating with a legal scholar, we also assessed the compliance of SST and noted that SST infringes both the GDPR and the ePD. Such a technique introduces several challenges. It can be difficult to detect such behavior and thus it is harder to block it. Server-side tracking is also making it difficult to discern the purposes of cookies, and the origin of resources which will render impossible for an auditor to determine whether the processing is compliant with current regulations. We hope that our research helps with addressing the security and privacy issues that we highlighted.

As part of future work, we aim to analyze the growth of SST over time, especially after the deprecation of third-party cookies. We also intend to explore a new avenue by immersing ourselves in the server-side environment. This involves establishing our own SST server with an online provider to gain insights from an internal perspective.

ACKNOWLEDGMENTS

This work has been supported by the ANR 22-PECY-0002 IPOP (Interdisciplinary Project on Privacy) project of the Cybersecurity PEPR and by the Hauts-de-France region in the context of the ASCOT project of the STaRS framework.

REFERENCES

[1] 2019. Judgment in Case C-673/17 Bundesverband der Verbraucherzentralen und Verbraucherverbände – Verbraucherzentrale Bundesverband eV v Planet49 GmbH. http://curia.europa.eu/juris/documents.jsf?num=C-673/17.

[2] 2020. Orange România case (Case C-61/19), ECLI:EU:C:2019:801. https://curia.europa.eu/juris/document/document.jsf?text=&docid=233544&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=13705373.

[3] 29 Working Party. 2014. Opinion 9/2014 on the application of Directive 2002/58/EC to device fingerprinting. https://www.dataprotection.ro/servlet/ViewDocument?id=1089.

[4] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juárez, Arvind Narayanan, and Claudia Díaz. 2014. The Web Never Forgets: Persistent Tracking Mechanisms in the Wild. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 674–689.

[5] Gunes Acar, Marc Juárez, Nick Nikiforakis, Claudia Díaz, Seda F. Gürses, Frank Piessens, and Bart Preneel. 2013. FPDetective: dusting the web for fingerprinters. In *2013 ACM SIGSAC Conference on Computer and Communications Security (CCS'13)*. 1129–1140.

[6] California Consumer Privacy Act. 2020. California Consumer Privacy Act (Final Text of Proposed Regulations). https://www.oag.ca.gov/sites/all/files/agweb/pdfs/privacy/oal-sub-final-text-of-regs.pdf

[7] adguardCname 2022. CNAME-cloaked trackers – AdGuard team. https://github.com/AdguardTeam/cname-trackers.

[8] AdRevenueShare2021 2021. Google, Facebook, and Amazon to account for 64% of US digital ad spending this year – eMarketer. https://www.insiderintelligence.com/content/google-facebook-amazon-account-over-70-of-us-digital-ad-spending.

[9] AdSpending2021. 2022. Digital advertising spending worldwide 2021-2026 – Statista. https://www.statista.com/statistics/237974/online-advertising-spending-worldwide/.

[10] AEPD-Guid-21 2021. Guide on use of cookies. https://www.aepd.es/sites/default/files/2021-01/guia-cookies-en.pdf.

[11] Akamai 2023. Akamai website. https://www.akamai.com/company.

[12] Nasser Mohammed Al-Fannah, Wanpeng Li, and Chris J. Mitchell. 2018. Beyond Cookie Monster Amnesia: Real World Persistent Online Tracking. In *Information Security - 21st International Conference, ISC 2018, Guildford, UK, September 9-12, 2018, Proceedings (Lecture Notes in Computer Science, Vol. 11060)*, Liqun Chen, Mark Manulis, and Steve A. Schneider (Eds.). Springer, 481–501. https://doi.org/10.1007/978-3-319-99136-8_26

[13] Alexa websites 2021. Alexa websites. https://www.dropbox.com/scl/fo/s83a6mxo340ommat2asau/h?rlkey=svro2emp1fyv1d3pcrzwuqk75&dl=0.

[14] Assel Aliyeva and Manuel Egele. 2021. Oversharing Is Not Caring: How CNAME Cloaking Can Expose Your Session Cookies. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security* (Virtual Event, Hong Kong) *(ASIA CCS '21)*. Association for Computing Machinery, New York, NY, USA, 123–134. https://doi.org/10.1145/3433210.3437524

[15] Anthony Chavez. 2022. Expanding testing for the Privacy Sandbox for the Web – Google Blog. https://blog.google/products/chrome/update-testing-privacy-sandbox-web/.

