IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

ePrivo.eu: An Online Service for Automatic Web Tracking Discovery

ISMAEL CASTELL-UROZ, ISMAEL DOUHA-PRIETO, MERITXELL BASART-DOTRAS, POL MESEGUÉ-MOLINA, AND PERE BARLET-ROS
Universitat Politècnica de Catalunya, Barcelona, Spain (e-mail:name.surname@upc.edu)
Corresponding author: Ismael Castell-Uroz (e-mail: ismael.castell@upc.edu).

This publication is part of the Spanish I+D+i project TRAINER-A (ref. PID2020-118011GB-C21), funded by MCIN/AEI/10.13039/501100011033.

ABSTRACT Given the pervasiveness of web tracking practices on the Internet, many countries are developing and enforcing new privacy regulations to ensure the rights of their citizens. However, discovering websites that do not comply with those regulations is becoming very challenging, given the dynamic nature of the web or the use of obfuscation techniques. This work presents ePrivo, a new online service that can help Internet users, website owners, and regulators inspect how privacy-friendly a given website is. The system explores all the content of the website, including traffic from third parties and dynamically modified content. The ePrivo service combines different state-of-the-art tracking detection and classification methods, including TrackSign, to discover both previously known and zero-day tracking methods. After 6 months of service, ePrivo detected the largest browsing history trackers and more than 40k domains including cookies with a lifespan longer than one year, which is forbidden in some countries.

INDEX TERMS Fingerprinting, Privacy, Web Tracking, Web Embedding

I. INTRODUCTION

IN a world where the collection of personal data has a tremendous impact on the revenue of many companies (the size of the online data analytic market is expected to reach $10.73B by 2026 [1]), the Internet has become a place where almost every website (>95% [2]) tries to obtain personal information about their users by means of web tracking methods. The loss of privacy that this practice entails is not always perceived but, when it is known, it is seen as inevitable. However, only the most obvious web tracking systems (e.g., cookies) are commonly known. Most companies use complementary fingerprinting methods hidden in the background to better identify users, which can be used to infer sensible information, such as gender, race, income or even political ideologies. Thus, many users are not aware that their actions on the Internet can (and will most likely) affect other aspects of their lives. The personal information collected using web tracking techniques has been used, for example, to decide whether to give someone a loan depending on the financial status of their Facebook friends, or accepting or denying someone's insurance coverage based on their lifestyle and hobbies, among many other examples [3].

Fortunately, in the last few years, governments have started to grasp the importance of privacy on the Internet. For example, the EU in 2018 and California in 2020 enforced new mandatory regulations focused on online privacy. China published similar regulations just a few months ago. All those regulations define some useful policies to defend users' privacy from such questionable usages, or at least to obtain the user's consent in order to collect their personal information (Section II-B). However, due to the complex web of interconnections that form the Internet ecosystem, determining whether a website complies with those regulations is not an easy task. Thus, regardless of all the previous work done on the topic, there is still a lack of tools that can help users as well as other privacy-concerned actors understand the data collection processes being performed underneath any given website.

In this work, we present ePrivo [4], an electronic privacy (e-privacy) observatory that reports privacy-threatening information about millions of online resources. ePrivo works by examining the HTML and JavaScript code of all the elements included in any website, looking for pieces of code used for tracking purposes. The detection system is based on a state-of-the-art algorithm (TrackSign [2]) that looks for common tracking code shared among many websites. ePrivo also automatically classifies the detected tracking code according to the particular method used by the website to collect personal information (e.g., cookie type, fingerprinting method). ePrivo is a web-based public service that can help

improve the privacy ecosystem of the Internet by letting regulators easily inspect any website and find common privacy threats against the current regulations. Moreover, ePrivo can be used by any person willing to know more about the privacy threats present in the websites they frequently visit.

The rest of the paper is organized as follows: Section II presents the required background and the related work. Section III describes the architecture of the system and the web tracking detection and classification algorithms behind ePrivo. In section IV we present some common use cases of ePrivo to explore privacy aspects of the Internet. Section V includes some illustrative results obtained with ePrivo, while Section VI briefly discusses the future of privacy-preserving methods within the current Internet context. Finally, Section VII concludes the paper and presents future work.

II. BACKGROUND AND RELATED WORK

A. WEB TRACKING

Web tracking is a collection of techniques to identify and follow users' activities on the Internet. Although web tracking technologies were initially designed to provide a better browsing experience, nowadays they are extensively used to collect large amounts of personal information from our online activity, including information about our online browsing, searches, purchases, or even people with whom we are in contact. Web tracking can be used for beneficial purposes like maintaining an open session or keeping track of the contents of our online shopping basket. However, it has been demonstrated that this information is also used for many purposes, such as credit evaluation, personal assessment and price discrimination [3]. Moreover, the existence of the so-called *data-brokers* (companies whose goal is to do business exclusively with the collected personal data) is usually unknown by the common user. Data-brokers act as third-parties, acquiring, summarizing, and reselling personal data similarly to any other exchange currency. However, the data is often incomplete, obsolete, or even erroneous, as they can mix information from different people who shares some identifiers (e.g., name and surname) or incorrectly label people based on partial information. As they usually do not interact directly on websites (or their interaction is hidden by several layers of ever-changing domains), in many cases, obtaining clear and easy-to-understand information with respect to the type of personal information collected and the way it is used is extremely complicated or even impossible.

Two of the most prevalent methods to collect the data are by means of *cookies* or *user fingerprinting*. A cookie is a small piece of code that the server sends to the client, and it remains stored in the browser's cache. Each time the user opens the same website, the corresponding cookie is sent back to the server. Cookies are the most common web tracking method, as they permit precisely identifying the same user across sessions. Nowadays, most websites use cookies, and their existence is usually known even by non-expert users. On the other hand, fingerprinting techniques are used to precisely identify a person by generating an identifier based on a combination of characteristics present in the computer used to access the Internet (e.g., CPU, OS, Browser). Considering that every piece of software installed in the computer modifies its characteristics in a measurable way, it is very unusual to have two different people with the same exact combination of characteristics, making fingerprinting methods among the most powerful tools to identify users. In fact, some fingerprinting methods are so powerful that they can be used alone, although usually the identification precision increases using a combination of different fingerprinting techniques. Due to the difficulty of implementing such algorithms, it is very uncommon for a website to use a completely custom web tracking system. Usually, websites include generic web tracking services that load their libraries within their code. This is the so-called "web embedding", calling third-party resources and code inside the main website.

