# Multi-country Study of Third Party Trackers from Real Browser Histories

Xuehui Hu
*King's College London*

Guillermo Suarez de Tangil
*King's College London*

Nishanth Sastry
*University of Surrey*
*King's College London*
*The Alan Turing Institute*

***Abstract*—This paper aims to understand how third-party ecosystems have developed in four different countries: UK, China, AU, US. We are interested in how wide a view a given third-party player may have, of an individual user's browsing history over a period of time, and of the collective browsing histories of a cohort of users in each of these countries. We study this by utilizing two complementary approaches: the first uses lists of the most popular websites per country, as determined by Alexa.com. The second approach is based on the real browsing histories of a cohort of users in these countries. Our larger continuous user data collection spans over a year. Some universal patterns are seen, such as more third parties on more popular websites, and a specialization among trackers, with some trackers present in some categories of websites but not others. However, our study reveals several unexpected country-specific patterns: China has a home-grown ecosystem of third-party operators in contrast with the UK, whose trackers are dominated by players hosted in the US. UK trackers are more location sensitive than Chinese trackers. One important consequence of these is that users in China are tracked lesser than users in the UK. Our unique access to the browsing patterns of a panel of users provides a realistic insight into third party exposure, and suggests that studies which rely solely on `Alexa` top ranked websites may be over estimating the power of third parties, since real users also access several niche interest sites with lesser numbers of many kinds of third parties, especially advertisers.**

***Index Terms*—browser privacy, web tracking, cookies**

## 1. Introduction

Web advertising has evolved considerably over the last decade. Publishers and advertisers currently leverage Web technology to track users' browsing histories and make advertising even more targeted, and therefore more profitable. This is supported by many of the most popular websites that willingly embed this technology into their sites to monetize the content they host. This technology basically allows publishers to obtain a unique identifier of the visitor of a site, which is then used to match the user across other websites. Although there are many ways in which a tracker can associate unique identifiers to visitors, current efforts are largely based on the DART (Dynamic Advertising Reporting and Targeting) initiative launched by DoubleClick [1]. Here, unique third-party cookies are left in the browser of the user when she visits a website with a tracker embedded. With the scaled size of the advertisement industry, this poses a risk to the privacy of the users and leads their browsing history being shared in some shape or form with publishers and advertisers.

Related works in the area have recently looked at this problem and they have provided a good understanding on how the tracking technology works and the underlying privacy issues [2]–[5], including mobile tracking [6]. However, their analysis only look at a slice of the problem: either because they look at a specialized third-party network [7], or because they look at the problem from a holistic perspective without considering users' browsing patterns [8]. For instance, authors in [3] look at advertisements alone, [2] and [8] do not look at the population segmentation, and [5] does not quantify how third-party categories change over time. A central aspect of understanding the tracking ecosystem is characterizing the different trackers users came across. This is a challenging process as the third-party ecosystem is complex, highly dynamic, and — in some cases — localized [9]. Our study differs from other works in the scope of our analysis. Here, we consider general-purpose third-party domains with a fine-grained categorization. We also consider, as a key distinction, targeted population segments (e.g., Chinese users) in different locations (i.e., domestic users vs. users abroad). We not only provide a comparison of third parties across countries and categories, but also insights into the causes of differences based on a broader data collection.

To better understand the magnitude of the tracking problem, we present the following main contributions. First, we build technology that can capture to what extent third-party trackers are profiling users as they browse the Web. We do this by extending a popular Firefox extension, called Lightbeam, in two directions: i) enabling the support of fine-grained cookie logging and porting it to Chrome, and ii) integrating an automated browsing system into it. Our extension is available in the Chrome Store as the "Thunderbeam-Lightbeam for Chrome" plugin[1] and has seen 2,384 installs (as of May 31, 2020). Second, we provide an improved categorization (15% improvement) of the type of third-party providers by employing a number of heuristics and using several online resources. We freely make available the resulting "Tracking the Trackers categorization list"[2]. Finally, we study the interplay between user location and the overall number of third parties observed using a twofold approach: with an automated controlled experiment and a user study. In particular, we look at third-party domains in five different population

[1] https://tiny.cc/lightbeam-chrome-plugin
[2] https://tiny.cc/tracking-trackers-list

segments: Australia (AU) users, United States (US) users, United Kingdom (UK) users in UK, Chinese domestic users (CN), and Chinese users located in UK (CN-UK). We aim at answering the following research questions, which we present together with our main findings:

**RQ1: Is the number of trackers per site affected by the popularity of the website, as well as its category?** §3 demonstrates specialisation of third parties across categories. Thus, third party actors are more easily able to track individual users (who fit their specialisation areas) across time, than a diverse cohort of users simultaneously.

**RQ2: Are there country-specific third parties?** §4 finds specialized actors that track users only in a given location (e.g., CN but not UK or vice versa). In contrast with UK websites, whose third party providers are mostly US-based, CN is dominated by local third party providers.

**RQ3: Do all countries experience the same amount of tracking?** §4 also shows UK users are tracked *more* than in China — the dominance of players like Google results in individual players obtaining a large coverage of users' browsing patterns. China's third party ecosystem is more decentralized; which results in diminished visibility and coverage of individual third parties. However, there are *fewer* social third parties targeting UK users than CN users. Also, we observe that Google manages to obtain non-trivial, albeit diminished, coverage of users in China through some of its domains which are not blocked.