[16] Article 29 Working Party. 2012. Opinion 04/2012 on Cookie Consent Exemption (WP 194).

[17] Article 29 Working Party. 2013. Opinion 03/2013 on purpose limitation (WP203).

[18] Article 29 Working Party. 2013. Working Document 02/2013 providing guidance on obtaining consent for cookies', (WP208).

[19] Mika D Ayenson, Dietrich James Wambach, Ashkan Soltani, Nathan Good, and Chris Jay Hoofnagle. 2011. *Flash Cookies and Privacy II: Now with HTML5 and ETag Respawning*. Technical Report. Available at SSRN: https://ssrn.com/abstract=1898390orhttp://dx.doi.org/10.2139/ssrn.1898390.

[20] Pouneh Nikkhah Bahrami, Umar Iqbal, and Zubair Shafiq. 2022. FP-Radar: Longitudinal Measurement and Early Detection of Browser Fingerprinting. *Proc. Priv. Enhancing Technol.* 2022, 2 (2022), 557–577.

[21] Ben Fisher. 2020. Improve performance and security with Server-Side Tagging – Google Blog. https://blog.google/products/marketingplatform/360/improve-performance-and-security-server-side-tagging/.

[22] bmj-policy 2023. Bmj privacy policy. https://www.bmj.com/company/legal-information/bmj-cookie-policy/.

[23] Károly Boda, Ádám Máté Földes, Gábor György Gulyás, and Sándor Imre. 2011. User Tracking on the Web via Cross-Browser Fingerprinting. In *16th Nordic Conference on Secure IT Systems, NordSec 2011*. 31–46.

[24] Yinzhi Cao, Song Li, and Erik Wijmans. 2017. (Cross-)Browser Fingerprinting via OS and Hardware Level Features. In *24th Annual Network and Distributed System Security Symposium, NDSS 2017, San Diego, California, USA, 26 February - 1 March, 2017*.

[25] Quan Chen, Panagiotis Ilia, Michalis Polychronakis, and Alexandros Kapravelos. 2021. Cookie Swap Party: Abusing First-Party Cookies for Web Tracking. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 2117–2129. https://doi.org/10.1145/3442381.3449837

[26] CJEU. 2022. JUDGMENT OF THE COURT (Grand Chamber) In Case C-184/20). https://curia.europa.eu/juris/document/document.jsf?docid=263721&doclang=EN.

[27] CNIL-donnee-sante 2021. Qu'est-ce ce qu'une donnée de santé ? https://www.cnil.fr/fr/quest-ce-ce-quune-donnee-de-sante. Accessed on 18 May 2021..

[28] ContextualCriteo 2022. Contextual advertising – Criteo. https://www.criteo.com/solutions/contextual-advertising/.

[29] cookiepedia 2023. Cookiepedia Official website. https://cookiepedia.co.uk/classify-cookies.

[30] Court of Justice of the European Union. 2016. Case 582/14 – Patrick Breyer v Germany. ECLI:EU:C:2016:779.

[31] Ha Dao, Johan Mazel, and Kensuke Fukuda. 2021. CNAME Cloaking-Based Tracking on the Web: Characterization, Detection, and Protection. *IEEE Trans. Netw. Serv. Manag.* 18, 3 (2021), 3873–3888. https://doi.org/10.1109/TNSM.2021.3072874

[32] Anupam Das, Gunes Acar, Nikita Borisov, and Amogh Pradeep. 2018. The Web's Sixth Sense: A Study of Scripts Accessing Smartphone Sensors. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS 2018, Toronto, ON, Canada, October 15-19, 2018*, David Lie, Mohammad Mannan, Michael Backes, and XiaoFeng Wang (Eds.). ACM, 1515–1532. https://doi.org/10.1145/3243734.3243860

[33] Nurullah Demir, Daniel Theis, Tobias Urban, and Norbert Pohlmann. 2022. Towards Understanding First-Party Cookie Tracking in the Field. In *Sicherheit, Schutz und Zuverlässigkeit: Konferenzband der 11. Jahrestagung des Fachbereichs Sicherheit der Gesellschaft für Informatik e.V. (GI), Sicherheit 2022, Karlsruhe, Germany, April 5-8, 2022*, Christian Wressnegger, Delphine Reinhardt, Thomas Barber, Bernhard C. Witt, Daniel Arp, and Zoltán Ádám Mann (Eds.). Gesellschaft für Informatik, Bonn. https://doi.org/10.18420/sicherheit2022_01