Currently, the most popular solution to fight web tracking are the so-called Adblockers (e.g., [5]–[7]), small browser plugins intercepting in real time every URL being loaded and comparing them to manually curated block lists containing all the known privacy-threatening domains. The research community has been actively looking for ways to improve the scene. Recent advances have tried to block non-requested data collection algorithms by means of machine learning methods that detect specific kinds of tracking algorithms. For instance, in [8] Ikram et al. applied a one-class learning algorithm to the website code to find features that can identify some web tracking methods. Iqbal et al. in [9] apply random forests on a subset of features extracted from both, network requests and website code, to detect a subset of fingerprinting methods. In [10], they present a new system called FP-Inpector, that applies static analysis to the AST representing the JavaScript code, where they look for a subset of API calls commonly used for tracking. Unfortunately, static analysis is usually vulnerable to obfuscation and similar techniques. To solve the problem, they introduce complementary dynamic analysis systems in order to recover the original features, greatly improving the system's accuracy to find web tracking using those API calls. However, machine learning solutions suffer from overspecialization, detecting only the tracking systems used during the training phase. In contrast, TrackSign [2], the detection method used by ePrivo, computes statistically with a high degree of accuracy (more than 92%) the probability of the website code to include web tracking algorithms, regardless of the web tracking system being used.

B. ONLINE PRIVACY REGULATIONS

Given the pervasiveness of web tracking algorithms on the Internet (>95%) [2], many countries are starting to develop and enforce policies and regulations to protect their citizens' privacy. During the last years, some research works (e.g., [11]–[13]) have focused on the impact of those regulations in the economic and website environment. Others looked at the evolution and characteristics of the new consent systems [14] or presented some privacy-preserving techniques [15] avail-

TABLE 1. Privacy policies comparison

	CCPA	eDP + GDPR	PIPL
Year	2020	2002 / 2018	2021
Under protection	People born in California	People born or living in EU	People born or living in China
Target	For-profit companies (with conditions)	For-profit and non-profit national or private	For-profit and non-profit national or private
Area of effect	State of California	World	World
Affects cookies	Yes	Yes	Yes
Affects fingerprinting	Yes	Yes	Yes
Requires active consent (reject by-default)	No	Yes	Yes
Forces service without consent	No	Yes	Yes
Opt-out available	Partially (only data selling processes)	Yes	Yes
Cookie maximum living period	Undefined	Per country (usually 12 to 24 months)	Undefined
Policies on automated algorithms (e.g., price discrimination, personalized ads)	No	Yes	Yes
Limits data retention	No	Yes	Yes

able for website companies to comply with them. However, not many works include information about which data and web tracking collection methods are legally permitted, being mostly unknown by the non-expert audience. As the information obtained by means of ePrivo, the tool presented in this paper, may be useful to detect online services not complying with these regulations, we include a brief introduction to the most relevant legal aspects regarding web tracking methods to allow the reader to detect non-permitted mechanisms. We include information about the three most important privacy regulations by both, volume of users and economic impact: the USA, Europe, and China. We refer the interested reader to the regulations documentation [16]–[19] for a more detailed inspection of other legal aspects not directly related to web tracking methods. Table 1 summarizes their main differences.

1) **USA** (CCPA)

Currently, there is not yet a regulation that provides federal protection of personal information in the USA. Recently, a new law called the *American Data and Privacy Protection Act* [20] (ADPPA) aimed at this objective has been presented, but it still needs to pass the House and Senate and get White House support to prosper. However, while there is not yet a federal law, there are several state laws in place. Among the most comprehensive is the *California Consumer Privacy Act* [16] (CCPA), which took effect on January 1st, 2020, in the state of California. In 2023, a modification of the CCPA called *California Privacy Rights Act* [21] (CPRA) is expected to come into force.

The CCPA specifies the right of California's citizens to know about the personal information a business collects about them and how it is used and shared. Personal information is defined as "information that identifies, relates to, describes, is capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household". Thus, cookies and other web tracking methods are considered personal information. There are exemptions for health and financial information, which have their own specific regulations. Regarding the methods enforced to comply with the regulations, cookie consent is not a mandatory requirement for CCPA compliance. The only requirement is to provide a clear statement on the website and an opt-out for processes that sell personal information, with one year being the shortest opt-out period. Moreover, the CCPA only applies to for-profit entities doing business in California that satisfy some prerequisites about the revenue and quantity of personal information collected.

2) **EU** (ePD and GDPR)

The *ePrivacy Directive* [17] (ePD) and the *General Data Protection Regulation* [18] are the European answer to protecting its citizens' privacy. The ePD is the primary legal source on cookies' regulation through Article 5(3). It was developed in 2002 and amended in 2009 and stipulates limits to online identification algorithms regardless of whether the information stored or accessed on the user's terminal equipment is personal data or not. For consent and information requirements, the GDPR, a regulation that came into force on May 25th of 2018, applies. The ePrivacy Regulation [22] (ePR), a new regulation that supersedes the eDP, is currently being prepared to be incorporated in the near future.

ePD stipulates that websites can collect personal information as long as the users are provided with clear and comprehensive information about the purposes of the processing and are offered the right to refuse such processing (art. 5.3). The GDPR complements the ePD and delimits the scenarios where these data can be legally collected (art. 6). Moreover, it establishes that the consent must be "freely given, specific, informed, and unambiguous", and that requests for consent must be "clearly distinguishable from other matters" and presented in "clear and plain language" (art. 7). Generally, users shall have the right not to be subject to a decision based solely on automated processing when it significantly affects them, including profiling generated from collected data (art. 22). GDPR applies to all the data collectors or processors established in the EU, and to those established outside the EU but offering services or monitoring subjects inside the EU (art. 3). Regarding the data, the user has the right to

request copies, rectifications, or deletions, executing the so-called *right to be forgotten*. On those demands, the company has the obligation to review and act within one month of the request.

3) China (PIPL)

The *Personal Information Protection Law* is a new privacy-centered law prepared by China, effective since November 2021. The information presented here is extracted from the English translation done by Stanford University, available at [19]. Although not definitive, the translation is based on the second review draft of the regulations and is a good approximation of the final text.

PIPL defines personal information as any kind of information, electronically or otherwise recorded, related to an identified or identifiable natural persona within the borders of the country, even if the collection process is executed abroad. Thus, cookies and other web tracking methods are included in the regulation. For the common scenario, websites must obtain freely, voluntarily, and explicitly the consent of the user on a fully informed basis. PIPL also establishes other scenarios where personal information can be collected, including some controversial ones such as implementing news reporting, public opinion supervision, and other such activities for the public interest within a reasonable scope.

Similar to the GDPR, special treatment is given to automated decision-making algorithms, such as personalized advertising and similar systems, that must use the data in a transparent and fair way. Moreover, the user has the right not to be the target of advertisements based on an individual's characteristics. Finally, the PIPL explicitly mentions the possibility of collecting data from recognition equipment in public venues (e.g., security cameras), although it specifies that it can only be done for public security reasons.