**RQ4: Do trackers use traffic discrimination?** We find websites have dynamic strategies to load different trackers over time (§3). These strategies are segmented based on the location the user connects from and the type or category of website the user connects to. By combining a more common study of different websites based on Alexa rankings with a study of real browsing histories of a panel of users, we come to the conclusion that Alexa based studies may be systematically over estimating the amount of tracking that individual users may experience.

## 2. Data collection methodology

Our study develops two new datasets to explore the third-party ecosystem: one based on real-world browser histories and another one based on the most popular Alexa websites. The first dataset looks at three user groups (UK users in UK, Chinese users in China, and Chinese users in UK) and we compare the structure of third party networks across groups. This is complemented by users from two other countries, Australia and USA. The China and UK data is longitudinal for over a year, and the Australia and US data represents nearly one month of activity. This dataset is used to answer RQ2 and RQ3 in §4.

Second, we contextualize our findings by looking at the prevalence of third parties in some of the most popular sites (according to Alexa) in an automated fashion. The rationale behind the second experiment is to understand how representative our first dataset is. We show this in Fig. 1, which depicts the cumulative distribution functions of the data by country. We observe that the distribution of the numbers of third parties per domain in our user dataset matches that of the topsites dataset for each country. This gives us confidence that the data obtained from our user group maps with behaviors of each of the populations. This second dataset answers RQ1 and RQ4 in §3. We access the Alexa top websites using Selenium in non-headless mode, to simulate user activity. This enables the collection of connections between first parties and all dynamic third parties on an active browser. We repeat the test 25 times per day over 7 days to confirm that change in numbers of dynamic TPs is less than 1% (c.f. §3.3).

In this section, we first describe the design of a browser plugin which collects our data (§2.1). We then describe how we categorize URLs, including new heuristics and improvements to deal with paucity of data about Chinese websites (§2.2), as well as methods employed for merging or distinguishing third parties (§2.3). We then discuss ethics considerations (§2.4), before giving an overview of the data we have collected (§2.5).

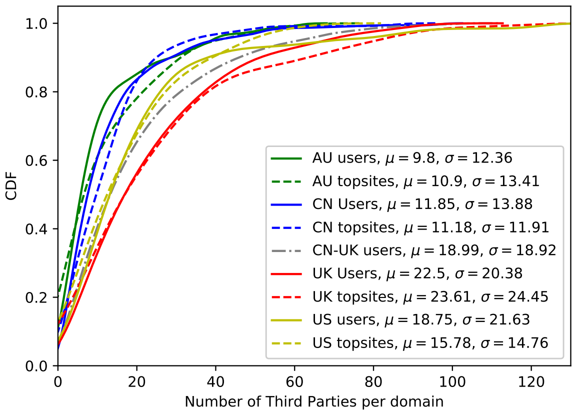


Figure 1. Number of third-party sites per domain in User Group data (*real users*) follows a similar distribution to the numbers of third parties on Alexa top500 (*topsites*) in each country. *UK users also have more third parties per domain than CN (China) and AU (Australia). CN-UK users visit both CN and UK websites, and also without the Great Firewall of China, and thus experience an intermediate number of third parties between purely CN and purely UK users. Surprisingly, both data from US users and US topsites show the fewer third-party number than UK.*

### 2.1. Data collection using browser plugin

For the data collection, we extend a popular Firefox extension called Lightbeam [10]. As detailed in §2.5, our add-on is installed both in an automated browsing system over OpenWPM [11] which we use for automated browsing of websites, as well as in the browsers of all the members of our user group (with University Ethics Approval). Many of our CN and UK user group prefer to use Google Chrome, so we also extended the Firefox extension to work on Google Chrome, to fit unobtrusively into the browsing habits of our users. Making the extension work on Google Chrome involved tackling several challenges, which we outline below.

Lightbeam not only allows users to log their browsing history, but it also provides support to track third-party networks *across* sites. Lightbeam was based on an add-on called Collusion developed by Mozilla in 2012 [12]. Branching from Collusion, there is an add-on called Disconnect [13] for Google Chrome browsers. However, as opposed to Lightbeam, Disconnect can only log trackers in an uncontextualized manner (i.e., by looking at websites individually). Thus, Disconnect does not support tracking third parties across sites since there is no direct mechanism

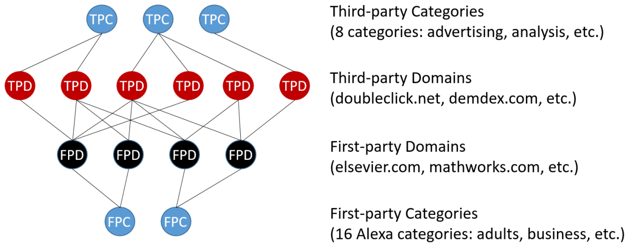

TPC
TPD
FPD
FPC
Third-party Categories
(8 categories: advertising, analysis, etc.)
Third-party Domains
(doubleclick.net, demdex.com, etc.)
First-party Domains
(elsevier.com, mathworks.com, etc.)
First-party Categories
(16 Alexa categories: adults, business, etc.)

Figure 2. Tracking Model: Users visit First Party Domains (FP or FPD) who then load third party domains (TPD or TP). FPs are categorised into sixteen categories based on `Alexa`. TPs are categorised using eight different lists of third party blockers, and further enhanced by a list of 1,685 manually verified third parties compiled by us.

to capture and match the correspondence of first-party and third-party requests in Chrome.