[34] Yana Dimova, Gunes Acar, Lukasz Olejnik, Wouter Joosen, and Tom Van Goethem. 2021. The CNAME of the Game: Large-scale Analysis of DNS-based Tracking Evasion. *Proceedings on Privacy Enhancing Technologies* 2021, 3 (7 2021), 1–19. https://doi.org/10.2478/popets-2021-0053

[35] disconnect 2023. Disconnect Official website. https://disconnect.me/.

[36] DPC-Guid-20 2020. Guidance note on the use of cookies and other tracking technologies. https://www.dataprotection.ie/sites/default/files/uploads/2020-04/Guidance%20note%20on%20cookies%20and%20other%20tracking%20technologies.pdf.

[37] easylist 2023. EasyList filter lists. https://easylist.to/.

[38] easyprivacy 2023. EasyPrivacy filter lists. https://easylist.to/easylist/easyprivacy.txt.

[39] Peter Eckersley. 2010. How Unique is Your Web Browser?. In *Proceedings of the 10th International Conference on Privacy Enhancing Technologies (PETS'10)*. Springer-Verlag, 1–18.

[40] Steven Englehardt and Arvind Narayanan. 2016. Online Tracking: A 1-Million-Site Measurement and Analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (Vienna, Austria) *(CCS '16)*. Association for Computing Machinery, New York, NY, USA, 1388–1401. https://doi.org/10.1145/2976749.2978313

[41] Steven Englehardt and Arvind Narayanan. 2016. Online Tracking: A 1-million-site Measurement and Analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security ACM CCS*. 1388–1401.

[42] 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 WWW 2015*. 289–299.

[43] ePD-09 2009. Directive 2009/136/EC of the European Parliament and of the Council of 25 November 2009 amending Directive 2002/22/EC on universal service and users' rights relating to electronic communications networks and services, Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector and Regulation (EC) No 2006/2004 on cooperation between national authorities responsible for the enforcement of consumer protection laws (Text with EEA relevance). Directive 2009/136/EC. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32009L0136.

[44] European Court of Justice. 2019. Case C-40/17 Fashion ID GmbH & Co.KG v Verbraucherzentrale NRW eV, ECLI:EU:C:2019:629.

[45] European Data Protection Board. 2007. Opinion 4/2007 on the concept of personal data (WP 136), adopted on 20.06.2007. https://ec.europa.eu/justice/article-29/documentation/opinionrecommendation/files/2007/wp136_en.pdf.

[46] European Data Protection Board (EDPB). 2010. Opinion 2/2010 on online behavioural advertising, 22 June 2010, WP171, p. 10.

[47] Michèle Finck and Frank Pallas. 2020. They Who Must Not Be Identified – Distinguishing Personal from Non-Personal Data Under the GDPR. *International Data Privacy Law* 10 (2020).

[48] Imane Fouad, Nataliia Bielova, Arnaud Legout, and Natasa Sarafijanovic-Djukic. 2020. Missed by Filter Lists: Detecting Unknown Third-Party Trackers with

Invisible Pixels. In *PETS 2020 - 20th Privacy Enhancing Technologies Symposium (PETs (Privacy Enhancing Technologies Symposium))*. Montréal, Canada. https://hal.inria.fr/hal-01943496

[49] Imane Fouad, Nataliia Bielova, Arnaud Legout, and Natasa Sarafijanovic-Djukic. 2020. Missed by Filter Lists: Detecting Unknown Third-Party Trackers with Invisible Pixels. *Proc. Priv. Enhancing Technol.* 2020. https://doi.org/10.2478/popets-2020-0038

[50] Imane Fouad, Cristiana Santos, Feras Al Kassar, Nataliia Bielova, and Stefano Calzavara. 2020. On Compliance of Cookie Purposes with the Purpose Specification Principle. In *IWPE 2020 - International Workshop on Privacy Engineering*. Genova, Italy, 1–8. https://hal.inria.fr/hal-02567022

[51] Imane Fouad, Cristiana Santos, Feras Al Kassar, Nataliia Bielova, and Stefano Calzavara. 2020. On Compliance of Cookie Purposes with the Purpose Specification Principle. In *2020 International Workshop on Privacy Engineering, IWPE*. https://hal.inria.fr/hal-02567022.