C. RELATED WORK

Detecting online web tracking is a key aspect of discovering websites not following the principles reviewed above. It has been a hot topic for many years, and consequently, there are many systems focused on detecting web tracking algorithms. In this section, we compare ePrivo with some other existing systems that inspect and analyze different privacy aspects of websites. We do not cover here *Adblockers* or other applications used to block web tracking in real-time, as they usually do not focus on discovering web tracking elements, but on blocking already-known tracking URLs. A detailed review of the most common Adblockers can be found in [23].

PrivacyScore [24] is an online service similar to ePrivo that permits you to inspect the privacy of a selected website. It shows information about the cookies and encryption characteristics of the website, as well as some server security settings. It also explores the third parties embedded on the website, looking for already-known tracking domains. However, the system does not inspect the website at the resource or code level, only looking for information directly available from the URLs and HTTP requests loaded by the website. *PrivacyMonitor* [25] is a small browser plugin that allows to see the privacy-friendliness of the website being accessed. However, its information is based exclusively on the privacy policies of the website. Moreover, this information is obtained manually. Privacy policy changes are sent back to the experts in charge of classifying the website based on their policy statements. *PolicyXRay* [26] is a tool to automatically read policies of websites and explore the relation between the third-party tracking present inside the web and the third-parties present within the policy terms. The system focuses only on the subset of websites that use one of the already known policy content providers included in the system. None of these systems explores website privacy from the resource or website code perspective, but only on the URL or policy level. In contrast, ePrivo automatically inspects the resources included on the website, looking for website code that performs web tracking. Thus, ePrivo can be seen as complementary to current solutions.

III. EPRIVO

This section introduces the architecture of ePrivo and deepens into the two most important modules: web tracking detection and automatic classification.

A. ARCHITECTURE

ePrivo is an open source solution deployed inside three independent servers, currently running over an Ubuntu Server 20.04 LTS. Figure 1 shows the architecture of the system. The main server stays in charge of continuously collecting and updating website information for the list of websites present inside the database. The secondary server remains unoccupied until a user queries a website not present inside the database, in which case it automatically scans the new website and fills the corresponding data. The time needed to explore the website's content depends on the number and size of its included resources. The larger the number of JavaScript files loaded, the longer it will take for the web tracking detection module to inspect them. For the average website with fewer than 25 JavaScript resources of between 100Kb and 200Kb, the system takes about 2 to 3 minutes to explore all of them. Each one of the servers contains a combination of three different modules:

1) Collection module

This module is in charge of data collection. To this end, it uses ORM [27], an open source web mapping framework that explores the web as a graph containing information about each domain homepage, the URLs loaded by that domain, the online resources contained inside those URLs (e.g., JS or HTML files), and the website code inside those resources. Fig. 2 depicts the typical structure of the graph used by ePrivo. All the URL information (e.g., HTTP request, headers) and the resource information (e.g., file type, size, hash) are stored inside a MySQL database. The SHA1 hash value of the resource is used as its identifier, allowing ePrivo to deduplicate the same resource at different URLs.

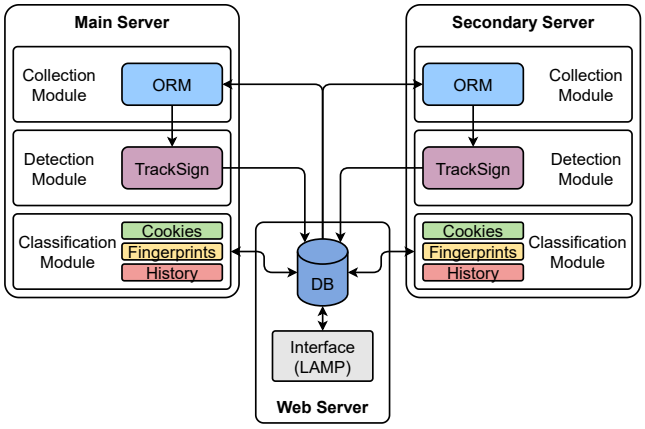


FIGURE 1. ePrivo architecture: The current architecture is based on 3 servers. The main server is in charge of constantly collecting and updating website information, the secondary server dedicated to human generated website searches and the last one to the website interface. All the website and tracking information is stored inside a common database.

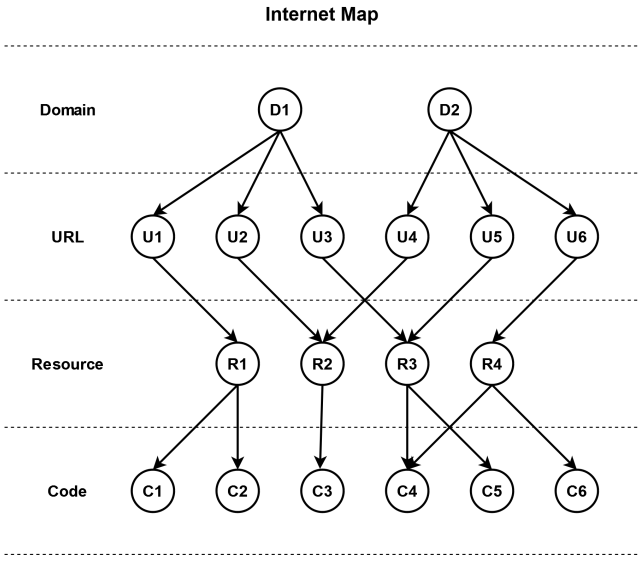


FIGURE 2. ePrivo internet map visualization. Divided in 4 different layers. Domain layer contains information about the homepage of the website. The URL layer saves the details of the URLs loaded by the domain. Resource layer deduplicates files loaded in each URL. Code layer contains information about the code pieces that compose a resource.

Moreover, the code of the JavaScript and HTML files is also stored and linked with its corresponding resource. Note that this design differs slightly from the original design proposed in [2] adding a new fourth layer to detect other forms of tracking (e.g., cookies). For security reasons, in order to avoid collapsing the database by means of attacks that automatically explore millions of dynamic or mutable subdomains, ePrivo currently only works on level 2 domains. We plan to include other countermeasures to allow for third-level domain searches. Moreover, the system uses a timeout of 30 seconds to discard slow or non-working websites.

2) Detection module

The web tracking detection is done by means of TrackSign [2], a state-of-the-art unknown web tracking detection system. The algorithm finds code shared by multiple websites with a high probability of being tracking (details in Section III-B). The information about the code pieces labeled as tracking and each resource containing them is also saved inside the database.