Lightbeam does not work on Google Chrome because the mechanism that Lightbeam uses to tie the third-party requests to the first parties that initiate the loading of those third parties is only supported on Firefox. Specifically, when a HTTP request is made to any website, the WebRequest object in Firefox contains a property called `webextensions.api.webRequest.onBefore-Request.details.originUrl`. In a third-party HTTP request on Firefox, the `originUrl` gives the details of the first party that initiated the request. Unfortunately, `originUrl` is not supported on Google Chrome, which makes it difficult to tie third-party requests back to a given first party. We get around this restriction by noting that the `WebRequest` API on Chrome supports a `tabId` field, which identifies which tab made a request. We build and maintain a table of all the opened tabs, and use the URL loaded by the tab to obtain the first party information. We also use this table in both the Firefox and Chrome versions to build a better categorisation of first party and third-party websites than the "vanilla" Lightbeam plugin, similar to previous research [14]. In particular, we reduce errors in detecting third parties by relying on heuristics that better characterize the identity of the server [15], and using the correspondence between *tabIDs* to minimize the number of first-parties mistakenly classified.

## 2.2. Improved categorization database

Our modified Lightbeam plugin allows us to extract a relationship as shown in Fig. 2 between first-party domains (FPD) that a user visits and the third-party (TPD) domains that are loaded by those first parties. Notice the same third party (e.g., DoubleClick for ads, or Facebook for the 'like' button) may be loaded by more than one first party. Thus, third parties are in a uniquely privileged position to infer the *first party* browsing habits of users over time, or to understand the overlaps in browsing habits of cohorts of users.

To understand which *kinds* of third parties have this visibility and on which kinds of first party websites, we create first party and third party categories (FPC and TPC in Fig. 2). We categorise the first party sites into sixteen categories using `Alexa`'s categories — all but the 'world' category, as listed at: https://www.alexa.com/topsites/category. `Alexa` publishes the top 500 domains for each category. But category information is also available for other websites from the `siteinfo` page for that site (https://www.alexa.com/siteinfo/DOMAINNAME). We categorise third parties by their functions, as follows:

**Advertising Third Parties** have two subcategories: i) ads or services from other first-party business (e.g., sites from section), and ii) ads from third-party networks (e.g., *doubleclick.net*, *adkernel.com*, or *webspectator.com*).

**Analysis Third Parties** are services provided by web analytic companies. (For example, *google-analysis.com*, *hupso.com*, and *audienceinsights.net*.)

**Essential Third Parties** are domains that work as essential features of the site, such as the secure log-in, cloud storage of website resources, etc. (e.g., *bootcss.com*, *squixa.net*, and *commandersact.com*.)

**Malware Third Parties** are sites proven to be responsible of severe data breaches, containing adware, viruses, potential ransomware, etc. (Examples include: *msecnd.net*, *imrworldwide.com*, and *securestudies.com*.)

**Optimization Third Parties** provide services to optimize user experience like: supporting higher speed, automated and interactive marketing, etc. (E.g., *yieldoptimizer.com*, *maxymiser.net*, *bzgint.com*.)

**Redirect Third Parties** are domains that use HTTP redirection to divert the user from one website to another. Some of them are URL shorteners and third-party payment services. (e.g., *redirectingat.com*, *tinyurl.com*, and *clickredirection.com*.)

**Social Third Parties** refers to social broadcasting, news, plugin, media, etc. (For example, *metabroadcast.com*, *buzzfeed.com*, and *twitter.com*.)

**Tracking Third Parties** include all forms of tracking domains through embedded technology, web bugs, information collectors for providing third-party providers with the customer data. (e.g., *otracking.com*, *tctm.co*, and *zenfs.com*.)

We obtain this classification by merging eight different lists from different sources, as listed in Table 3. As might be expected, different lists may use slightly different names. We merge these to obtain the uniform nomenclature as listed above. Different lists mostly agree about the categories for specific third parties. Where there is a disagreement (e.g., a particular third party is categorized as advertising by one list, and as malware by another list), we use a majority rule to disambiguate. However, note that some third parties may legitimately serve multiple functions (e.g., Google may be in advertising and analytics; Facebook in social, tracking, and advertising). Thus, we count a third party in all relevant categories, unless a majority of lists overrules by providing a definitive single category. We further reduce errors on the detection of TPDs by relying on a number of heuristics that better characterize the identity of the server [15].

Finally, to get around limitations of current Web tracker lists (which are especially poor for Chinese websites [16]), we also manually classify TPDs which are not in any of the known lists. Our manual categorization involves visiting the website of the third party vendors to check the "Home" and "About" pages, checking the JavaScript used by the third parties, and querying Open Source Intelligence for vulnerable/malicious indicators. Overall, we identify a significant number of third-party domains (1,685). We refer to this manual annotation as "Ourlist" in Table 3.

## 2.3. Disambiguation of Third Parties

In estimating how much of a user's browsing history a single entity may be privy to, it is insufficient to just use the domain name of the third party — a single entity may simply employ multiple domain names, either to explicitly hide the extent of tracking, or as a result of organic discrepancies arising in domain name usage (such as due to mergers). For instance, Google owns multiple other domain names such as `doubleclick.net` and `google-analytics.com`. To disambiguate such cases, we follow previous work [17] and merge third parties if they are controlled by the same Authoritative DNS Server (ADNS).