[52] Imane Fouad, Cristiana Santos, Arnaud Legout, and Nataliia Bielova. 2022. My Cookie is a phoenix: detection, measurement, and lawfulness of cookie respawning with browser fingerprinting. In *PETS 2022 - 22nd Privacy Enhancing Technologies Symposium*. Sydney, Australia. https://hal.archives-ouvertes.fr/hal-03218403 Accepted at the 22nd Privacy Enhancing Technologies Symposium (PETS 2022).

[53] GDPR 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) (Text with EEA relevance). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32016R0679.

[54] 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 *WWW2018 - TheWebConf 2018 : 27th International World Wide Web Conference*. Lyon, France, 1–10. https://doi.org/10.1145/3178876.3186097

[55] Google3PCookiesDeath 2024. Update on the plan for phase-out of third-party cookies on Chrome – The Privacy Sandbox blog. https://privacysandbox.com/intl/en_us/news/update-on-the-plan-for-phase-out-of-third-party-cookies-on-chrome/.

[56] GoogleHealth 2024. Healthcare and medicines – Google Advertising Policies Help. https://support.google.com/adspolicy/answer/176031.

[57] GoogleSST 2023. Google Server-side tagging. https://developers.google.com/tag-platform/tag-manager/server-side.

[58] IABSDA 2024. Seller Defined Audiences – IAB Tech Lab. https://iabtechlab.com/sda/.

[59] IABSDA2 2024. Audient Taxonomy 1.1 – IAB Tech Lab. https://github.com/InteractiveAdvertisingBureau/Taxonomies/blob/main/Audience%20Taxonomies/Audience%20Taxonomy%201.1.tsv.

[60] IdentityGraphsAWS 2022. Identity Graphs on AWS – AWS. https://aws.amazon.com/neptune/identity-graphs-on-aws/.

[61] IdentityGraphsLiveRamp 2022. Identity Resolution – LiveRamp. https://liveramp.com/identity-resolution/.

[62] ItalianDPA-cookiewall-2021 2021. *Guidelines on the use of cookies and other tracking tools*. Technical Report. GuidelinesontheuseofcookiesandothertrackingtoolsIt.

[63] John Wilander. 2021. Introducing Private Click Measurement, PCM – WebKit Blog. https://webkit.org/blog/11529/introducing-private-click-measurement-pcm/.

[64] Justin Schuh. 2020. Building a more private web: A path towards making third party cookies obsolete – Chromium Blog. https://blog.chromium.org/2020/01/building-more-private-web-path-towards.html.

[65] Pierre Laperdrix, Nataliia Bielova, Benoit Baudry, and Gildas Avoine. 2020. Browser Fingerprinting: A Survey. *ACM Trans. Web* 14, 2, Article 8 (apr 2020), 33 pages. https://doi.org/10.1145/3386040

[66] Pierre Laperdrix, Walter Rudametkin, and Benoit Baudry. 2016. Beauty and the Beast: Diverting modern web browsers to build unique browser fingerprints. In *37th IEEE Symposium on Security and Privacy (S&P 2016)*. https://hal.inria.fr/hal-01285470

[67] Adam Lerner, Anna Kornfeld Simpson, Tadayoshi Kohno, and Franziska Roesner. 2016. Internet Jones and the Raiders of the Lost Trackers: An Archaeological Study of Web Tracking from 1996 to 2016. In *25th USENIX Security Symposium (USENIX Security 16)*. USENIX Association.

[68] manifestv3AdGuard 2022. AdGuard publishes the world's first ad blocker built on Manifest V3 – AdGuard. https://adguard.com/en/blog/adguard-mv3.html.

[69] Martin Thomson. 2022. Privacy Preserving Attribution for Advertising – Mozilla Blog. https://blog.mozilla.org/en/mozilla/privacy-preserving-attribution-for-advertising/.

[70] mcafee 2023. McAfee categorization service. https://sitelookup.mcafee.com/.