3) Classification module

This module is executed over the tracking information found by the detection system in order to automatically classify some web tracking systems included in the suspicious resources. It explores the JavaScript pieces of code found by TrackSign looking for already known tracking patterns. If found, ePrivo automatically categorizes the type of tracking detected according to the pattern characteristics explained in Section III-C. Note that many web tracking resources contain tracking systems not included within our initial classification modules. Thus, those resources will appear as *other tracking* in the website interface. Our future work includes exploring and developing new modules to automatically classify a bigger spectrum of web tracking systems.

4) Interface

The interface runs over a LAMP server (Linux, Apache, MySQL, and PHP), consulting directly the database for website information. Fig. 3 presents the ePrivo website front end. On the left, we can observe a graph of the URL structure of the selected website, including information about their parent-child relationship (edges) and tracking (colors). Isolated nodes are resources loaded directly by means of *XMLHTTPRequest* calls or server redirects. We can also colorize the nodes based on their file type or if they pertain to the first-party domain or a third-party website. The right column presents information about all the tracking methods detected by ePrivo and their number of appearances within the website resources, as well as some other useful information about the web. Finally, ePrivo also computes an intrusion level for each website, pondering the number of different web tracking methods used to collect personal information and the total number of each of them. The greater the number, the more intrusive the website is.

B. WEB TRACKING DETECTION

ePrivo is based on TrackSign [2], a state-of-the-art web tracking detection system that is able to detect completely unknown tracking. TrackSign works by computing heuristics for the website code in a deterministic way, looking for breaking points that are common within different files. Many tracking systems (e.g., Google Analytics) are shared by numerous web services. Thus, by looking at those code pieces shared by different domains, we can narrow down the web tracking search to the resources with the highest probability of including them. Then, accounting for the number of resources that contain the same code pieces and

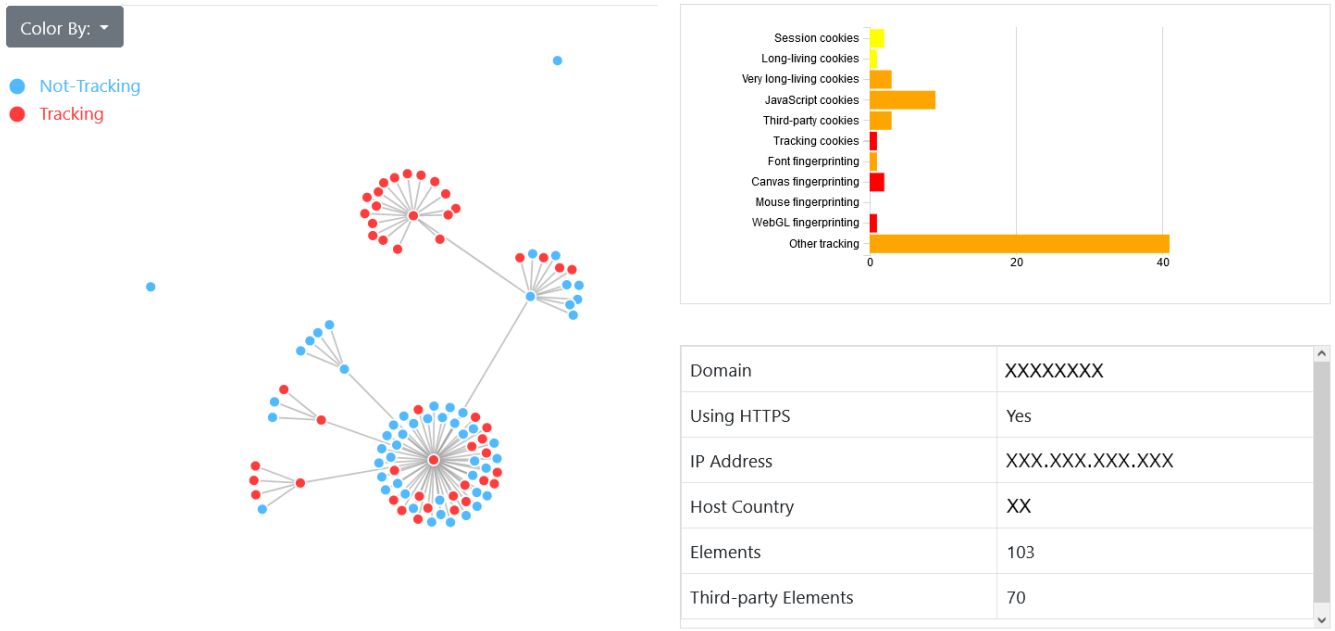


FIGURE 3. ePrivo main interface screenshot: The title states the overall intrusion level. The left graph depicts all the URLs (nodes) loaded by the website. The edges represent the parent URL from which it was called. It can be colored by resource type, resource origin (1st or 3rd party) or tracking resources. The right column shows graphically the detailed information about the web tracking performed by the website, as well as some of its characteristics.

TABLE 2. Web tracking detection comparison: TrackSign vs. uBlock Origin

	Total	Tracking (uBlock Origin)	Tracking (TrackSign)
Domains	1,383,057	984,830 (71.2%)	1,295,433 (93.66%)
URLs	76,492,732	6,246,049 (8.16%)	10,775,629 (14.08%)
Resources	5,968,622	2,369,814 (39.7%)	3,189,992 (53.44%)

are already known to include tracking, we can compute the probability of containing tracking for the selected code piece. In other words, if a piece of code is mostly present in files that perform tracking, it is most probably used for web tracking purposes. This probabilistic approach lets TrackSign propagate the tracking information, automatically labeling resources that contain the newly detected web tracking code. Thus, we can detect new resources using already-known web tracking techniques as well as new tracking systems on unknown URLs sharing the tracking code. The initial labeling process is done by means of URL tracking lists containing the most current information available (e.g., [28], [29]). We refer the interested reader to [2] for a more detailed explanation of the algorithm used by TrackSign.

According to our findings, TrackSign is able to detect thousands of unknown web tracking resources while preserving a detection accuracy of about 92%. Note that TrackSign works equally to find stateful technologies such as cookies and stateless technologies such as fingerprinting techniques. However, the detection system is not able to directly differentiate between them, as it works by looking only at the probability of the code being used for tracking or not. Consequently, we need a post-processing phase to classify the found web tracking and discover which method is being used. Thus, ePrivo uses TrackSign to find the tracking code pieces and complements it with several modules to discover the underlying web tracking methods. Table 2 compares the tracking found by TrackSign and uBlock Origin, one of the most popular methods based on the traditional pattern lists, over almost 1.4 million websites. The results show that TrackSign is able to propagate the tracking information and find millions of unknown tracking URLs and resources.

C. WEB TRACKING AUTOMATIC CLASSIFICATION

In order to discover which tracking methods are being used to obtain user personal information, ePrivo automatically classifies some of the most common web tracking algorithms. It analyzes the JavaScript files detected by TrackSign as well as the HTTP network requests in order to classify three different web tracking methods: (i) cookie-based algorithms, (ii) some common fingerprinting-based methods, and (iii) the browsing history tracking. As already mentioned, we expect to progressively add support for more methods. This section presents each of the automatically classified algorithms as well as the rules and patterns followed to do it.