Secondly, in estimating the loss of privacy, we also take into account possible data sharing through cookie synchronization [18] as a mechanism to establish a "data sharing tunnel" between different third-party vendors. Cookie synchronization can be detected by correlating unique userIDs embedded in cookies stored by different third parties. In this paper, we apply a cookie synchronization model as described in Appendix A. Our methodology follows the guidelines given in [18], although we also consider 8-character string for the length of shared userIDs. This is because we experimentally observed cookie synchronization over values of that length.

## 2.4. Research Ethics

We ensure our research is ethical by following the guidelines from The Belmont Report [19]. First, we do not request personally identifiable data such as name, nor do we collect information that could be use to identify users like IP addresses. We do not collect any sensitive information (e.g., age or gender) except for the browsing history. Out of all the data one can extract from the browsing history, we only retain first-party and third-party domains and the corresponding categorization of the URL. This means that we do not collect URL parameters, which may include username, passwords or other identifiable information.

We have taken steps to guarantee users: 1) willingly share their anonymized browsing history with us, 2) understand the purpose of this study, and 3) know their rights. The information sheet and consent form provided to participants of our real-user data collection have been reviewed by our Institutional Review Board.[3] Furthermore, our released Chrome extension provides a Privacy Policy with information about user's rights, including withdrawal.

## 2.5. Dataset

Using our modified Lightbeam plugin and our categorisation database, we have collected two datasets. The first one is obtained after collecting anonymized browsing histories through our Lightbeam add-on from 16 different users weekly between January 5, 2018 and January 2019 — 9 users in the UK and 7 users in China (CN). Here, 3 of the 9 UK users are of Chinese origin and tend to also visit Chinese websites from the UK. We term these users as CN-UK, and they offer a unique perspective on the tracking done by Chinese websites to users out of China. In a second stage, four more users from Australia (AU) and United States (US) enrolled in our study via the Chrome Store. These users also provide data on a weekly basis for a period of three weeks.

TABLE 1. FIRST-PARTY (TP) AND THIRD-PARTY (TP) DATA COLLECTED FROM OUR PARTICIPANTS.

| User Group | Records of FP | Records of TP |
|---|---|---|
| UK users | 8416 | 113,003 |
| (include CN-UK users) | (2680) | (36,209) |
| CN users | 6144 | 74,313 |
| US users | 392 | 4450 |
| AU users | 104 | 820 |
| Total | 15,323 | 192,586 |

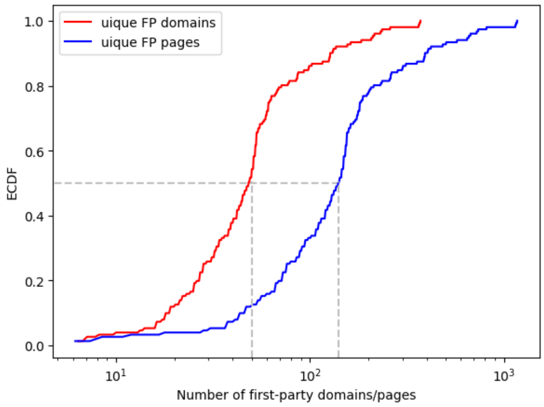


Figure 3. CDF of the number of unique first-party (FP) web domains/pages visited weekly overall.

All participants stay online for more than 5 hours a day, more than 5 days a week. Table 1 shows an overview of the data collected from these users. In total, we have gathered 11,232 unique third party domains from the browsing logs of our test users in the first stage, between Jan 5, 2018 when the collection started, and Jan 1, 2019, the last date reported in this paper. Of these, 1,685 sites were manually identified as described above, yielding a 15% improvement over the union of previously known lists. In total, we are able to successfully categorise nearly 90% of third parties for the UK users and 70% of the third parties for Chinese users. And after the second stage (publish of our extension), we have gained over 15,323 first parties and more than 19k third-party records from four countries (continents).

To understand the browsing habits of our users, we measure their browsing activity by looking at the average number of visits per week. In particular, we look at the number of pages (URLs) visited and the unique number of second-level domain names they connect to. Fig. 3 shows that our users are all active browsers, visiting tens of sites per week and multiple URLs per domain (about three URLs per domain on average). The trend displayed is uniform but diverse, with the bulk of the users

[3] The consent process has been vetted by the King's College London Research Ethics Committee. The criteria for approval can be found here: https://bit.ly/2XHiwT8

TABLE 2. TOP5 FIRST-PARTY DOMAINS IN OUR REAL-USER DATABASE

| UK | China | Australia | US |
|---|---|---|---|
| {college.ac.uk} | tmall.com | aliexpress.com | google.com |
| google.com | baidu.com | google.com | {college.edu} |
| microsoft.com | weibo.com | renren.com | linkedin.com |
| wordpress.com | taobao.com | ebay.com.au | wix.com |
| stackexchange.com | qq.com | youtube.com | pinterest.com |

browsing like the average and some users browsing either a slightly smaller or large number of sites. The dashed lines point to the average browsing habits of half of the users. Half of the users visit at least 50 unique first-party domains and 140 unique web pages each week. Table 2 summarizes the most frequently visited first-party domains per country. A key distinction on the way we compute the frequency of visited domains with respect to `Alexa.com` is that we look at second-level domain names.[4] Instead, Alexa looks at third-level domain names and therefore cases like *tmall.com*, the most frequently visited site for Chinese users, is treated as six different sites in Alexa (e.g., *login.tmall.com*, or *page.tmall.com*, to name a few). This way we better capture the organization that registers a domain name, making our data more analytically usable.