[71] Vikas Mishra, Pierre Laperdrix, Antoine Vastel, Walter Rudametkin, Romain Rouvoy, and Martin Lopatka. 2020. Don't Count Me Out: On the Relevance of IP Address in the Tracking Ecosystem. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 808–815. https://doi.org/10.1145/3366423.3380161

[72] Keaton Mowery and Hovav Shacham. 2012. Pixel Perfect: Fingerprinting Canvas in HTML5. In *Proceedings of W2SP 2012*, Matt Fredrikson (Ed.). IEEE Computer Society.

[73] Nick Nikiforakis, Alexandros Kapravelos, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. 2013. Cookieless Monster: Exploring the Ecosystem of Web-Based Device Fingerprinting. In *IEEE Symposium on Security and Privacy, SP 2013*. 541–555. https://doi.org/10.1109/SP.2013.43

[74] Lukasz Olejnik, Minh-Dung Tran, and Claude Castelluccia. 2014. Selling off User Privacy at Auction. In *21st Annual Network and Distributed System Security Symposium, NDSS 2014, San Diego, California, USA, February 23-26, 2014*.

[75] OpenWPM 2022. Information stored by OpenWPM. https://github.com/mozilla/OpenWPM.

[76] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos Markatos. 2019. Cookie Synchronization: Everything You Always Wanted to Know But Were Afraid to Ask. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1432–1442. https://doi.org/10.1145/3308558.3313542

[77] Ratcliff-Obershelp 2024. Ratcliff-Obershelp algorithm. https://xlinux.nist.gov/dads/HTML/ratcliffObershelp.html.

[78] Abbas Razaghpanah, Rishab Nithyanand, Narseo Vallina-Rodriguez, Srikanth Sundaresan, Mark Allman, Christian Kreibich, and Phillipa Gill. 2018. Apps, Trackers, Privacy, and Regulators: A Global Study of the Mobile Tracking Ecosystem. In *Network and Distributed System Security Symposium, NDSS*.

[79] Tongwei Ren, Alexander Wittman, Lorenzo De Carli, and Drew Davidson. 2021. An analysis of first-party cookie exfiltration due to cname redirections. *Workshop on Measurements, Attacks, and Defenses for the Web (MADWeb)* (2021).

[80] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. 2012. Detecting and Defending Against Third-Party Tracking on the Web. In *Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2012*. 155–168.

[81] SameOriginPolicy 2010. Same Origin Policy. https://www.w3.org/Security/wiki/Same_Origin_Policy.

[82] Iskander Sanchez-Rola, Matteo Dell'Amico, Davide Balzarotti, Pierre-Antoine Vervier, and Leyla Bilge. 2021. Journey to the Center of the Cookie Ecosystem: Unraveling Actors' Roles and Relationships. In *2021 IEEE Symposium on Security and Privacy (SP)*. 1990–2004. https://doi.org/10.1109/SP40001.2021.9796062

[83] Cristiana Santos, Nataliia Bielova, and Célestin Matte. 2020. Are cookie banners indeed compliant with the law? Deciphering EU legal requirements on consent and technical means to verify compliance of cookie banners, journal=Technology and Regulation. (2020), 91–135. https://doi.org/10.26116/techreg.2020.009

[84] Ashkan Soltani, Shannon Canty, Quentin Mayo, Lauren Thomas, and Chris Jay Hoofnagle. 2010. Flash Cookies and Privacy. In *AAAI Spring Symposium: Intelligent Information Privacy Management*.

[85] SST1stParty 2023. Map a custom domain to your service – Server-side tagging fundamentals from Google. https://developers.google.com/tag-platform/learn/sst-fundamentals/8-upgrade-infrastructure#map-a-custom-domain-to-your-service.

[86] t-mobile 2023. T-mobie Manage cookies page. https://www.t-mobile.com/dns.

[87] t-mobile 2023. T-mobile website. https://www.t-mobile.com/.

[88] t-mobile-policy 2023. T-mobie privacy policy. https://www.t-mobile.com/privacy-center/privacy-notices/t-mobile-privacy-notice.

[89] Transparency29WP 2018. "Guidelines on transparency under Regulation 2016/679, WP260 rev.01. https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=622227.

[90] ublockSST 2022. Static filter syntax – uBlock Origin. https://github.com/gorhill/uBlock/wiki/Static-filter-syntax.

[91] UK DPA. 2020. Guidance on the rules on use of cookies and similar technologies', 2020.