1) Cookie-based algorithms

Historically, the only way to set up cookies before the creation of client-based languages such as JavaScript was by adding the flag *set-cookie* inside the network headers of the HTTP request. Nowadays, another common way to create and consult cookies is by means of JavaScript API calls.

Thus, to detect cookie usage, we inspect the network headers of all the online resources loaded by a website and the JavaScript file's code, looking for those API calls. Whenever we find one, we also check the expiration date or maximum age of the cookie and the domain that created it. Based on the information obtained, we divided the cookie tracking by type into 6 groups:

- **Session cookies**: Non-persistent cookies that are automatically deleted when the browser is closed. This category is the more privacy-friendly web tracking method, as the amount of information that can be collected is limited to the current session.
- **Long-living cookies**: Long-living cookies are characterized by having a lifespan longer than three months. Maintaining cookies for a period longer than 3 months is a bad practice, as the website is constantly receiving information about the user, enabling them to create accurate profiles of their users' habits.
- **Very long-living cookies**: We consider a cookie to be excessively long-living when it is persistent for a period over 1 year. Moreover, as shown in Table 1, it goes against some country regulations, like, for instance, France [30], where no analytic cookie can have a living period of 13 or more months. Thus, this kind of cookie can incur fines for the companies utilizing them.
- **JavaScript cookies**: Cookies mainly setup or obtained using JavaScript code. JavaScript permits to not only recover the information already present in the cookie but to regenerate cookies based on other information stored in the computer (the so-called *evercookies*).
- **Third-party cookies**: All cookies created by a domain different from the one accessed by the user are considered third-party cookies.
- **Tracking cookie**: A tracking cookie corresponds to a long-living or very long-living cookie pertaining to a third-party website. This kind of cookie allows third-party services to track each user on every website where the third party is present, allowing them to acquire a huge amount of personal information.

2) Fingerprinting-based algorithms

ePrivo automatically classifies four intrusive and precise fingerprinting methods:

- **Font fingerprinting**: Font fingerprinting is based, as its name implies, on fonts. This tracking method obtains information about the fonts installed on the target device in order to generate a unique identifier. By default, several fonts come pre-installed inside the operating system. However, it is not uncommon that other software automatically adds its required fonts at the time of installation. Thus, since it is unusual that two different people have exactly the same software and the same versions installed on their devices, the available fonts form a pattern that can be used to differentiate their hosts. To detect font fingerprinting, ePrivo looks for website code that tries to check for available fonts. The most common font fingerprinting method is by means of trial and error, as the browser uses a default font if the requested one is not present in the system. The font fingerprinting code requests a list of different fonts and checks their availability by means of two rendering API calls (.offsetHeight and .offsetWidth). Based on the analysis performed from the results obtained and a manual validation, we determined that if a resource makes thirty or more font checks while executing a call to at least one of the two commented API endpoints, the domain is performing font fingerprinting.
- **Canvas fingerprinting**: Canvas is an HTML element introduced in HTML5 that allows to load images, create drawings, and create charts dynamically in a web browser. Canvas fingerprinting takes advantage of the element, abusing its power and generating an identifier that allows to identify the device used. In order to detect it, we look for the API call .toDataURL(). This API call returns a URI containing information about the canvas element (e.g., size of the text, color, distance between characters) and how it was rendered. This information is different for each device, as it greatly varies depending on the hardware and installed software. Even a software update can render changes inside the data. Thanks to this information and how sensitive it is to small changes, it is possible to track client devices.
- **Mouse fingerprinting**: Mouse fingerprinting is a tracking method that abuses the possibilities of the browser to obtain information about the pointer. This information can be used to perform static performance tests or inspect usability problems. However, it can also be used to track the user, as every person has a slightly different way of using the pointer (usually a mouse, although trackpads also fit into this tracking technique). In order to detect it, ePrivo searches functions listening for pointer actions like *click*, *scroll*, *hold*, *drag*, *release* and so on. Those functions are called *listeners*, and they are the common way to collect mouse information. By looking for those listeners, we can detect resources collecting mouse information. To differentiate between legitimate and non-legitimate mouse data collection, ePrivo inspects if the collected mouse information is sent back to a third-party domain, a common sign of fingerprinting systems.
- **WebGL fingerprinting**: WebGL fingerprinting is a tracking technique that obtains information about the GPU and CPU present in the device by means of the WebGL API engine and two globally defined WebGL variables. The variables contain information about the exact version of the GPU and the browser, while API functions results differ depending on the GPU present in the device. ePrivo looks for the variable appearance and automatically labels the code accessing it as WebGL fingerprinting. Regarding WebGL API functions, they can be used to legitimately render figures on websites.

However, it is uncommon to use a high number of different API calls to render images, as usually with only a few of them you can obtain the desired results. Thus, our hypothesis was that using many of them on the same website was indicative of WebGL fingerprinting. From the top 10k most popular websites, we manually inspected the subset of JavaScript files containing a high number of API calls (27 files) and found that all websites calling more than 60 API calls performed WebGL tracking. Consequently, ePrivo looks to see if at least 60 of the 88 available functions implemented are called by any website and automatically classifies them as WebGL fingerprinting.

3) Browsing History tracking

The browsing history tracking mechanism is based on collecting information about the user on as many websites as possible. To this end, the tracking company usually offers some kind of functionality that is attractive to the user and forces the website interested in integrating this new characteristic to load a resource from the tracking company. From that point on, the third-party tracker has access to all the people that visit the website. When a third-party tracker is present on many important websites, it can collect an extremely accurate browsing history of each user. The most representative examples are social networks. It is mostly unknown by the common user that social networks like *Facebook*, *Twitter* or *Instagram* obtain the URL of each website accessed by the user where there is one of their social interaction buttons. Based on that information, they create profiles, grouping users (even for people that do not have an account on their services) based on their history and website content. Inspecting our visited content in detail can bring to light information such as race, age, hobbies, purchasing patterns, dislikes, or even political ideologies. This type of web tracking can be considered the most intrusive of all of them. To detect third-party trackers, ePrivo aggregates all the tracking resources linked by URLs pertaining to the same domain and counts the number of websites opening any of those resources.

IV. APPLICATIONS OF EPRIVO

This section introduces some illustrative use cases where ePrivo can help to discover privacy vulnerabilities. Each of the presented use cases focuses on a different actor potentially interested in privacy aspects.