The second dataset has been obtained using a controlled experiment. In particular, we have used Selenium instrumented with our Lightbeam plugin to crawl popular sites from `Alexa.com`.[5] For the purpose of this paper, we mainly leverage the `Alexa top2000` global websites,[6] the `top500` categorized sites,[7] and `top500` national sites from each country.[8]

Fig. 1 also highlights some interesting results: CN and AU users display significantly *fewer* third-party domains than UK users. This also applies to CN-UK. Note that CN-UK users visit first-party sites from both UK and CN without the restriction of the Great Firewall in China. Thus, the number of third-party providers hitting CN-UK users is higher than for CN users alone — yet, not as high as UK users because many of the websites they visit are Chinese websites, with lower levels of tracking as discussed later. We also observe that the amount of third parties in US is lower than UK users. This implies that the popularity of third-party providers is in general higher across users located in UK, and that the demographics and browsing habits of the user (not just the location) play an essential role. Related works have provided a comparisons between region-specific third parties [20]. However, their analysis classifies regions by language. Thus, differences between US, Australia and the UK do not emerge.

The rest of this paper explores the implications of this finding. We examine where trackers are most effective: which categories of sites are they most prevalent in; how well they can track individual users as well as cohorts of users. This reveals differences across sites and categories

TABLE 3. NUMBER OF DOMAINS IN EACH THIRD PARTY CATEGORY OBSERVED IN OUR USERS, AS IDENTIFIED BY EIGHT DIFFERENT THIRD PARTY DATABASES. FOR EACH SOURCE AND CATEGORY WE LIST IN PARENTHESIS THE NUMBER OF THIRD PARTIES.

| | Num | Source Database |
|---|---|---|
| Advertising | 9110 | Disconnect(1588) [21], Webpage Toaster(1013) [22], EasyList(7580) [23], pgl(2816) [24], **OurList** (551) |
| Analysis | 552 | Disconnect(275), Webpage Toaster (308) |
| Essential | 965 | Disconnect(530), Webpage Toaster(515), **OurList** (47) |
| Malware | 68 | ZeuS Tracker [25], MalwareTips [26], **OurList** (68) |
| Optimization | 582 | Webpage Toaster(394), **OurList** (188) |
| Redirect | 31 | MalwareTips, **OurList** (31) |
| Social | 157 | Disconnect(59), Webpage Toaster(73), **OurList** (68) |
| Tracking | 3128 | Webpage Toaster (74), EasyList (2088), Better(604) [27], WhoTracks.me (1130) [28], **OurList** (732) |
| Total | 11,232 | Of this, 1685 were manually checked and added to "OurList", which *strengthens* the Chinese tracker list |

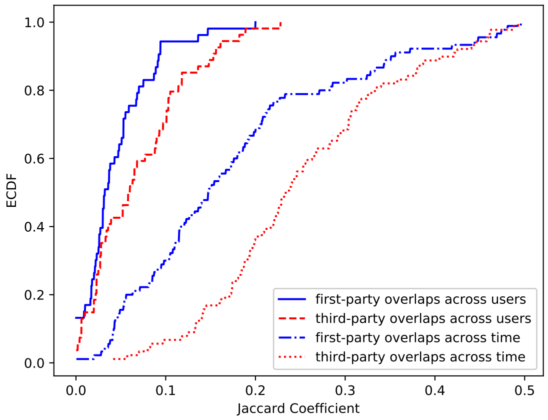


Figure 4. Overlaps across users and time in our real-user database. *Third parties obtain broader coverage of browsing habits than first parties.* This is true both for individual users over time and across users within the cohorts we study.

that are common in both countries (§3). Then we study differences between tracking across countries (§4).

## 3. Tracking patterns in UK and China

This section provides an overview of the magnitude of the tracking problem, and describes patterns common across both China and UK. We study the capability of third-party networks to track individual users over time, and a group of different individuals over a single time period (§3.1). Then we study where tracking is more prevalent, by exploring tracking on first party sites in different categories and with different popularity ranks (§3.2). Finally, we look at how the time spent on a web site affects the tracking strategies (§3.3).

### 3.1. Measuring tracking with overlaps

Trackers derive their power from obtaining a panoramic overview of browsing habits. We can extract a measure of this by studying overlaps in the first and third party domains across users and over time. Intuitively, this measure shows the similarity between two sets of browsing behaviors. There are two main applications to this. First, it can be used to measure how much overlap there is between the browsing behaviors of the same user

[4] There is one exception to this, we look at the third-level domain in the UK to take into account the type of entity (e.g., *ac.uk* or *co.uk* for academic and commercial domains respectively).

[5] https://www.alexa.com/topsites, which provides researchers with top websites across countries and categories

[6] Global Alexa top: aws.amazon.com/cn/alexa-top-sites/

[7] Alexa by category: www.alexa.com/topsites/category

[8] Alexa by country: www.alexa.com/topsites/countries

at two points in time [29]. Here, this measure can show to what extent a third-party network can track the user across the Web during that period. For instance, if there is a high overlap of third-party cookies across time, the issuer of the cookie will be capable of inferring most of the browsing history of the user during that period. Second, it can be used to measure how much overlap there is between the browsing behaviors of two users (or groups of users). This tells how similar two users are and, when looking at the third-party overlap, it can give a notion of how well a third-party network can track a population or cohort of users. It can also be used to compare the different user groups by country or location for instance.