[92] Vinay Goel. 2021. An updated timeline for Privacy Sandbox milestones – Google Blog. https://blog.google/products/chrome/updated-timeline-privacy-sandbox-milestones/.

[93] Vinay Goel. 2022. Get to know the new Topics API for Privacy Sandbox – Google Blog. https://blog.google/products/chrome/get-know-new-topics-api-privacy-sandbox/.

[94] Vera Wesselkamp, Imane Fouad, Cristiana Santos, Yanis Boussad, Nataliia Bielova, and Arnaud Legout. 2021. In-Depth Technical and Legal Analysis of Tracking on Health Related Websites with ERNIE Extension. In *Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society* (Virtual Event, Republic of Korea) *(WPES '21)*. Association for Computing Machinery, New York, NY, USA, 151–166. https://doi.org/10.1145/3463676.3485603

[95] whois 2023. Whois library. https://pypi.org/project/whois/.

[96] David Zeber, Sarah Bird, Camila Oliveira, Walter Rudametkin, Ilana Segall, Fredrik Wollsén, and Martin Lopatka. 2020. The Representativeness of Automated Web Crawls as a Surrogate for Human Browsing. In *Proceedings of The Web Conference 2020*. 167–178.

11 APPENDIX

A Cookies lifetime

We analyze the lifetime of the cookies used by cloaked trackers and SST domains. Figure 5 presents the results of our analysis. For each domain, we consider the maximum lifetime of cookies set or received by the domain. Our findings reveal that 401(84.60%) out of 474 cloaked trackers utilize cookies with an expiration date exceeding 30 days, a threshold commonly employed by related works as a filter for the expiration date [40]. Additionally, we observed that 31 (96.87%) out of 32 detected SST domains have cookies with expiration dates higher than 30 days.

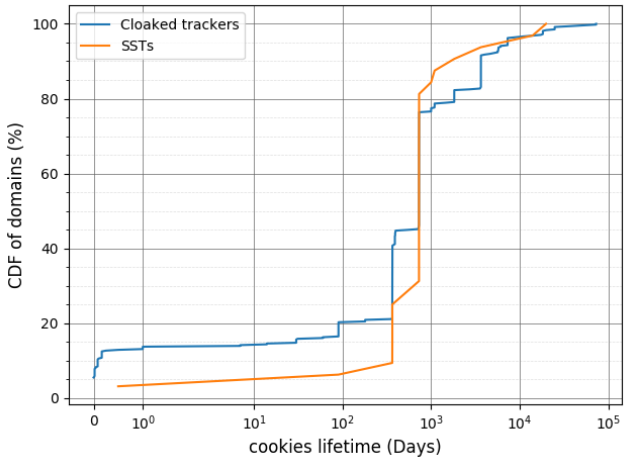


Figure 5: Cookies lifetime.

B SST websites categorization

Table 5 presents the top 5 categories of websites including SST. We analyzed the category of the visited 7,367 websites, and found that Internet services, Finance/banking, and Online shopping rank among the top 10 visited categories, indicating high user usage of these categories. Motor Vehicles and Online shopping are the categories that present the highest inclusion of SST within their category.

C Machine characteristics

Table 6 presents the characteristics of *Pre-SST crawl* and *Post-SST crawl* used in our study.

Website category	SST Prevalence	Category prevalence	Inclusion
Online Shopping	17.86%	4.88%	1.38%
Finance/Banking	14.29%	5.7%	0.95%
Business	10.71%	7.07%	0.58%
Motor Vehicles	10.71%	0.55%	6.82%
Internet Services	7.14%	5.90%	0.46%

Table 5: Top 5 category of websites including SST. *SST Prevalence*: refers to the percentage of websites employing SST out of the total 28 detected SST websites instances. *Category prevalence*: represents the category prevalence out of the studied 7,367 websites. *Inclusion*: indicates the percentage of SST websites within the same category.

Characteristics	*Pre-SST crawl*	*Post-SST crawl*
Visited websites	Alexa top 10,000 [13]	Alexa top 10,000 [13]
Visited webpages	Landing pages	Landing pages
Date of the crawl	March 2020	May 2022
Crawl state	Stateless	Stateless
OpenWPM version	OWPM 0.14.0	OWPM 0.19.1
Browser	FF86.0.1	FF95
Location	EU	EU

Table 6: Crawls Characteristics.