A. INTERNET USERS

As shown in [31], about 93% of Internet users think that it is important to be in control of who can get information about them. Thanks to the new privacy regulations (e.g., Sec. II-B), currently websites usually ask for explicit user consent to collect their data by means of cookies or other web tracking systems. However, it is difficult to know if websites respect the selected privacy settings, as the communications are transparent to the user. For instance, in 2014 the *Electronic Frontier Foundation* applied the findings obtained by Acar et al. in [32] to discover that the White House website included a canvas fingerprinting web tracking system executed prior to obtaining the user's consent for data collection [33]. Similarly, by means of ePrivo anyone can explore a website and easily detect the presence of cookies that collect information by default without giving explicit consent.

B. WEB DEVELOPERS

Nowadays, many websites include external online services or load resources hosted in third-party servers to customize the browsing experience. However, these external resources may contain hidden web tracking systems unknown to the web developer. Thus, loading any of them may render the website non-compliant with their own privacy regulations without their knowledge. For instance, in the White House example presented above, the privacy policies stated literally that "*we do not knowingly use third-party tools that place a multi-session cookie prior to the user interacting with the tool*". As explained at [33] the web tracking algorithm was executed by a third-party script, and there was no evidence that the White House knew that it was being run. A web developer can use ePrivo to automatically inspect their own website and detect web tracking systems in third-party elements, helping them to maintain their websites in compliance with the current policy regulations.

C. POLICY REGULATORS

Policy regulators can use ePrivo as a preliminary filtering solution. The system can be used to detect and account for web tracking systems running in the background of any given website. Applying their experience and knowledge about non-legally permitted systems as well as studying the privacy policies stated on the website can lead them to find new websites that are non-compliant with the regulations. Moreover, the presence of the same tracking resources across multiple websites can also be used to discover unknown third-party services, including web tracking systems, that could potentially infringe the law. Finally, ePrivo also accounts for the number of websites where each domain is present as a third party, allowing policy regulators to easily discover the most pervasive actors on the Internet.

V. RESULTS

This section presents some general results extracted from the dataset available in ePrivo. The results are based on data obtained from more than 1.5 million websites selected from a combination of the *Alexa* [34] and *Majestic Million* [35] most popular websites. All the data was acquired during the period between August 2021 and February 2022.

Fig. 4 shows the popularity of all the tracking methods found by ePrivo. *Detections* correspond to the number of resources or HTTP requests that contain tracking. *Domains* are the number of domains that loaded one of those resources, or URLs. As expected, the most popular web tracking methods are based on cookies, especially *JavaScript*

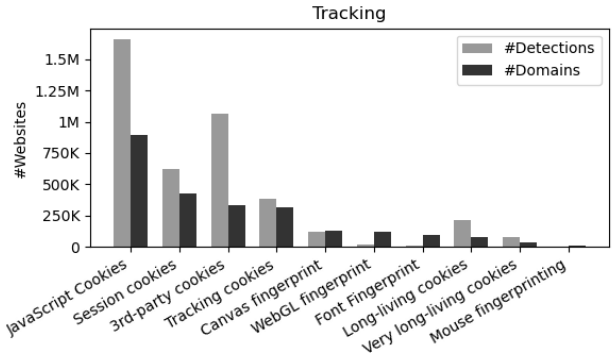


FIGURE 4. **Tracking distribution**: **#Detections** represent the number of tracking resources or tracking HTTP requests found. **#Domains** represent the number of domains that loaded one of those resources or URLs.

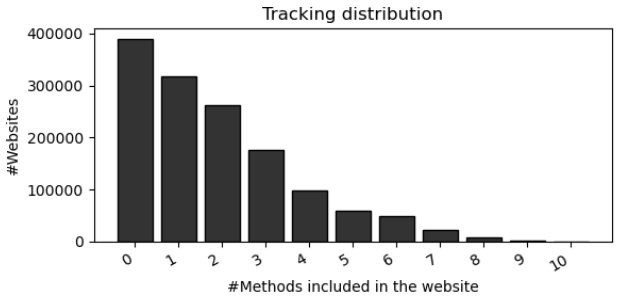


FIGURE 5. **#Trackers distribution**: The figure shows the distribution of the number of different tracking methods used by websites. Combining different methods greatly improves the identification of the user. Most websites use only a few different tracking methods. However, approximately 10% use more than half of the methods and about 0.6% use almost all of them.

cookies. The most common fingerprinting method is *Canvas fingerprinting*. Interestingly, we can clearly differentiate two patterns between cookies and fingerprinting. Cookie-based tracking techniques present a higher number of detections than the number of domains loading them. This is expected, as most websites load many cookies in the same visit. On the contrary, fingerprinting methods are usually contained in the same files, and those files are loaded by many websites. Thus, the number of files is smaller than the number of domains loading them. Note that there are almost 40k domains (2.6%) that still use *very long-living cookies*. As already commented, this kind of cookie lasts for more than one year, and they are forbidden by some regulations [30]. All the websites making use of these cookies are at risk of being pursued by privacy regulators.

Fig. 5 depicts the distribution of the quantity of tracking methods used by websites. A combination of different tracking methods usually improves the identification of the user. Thus, looking for a high number of different methods within the same website highlights the most privacy-invading services. Most websites use only a few different methods, mainly in the form of cookies. However, about 10% of websites include more than half of the inspected methods, and usually at least one or more fingerprinting methods. There are about 0.6% that load almost all the automatically classified methods (8 or more), including some very popular websites such as *wsg.edu*, *theaustralian.com.au* or *hulu.jp*. Note that these results are an underestimation, as currently ePrivo can only automatically classify a subset of the different web tracking methods, and some of them remain inside our "unknown" web tracking category. We plan on developing new automatic classification modules to decrease the quantity of unknown web tracking. Thus, this information should be revisited with each new iteration or new module deployed.

TABLE 3. Top browsing history trackers

Domain	Websites
googletagmanager.com	353160
google.com	296699
doubleclick.net	267973
facebook.net	238994
facebook.com	193217
gstatic.com	170619
cloudflare.net	127671
googlesyndication.com	125835
youtube.com	110123
googleadservices.com	102543
googleapis.com	82164
yandex.ru	75762
google-analytics.com	64885
wp.com	55154
twitter.com	43873
cloudfront.net	43741
googletagservices.com	35733
newrelic.com	33714
shopify.com	31262
hotjar.com	29986

Finally, table 3 shows the top 20 browsing history tracking domains. As expected, most of them pertain to very well-known online services present on multiple websites. Google is the predominant one, owning half of the most commonly linked domains and including different services such as analytics, advertising, tagging, and on-demand video systems. The rest of the list includes many other analytics and customized advertisement systems (e.g., newrelic.com, shopify.com, hotjar.com), as well as some social networks, with Facebook being the most common one only surpassed by the commented Google services.