In this paper, we use the Jaccard coefficient to measure the overlap between the two sites A and B. Formally, $J(A,B) = \sum overlaps / \sum websites = |A \cap B|/|A \cup B|$. We empirically observe that there is no correlation between Jaccard coefficient and the number of websites visited by different users in our dataset, implying that the amount of browsing of a user does not create a bias for this measure (at least in our dataset).

For each week, we compute the first- and third-party overlaps *across users*. The first-party overlap gives us a notion of how many users land in the same pages. The third-party overlap gives a picture of the extent to which third-party providers learn about similarities among users' browsing histories. We also measure the overlap of first- and third-party domains *across time* for individual users (i.e., the extent to which users revisit the same websites over different weeks, and the extent to which third parties know about the temporal visiting patterns of that user).

Fig. 4 depicts these overlaps. There are three take-aways: First, there is *higher* overlap in the third-party domain than in the first-party domain, both across different users and for individual users across time. This implies that third-party providers have a more comprehensive overview of browsing habits of individuals and cohorts of users than first party providers. Second, the overlaps of browsing histories over time for individual users is higher than the overlaps of browsing histories across different users, and this holds both for first-party and third-party domains. This implies that third parties are able to track individual users more effectively than tracking cohorts of users, and indicates a degree of specialisation (e.g., due to targeted advertising), whereby a third party may be interested in (or have visibility of) some users but not others. Finally, the difference between the first- and third-party overlap across cohorts of users is not as wide as the difference between first and third parties seen by individual users. This suggests that third parties do not have a massive advantage over first parties in understanding behaviors of cohorts of users. However, it should be noted that some of the largest domains, such as *Google* and *Facebook* act as both first and third parties depending on the context. For example, Google which can be a first party for search queries, is also a third party for analytic (Google Analytics) and for advertising (DoubleClick). As first party, Google has an extensive 64.2% coverage of the browsing histories of our UK user base. Thus some of the most common first parties may have much higher overview of users' browsing histories than third parties.

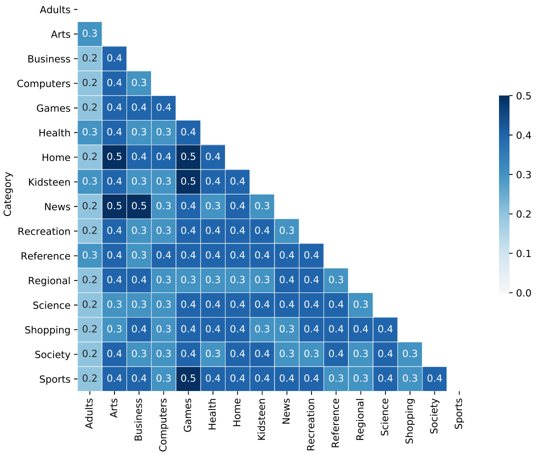


Figure 5. Jaccard Coefficient of third parties across `Alexa top500` websites by category after ADNS disambiguation.

TABLE 4. NUMBER OF THIRD PARTY DOMAINS SEEN IN `Alexa top500` PER CATEGORY BEFORE (#TP), AFTER ADNS (#ADNS) AND COOKIE SYNCHRONIZATION (#CSYNC) DISAMBIGUATION.

| Category (#TPs) | Category (#ADNS) | Rank | Category (#Csync) | Rank |
|---|---|---|---|---|
| News (3156) | News (396) | | Sports (207) | ↑ 1 |
| Sports (3051) | Sports (392) | | Recreation (201) | ↑ 4 |
| Business (3057) | Business (336) | | Shopping (198) | ↑ 8 |
| Arts (2814) | Arts (328) | | Business (177) | ↓ 1 |
| Home (2763) | Home (300) | | KidsTeen (171) | ↑ 2 |
| Recreation (2130) | Regional (280) | ↑ 3 | Home (165) | ↓ 1 |
| KidsTeen (2100) | Reference (268) | ↑ 7 | News (159) | ↓ 6 |
| Games (2043) | Society (268) | ↑ 2 | Arts (144) | ↓ 4 |
| Regional (1890) | Recreation (265) | ↓ 3 | Regional (135) | - |
| Society (1689) | Science (256) | ↑ 3 | Games (117) | ↓ 2 |
| Shopping (1641) | Shopping (248) | | Society (114) | ↓ 1 |
| Health (1578) | Games (246) | ↓ 4 | Computers (111) | ↑ 3 |
| Science (1491) | Computers (244) | ↑ 2 | Health (101) | ↓ 1 |
| Reference (1284) | KidsTeen (242) | ↓ 7 | Science (99) | ↓ 1 |
| Computers (1140) | Health (236) | ↓ 3 | Reference (54) | ↓ 1 |
| Adults (1134) | Adults (212) | | Adults (12) | - |

A different way to estimate the magnitude of tracking is to consider the extent to which trackers are shared among different kinds of websites. To study this, we make use of the categorisation of websites by `Alexa`. Fig. 5 shows the overlaps in third party domains between websites `Alexa top500` of different categories. The overlaps of third parties among most categories have a Jaccard coefficient in a tight band between 0.2 and 0.4, suggesting that in general, the category of a website does not make a huge difference to the presence or absence of particular trackers. However, there are notable exceptions: there are pairs of categories with an expected affinity (e.g., Kidsteen (kids & teens) and Games websites, News and Business, or Sports and Games) and these have a high Jaccard coefficient overlap of 50%. Another notable exception is the category of Adult sites, which have a very different ecosystem of third parties. These sites have a very low (≈ 0.2) overlap with most other categories of websites. This implies a degree of privacy for users visiting Adult websites, as called for by some regulators [30], and may be a consequence of explicit policies that some large trackers and mainstream advertisers have of not wanting to be associated with Adult sites.[9]