VI. DISCUSSION: ON THE FUTURE OF TRACKING

New technologies as well as new actors can have a major impact on the evolution of the Internet. Depending on this evolution, we plan to adapt the web tracking detection and classification algorithms used by ePrivo to follow the new advances in the area. In this section, we discuss some interesting aspects that can alter the web tracking paradigm in the near future.

A. EVOLUTION OF TRACKING

As seen in Section V, nowadays, most websites use cookies to track users. However, it is expected that many browsers, such as *Google Chrome*, *Brave* or *Opera* will start forbidding

third-party cookies as soon as 2022. The direct consequence of this fact is a possible improvement in privacy in some cases, as nowadays third-party tracking is the preferred way to monetize content and decide marketing strategies. However, first-party trackers will continue to be allowed to set cookies. This empowers big companies like *Amazon*, *Facebook* or *Google* that force users to log into accounts to use their services.

Despite this, Google is developing a new way to allow websites to know information about their users' interests while protecting their privacy at the same time. This new approach is the so-called *Topics API* [36], where the browser stores information about the user's topics of interest for 3 weeks, and websites can access a limited amount of this information to personalize the advertisements. This new API is still being developed and may impose restrictions and improve privacy on most websites. However, it still remains to be seen if third-party companies whose market is mainly based on personal data will be significantly affected, or most probably will simply update their collecting tools, giving as a result an increase in user fingerprinting.

B. THE FALL OF PROTECTION MEASURES

As commented in Section II-A, most current privacy-preserving tools work as small browser plugins that inspect in real time URLs loaded by the website and block or modify the website code to evade tracking mechanisms (e.g., AdBlockers, JavaScript blockers). However, many browsers are implementing the so-called *Manifest v3* [37]. Manifest v3 will prevent plugins from accessing the URL information in real time, making almost all the privacy protection tools currently available completely useless. Thus, in the near future, the only way to interact with websites during loading time will be to pass the browser a list containing a maximum of 50k URLs that you want to block (regular expressions or complete domains are not allowed), and the browser will block them.

This, again, empowers the browser's position to decide whether a URL has to be blocked or not. Needless to say, letting companies whose main income comes from targeted advertisements protect you does not guarantee your privacy. Manifest v3 is already supported in *Google Chrome*, and it is being implemented in *Mozilla Firefox*. In *Chromium*, it is expected to become mandatory during 2022.

VII. CONCLUSIONS AND FUTURE WORK

This work presented ePrivo, an e-privacy observatory designed to allow users, companies, and regulatory entities to inspect the online privacy details of any website. The system automatically explores the resources loaded by a website and detects and classifies all the privacy-threatening content. The detection method is based on TrackSign, a state-of-the-art algorithm that computes the probability of each website's code being used for web tracking purposes. After the detection, the system also automatically classifies and labels the most popular web tracking methods. We also presented some use cases where ePrivo can be used to discover privacy threats or non-compliant websites from the common user's, web developer's, and privacy regulator's perspectives. Moreover, the paper reviewed the background on web tracking and discussed the most important privacy regulations, which determine the data collection processes currently allowed by different countries.

Our future work includes developing more tracking classification modules to improve the labeling process, as well as a RESTful API to be able to access the web tracking information on demand and with higher granularity. We also expect to allow for level 3 domain searches in ePrivo and better correlate the web tracking information with the policy regulations by modifying the data collection module to interact with the policy consent banners. Consequently, the system should be able to differentiate between the tracking found by accepting and rejecting the website policies. Finally, we also expect to adapt ePrivo's algorithms to the evolution of online tracking systems as a response to some new paradigms, such as *Topics API* or *Manifest v3*.

ACKNOWLEDGMENT

We thank the anonymous reviewers for their comments, which led to significant improvements to this manuscript, especially in Section II-B.

This publication is part of the Spanish I+D+i project TRAINER-A (ref. PID2020-118011GB-C21), funded by MCIN/ AEI/10.13039/501100011033. This work is also supported by the Catalan Institution for Research and Advanced Studies (ICREA Academia).

REFERENCES

[1] A. M. Research, "Web Analytics Market Size, Share and Industry Analysis | Forecast 2026," Jul. 2020. [Online]. Available: https://www.alliedmarketresearch.com/web-analytics-market-A05971

[2] I. Castell-Uroz, J. Solé-Pareta, and P. Barlet-Ros, "TrackSign: Guided Web Tracking Discovery," in *Proc. 2021 IEEE Int. Conf. in Computer Communications (INFOCOM)*. San Diego, CA: IEEE, 2021.

[3] T. Bujlow, V. Carela-Español, J. Solé-Pareta, and P. Barlet-Ros, "A Survey on Web Tracking: Mechanisms, Implications, and Defenses," in *Proc. IEEE*, vol. 105, no. 8, Aug. 2017, pp. 1476–1510.

[4] "e-Privacy Observatory," May 2022. [Online]. Available: http://www.eprivo.eu/

[5] R. Hill, "uBlock Origin," Feb. 2020. [Online]. Available: https://github.com/gorhill/uBlock

[6] AdBlock Plus, "Adblock Plus," Feb. 2020, https://adblockplus.org/en/. [Online]. Available: https://adblockplus.org/en/

[7] Ghostery, "Ghostery Makes the Web Cleaner, Faster and Safer!" Feb. 2020. [Online]. Available: https://www.ghostery.com/

[8] M. Ikram, H. J. Asghar, M. A. Kaafar, A. Mahanti, and B. Krishnamurthy, "Towards Seamless Tracking-Free Web: Improved Detection of Trackers via One-class Learning," in *Proc. Privacy Enhancing Technologies*, vol. 2017, no. 1, Jan. 2017, pp. 79–99.

[9] U. Iqbal, P. Snyder, S. Zhu, B. Livshits, Z. Qian, and Z. Shafiq, "ADGRAPH: A Graph-Based Approach to Ad and Tracker Blocking," *IEEE Symp. Security and Privacy 2020*, p. 14, 2019.

[10] U. Iqbal, S. Englehardt, and Z. Shafiq, "Fingerprinting the Fingerprinters: Learning to Detect Browser Fingerprinting Behaviors," in *Proc. IEEE Security and Privacy (S&P'21)*, Virtual, May 2021, pp. 1143–1161.

[11] H. Li, L. Yu, and W. He, "The impact of gdpr on global technology development," *Journal of Global Information Technology Management*, vol. 22, no. 1, pp. 1–6, 2019.

[12] R. N. Zaeem and K. S. Barber, "The effect of the gdpr on privacy policies: Recent progress and future promise," *ACM Trans. Manage. Inf. Syst.*, vol. 12, no. 1, dec 2020.

[13] J. Sørensen and S. Kosta, "Before and after gdpr: The changes in third party presence at public and private european websites," in *Proc. 2019 The World Wide Web conference*. New York, NY, USA: Association for Computing Machinery, 2019, p. 1590–1600.