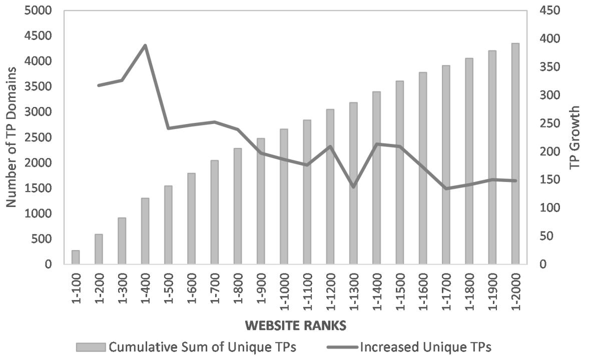


Figure 6. Growth rate of the number of third parties and the number of unique third parties in `Alexa top2000`, divided into bins of 100 sites by popularity ranks.

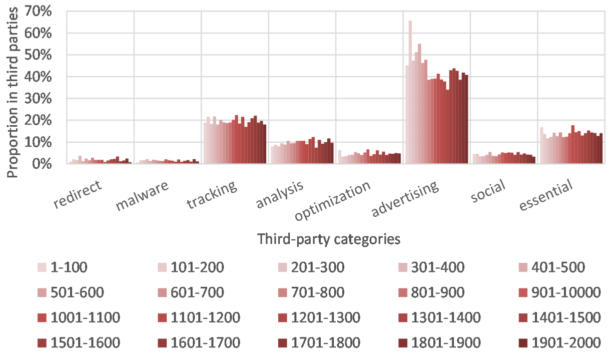


Figure 7. Proportion of third-party categories in `Alexa top2000`.

## 3.2. Impact of popularity rank & category

Inspired by the previous result of differences in overlaps between categories of websites, we next characterize to what extent the tracking varies among websites. First, in Table 4, we count numbers of third parties in the `Alexa top500` in different categories ranked by order. We establish that News websites have the highest numbers of third-party domains, and Adults the least numbers. In other words, the mainstream and accepted web browsing activity of reading news online has the highest amount of tracking and privacy violation. This confirms previous findings [8]. However, a limitation of previous works is that they do not take into account authoritative DNS (ADNS) and cookie synchronization, where two third parties might open a side channel to share data. Table 4 presents the number of third parties after disambiguation together with the relative change in the rank of the category. Assuming that merged entities share data, a decrease in their ranking means that users browsing pages in that category are more prone to be tracked than what was previously reported. We argue that when the diversity of third parties in a set of websites is reduced, single trackers then gain a better overview of a cohort. Our data shows more consolidation among trackers in KidsTeen, Health, Games or Recreation after coalescing by ADNS. Note that KidsTeen registers the largest rank decrease. Although websites in KidsTeen look like they have many smaller TP players, these are related entities and each player is bigger than what appears judging other works [8]. When looking at cookie synchronization, we observe that third parties in Sports, Recreation and Shopping categories share the largest number of cookies with other third parties. The content offered in sites under these categories enables targeted advertising and there is a greater incentive to share user's habits through cookie synchronization. The rest of the paper presents results after ADNS disambiguation, i.e., only third parties with different ADNS servers are recorded as distinct entities.

Next we ask whether websites of different levels of popularity also have different levels of tracking. Fig. 6 counts the number of unique third parties added as we go from the most popular websites (`Alexa` ranks 1–100) to less popular ones (up to `Alexa` rank 2000). As we move down the popularity rank, the cumulative number of unique third parties found continues to grow, indicating a vast and well developed ecosystem of third party providers. However, we find that the *number* of new third parties added for each 100 ranks plateaus out after an initial peak caused by a few of the popular websites. As a corollary, this means that the many academic papers which focus solely on `Alexa` ranked websites (e.g., [8], [31]–[33]) may be providing an *upper bound* on the amounts of tracking. For instance, while authors in [8] study 1M sites, they only sample top (100) sites in different categories. With some notable differences,[10] the number of third parties is over-approximated when considering sites in different popularity ranks. Equally, real users with browsing habits including specialist or niche-interest websites that are not among the most popular sites automatically tend to have lesser tracking (specially within the advertising industry). We confirm this by looking at our user group, where the number of third parties also plateaus out after an initial peak.

In Fig. 7, we seek to understand this result further by examining how different categories of trackers are used in websites of different popularity ranks. In most categories, there is not much difference in the relative proportion of trackers of that category among sites of different popularity ranks. However, advertising is one notable exception: the number of advertisers drops sharply after the top 1000 ranks, corresponding to the plateau of Fig. 7. Thus, the difference observed in number of trackers may be a result of financial pressures and incentives for online advertising, which pay more for more popular sites, and conversely, are less present in less popular sites.

## 3.3. Impact of loading time

Finally, we simulate an unusual experiment, to understand how tracking may change over time: We load `Alexa top100` websites in 100 Selenium instances and instruct Selenium to record all connections made after loading the site. We let every website run for seven days continuously and aggregate the observations on the time scale of each minute. This allows us to compute the per-minute rate of the increase or decrease in numbers of third parties in each third-party category. Table 5 shows the daily average increase in the numbers of trackers seen. Overall, we observe mostly positive values, indicating that

[9] E.g., see Taboola's policy https://bit.ly/2Vr9kQ9.