[14] C. Utz, M. Degeling, S. Fahl, F. Schaub, and T. Holz, "(un)informed consent: Studying gdpr consent notices in the field," in *Proc. 2019 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 973–990.

[15] N. Gruschka, V. Mavroeidis, K. Vishi, and M. Jensen, "Privacy issues and data protection in big data: A case study analysis under gdpr," in *Proc. 2018 IEEE Int. Conf. on Big Data*, 2018, pp. 5027–5033.

[16] "California Consumer Privacy Act (2020)," May 2022. [Online]. Available: https://www.oag.ca.gov/privacy/ccpa

[17] "ePrivacy Directive," Jul. 2002. [Online]. Available: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32002L0058:en:HTML

[18] "General Data Protection Regulation (2018)," May 2022. [Online]. Available: https://gdpr.eu/

[19] S. University, "Translation: Personal Information Protection Law of the People's Republic of China (effective nov. 1, 2021)," Oct. 2021. [Online]. Available: https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021/

[20] "American Data Privacy and Protection Act," Jun. 2022. [Online]. Available: https://www.congress.gov/bill/117th-congress/house-bill/8152/text

[21] "The California Privacy Rights Act of 2020," Nov. 2020. [Online]. Available: https://thecpra.org/

[22] "Proposal for an ePrivacy Regulation," Jan. 2017. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/eprivacy-regulation

[23] J. Mazel, R. Garnier, and K. Fukuda, "A comparison of web privacy protection techniques," *Computer Communications*, vol. 144, pp. 162–174, Aug. 2019.

[24] "PrivacyScore," May 2022. [Online]. Available: https://privacyscore.org/

[25] "PrivacyMonitor," May 2022. [Online]. Available: https://www.privacymonitor.com/

[26] T. Libert, "An Automated Approach to Auditing Disclosure of Third-Party Data Collection in Website Privacy Policies," in *Proc. 2018 World Wide Web Conference*, ser. WWW '18. Lyon, France: International World Wide Web Conferences Steering Committee, Apr. 2018, pp. 207–216.

[27] https://github.com/CBA-UPC/ORM, "Online Resource Mapper (ORM)," Jun. 2020. [Online]. Available: https://github.com/CBA-UPC/ORM

[28] "EasyPrivacy," May 2022. [Online]. Available: https://easylist.to/easylist/easyprivacy.txt

[29] "EasyList," May 2022. [Online]. Available: https://easylist.to/

[30] "Délibération n° 2020-092 du 17 septembre 2020 portant adoption d'une recommandation proposant des modalités pratiques de mise en conformité en cas de recours aux « cookies et autres traceurs »," Sep. 2020. [Online]. Available: https://www.cnil.fr/sites/default/files/atoms/files/ligne-directrice-cookies-et-autres-traceurs.pdf

[31] "American views about data collection and security," Sep. 2022. [Online]. Available: https://www.pewresearch.org/internet/2015/05/20/americans-views-about-data-collection-and-security/#few-feel-they-have-a-lot-of-control-over-how-much-information-is-collected-about-them-in-daily-life

[32] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz, "The Web Never Forgets: Persistent Tracking Mechanisms in the Wild," in *Proc. 2014 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '14. Scottsdale, Arizona, USA: Association for Computing Machinery, Nov. 2014, pp. 674–689.

[33] "White House Website Includes Unique Non-Cookie Tracker, Conflicts With Privacy Policy," Jul. 2014. [Online]. Available: https://www.eff.org/deeplinks/2014/07/white-house-website-includes-unique-non-cookie-tracker-despite-privacy-policy

[34] K. Cooper, "Alexa: Most popular website list," Jun. 2020. [Online]. Available: https://www.alexa.com/

[35] "The majestic million," Jun. 2022. [Online]. Available: https://majestic.com/reports/majestic-million

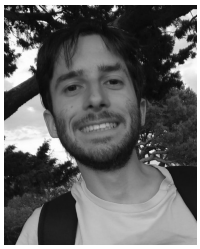
[36] "The Topics API," Jan. 2018. [Online]. Available: https://github.com/patcg-individual-drafts/topics

[37] "Manifest v3," May 2022. [Online]. Available: https://developer.chrome.com/docs/extensions/mv3/intro/

ISMAEL CASTELL-UROZ (ismael.castell@upc.edu) is a Ph.D. student at the Computer Architecture Department of the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, where he received the B.Sc. degree in Computer Science in 2008 and the M.Sc. degree in Computer Architecture, Networks, and Systems in 2010. He has several years of experience in network and system administration and currently holds a Projects Scholarship at UPC. His expertise and research interest are in computer networks, especially in the field of network monitoring, anomaly detection, internet privacy and web tracking.

ISMAEL DOUHA-PRIETO (ismael.douha@estudiantat.upc.edu) received the B.Sc. degree in Computer Science in 2021 and the Master's degree in Cybersecurity in 2021, both at the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain. He is currently working as a Cloud Security Engineer at NTT Data and his main research interest are in cloud, cybersecurity, and privacy. He was a core developer of the e-Privacy Observatory.

MERITXELL BASART-DOTRAS (meritxell.basart@estudiantat.upc.edu) received the M.Sc. in Cybersecurity in 2021, the B.Sc. in Network Engineering in 2020, and the B.Sc. in Aerospace System Engineering in 2020, at Universitat Politècnica de Catalunya (UPC), Barcelona, Spain. She is currently working as a Full Stack Developer at Kongsberg Maritime in Trondheim, Norway. Her research interests are in cybersecurity, computer networks, and internet privacy.



POL MESEGUÉ-MOLINA (pol.mesegue@estudiantat.upc.edu) is a M.Sc. degree student in Privacy and Cybersecurity at Universitat Oberta de Catalunya (UOC) and received the B.Sc. degree in Computer Science in 2021 at the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain. His interests are related to cybersecurity, internet privacy and decentralized networks. He was a front-end developer of the e-Privacy Observatory.

PERE BARLET-ROS (pere.barlet@upc.edu) is a Full Professor with the Computer Architecture Department of the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, and Scientific Director at the Barcelona Neural Networking Center (BNN-UPC). He received the M.Sc. and Ph.D. degrees in Computer Science from UPC, in 2003 and 2008, respectively. From 2013 to 2018, he was Co-founder and Chairman of the machine learning startup Talaia Networks. His research has been integrated in several open-source and commercial products, including Talaia Polygraph, Auvik TrafficInsights, Intel CoMo and SMARTxAC. In 2015, he was awarded as the best entrepreneur of the UPC School of Informatics (FIB). More recently he has been awarded by the ICREA Academia Programme (2023).