[10] E.g., Number of third parties remain steady across ranks for Government-related sites; numbers grow for Games and Shopping.

TABLE 5. ALEXA TOP 100 SITES ARE LOADED AND KEPT OPEN FOR 7 DAYS IN 100 SELENIUM INSTANCES. EACH ENTRY IN THE TABLE SHOWS THE DAILY AVERAGE OF THE PER-MINUTE RATE OF INCREASE IN THE NUMBERS OF THIRD PARTIES OF A PARTICULAR CATEGORY. (*OPT: OPTIMIZATION; AD: ADVERTISING)

| | redirect | malware | tracking | analysis | opt | ad | social | essential |
|---|---|---|---|---|---|---|---|---|
| day1 | 0.01% | 0.10% | 0.21% | 0.11% | 0.27% | 0.33% | 0.03% | 0.21% |
| day2 | 0.02% | 0.02% | 0.21% | 0.09% | 0.27% | 0.56% | 0.04% | 0.28% |
| day3 | 0.09% | 0.09% | 0.38% | 0.14% | 0.25% | 0.75% | 0.04% | 0.23% |
| day4 | 0.08% | 0.11% | 0.69% | 0.59% | 0.14% | 1.11% | 0.09% | 0.98% |
| day5 | 0.04% | 0.18% | 0.43% | 0.45% | 0.34% | 0.24% | 0.09% | 0.24% |
| day6 | 0.03% | 0.07% | 0.15% | 0.15% | 0.31% | 0.52% | 0.08% | 0.17% |
| day7 | 0.08% | 0.18% | 0.22% | 0.09% | 0.23% | 0.43% | 0.15% | 0.65% |

numbers of third parties keep increasing over time even after several days. The highest rate of increase is seen in the Advertising and Tracking categories of third parties, especially on days 3 and 4. Note that these two categories represent the largest fraction of third-party connections (c.f., Fig. 7). This seems to suggest a wide-spread practice of regular turnover of advertising and tracking third parties. To the best of our knowledge we are the first to report this behavior. We investigate how this behavior varies by country in Sec. 4.3.

In a small number of cases, we also observe that the tracker changes over time, over much longer time scales than the 7 day period of the above experiment, but visible in our year-long browser histories of our users. This change in trackers occurs due to a *renaming* of the tracking domain itself, i.e., a change in the domain name of the tracker. To exclude the influence of loading time in our analysis, we open the same websites and record the number of changes in third parties per minute over 30 minutes. We observe that there are changes over time, but these are not very prominent.

This occurs typically as a response to the third party domain being listed on a blocker's list such as uBlockOrigin, a free and open-source[11] browser add-on for the online ad-blocking. In response to this block, we observe that the old domain name is dropped, and a new domain name with a similar sounding name is registered. Fig. 8 lists a few examples. For example, shortly after `pussl4.com` is included in the uBlockOrigin list, that third party stops getting included in third parties loaded by first party sites, and instead is replaced by `pussl3.com`. Similarly, after `wikia-beacon.com` is blocked by uBlockOrigin, this is moved to `beacon.wikia-services.com`.

### 3.4. Key findings

By looking at third parties across different categories of websites, we note that there is a specialisation with higher overlaps between some categories than others. Furthermore, Adult websites have a lower overlap with most other categories, affording a degree of privacy. We also noted that numbers of trackers drop off as the popularity rank of a website decreases — real users, who may visit a significant number of niche interest websites outside the `Alexa` most popular lists tend to see fewer trackers. This have resulted on recent works [8], [31]–[33] reporting an *upper bound* on the amounts of tracking.

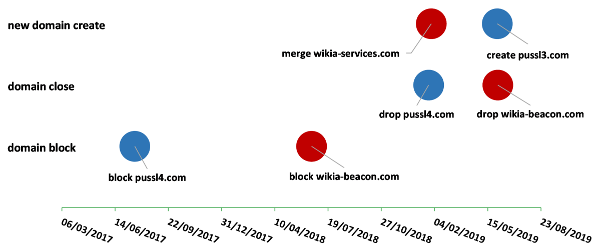


Figure 8. Business cycle of third-party domains, with a timeline typically displaying the following pattern: old domain blocked by listing in external database → old domain dropped (stops getting loaded by first parties) → usage of new domain starts (e.g., `pussl4.com` is replaced by `pussl3.com`) or merge with another domain (e.g., `wikia-beacon.com` is merged as a subdomain of `wikia-services.com`).

By looking at how third-party networks evolve over time, we note that there is an arms race that shapes the tracking ecosystem. On the one hand, we observe that first-party web sites have complex strategies that evolve over time, including loading different third-party technology over time scales of minutes, hours and days. Some of these strategies are likely motivated by the reactive nature of privacy-aware users that use blockers. On the other hand, we observe that the third-party trackers themselves change over time precisely when their domains are blacklisted by the blockers. This is done to increase survivability as dynamic loading strategies of first-parties will not favour loading third parties that are often blocked by their users.

In the rest of the paper we focus our attention on our user group to: i) drift away from reporting over approximated (`Alexa`) results, ii) to reduce the impact of popularity ranks and categories, iii) minimize the bias introduced by running experiments with different loading time in a context where real-time bidding might have an affect in our understanding of the tracking ecosystem.

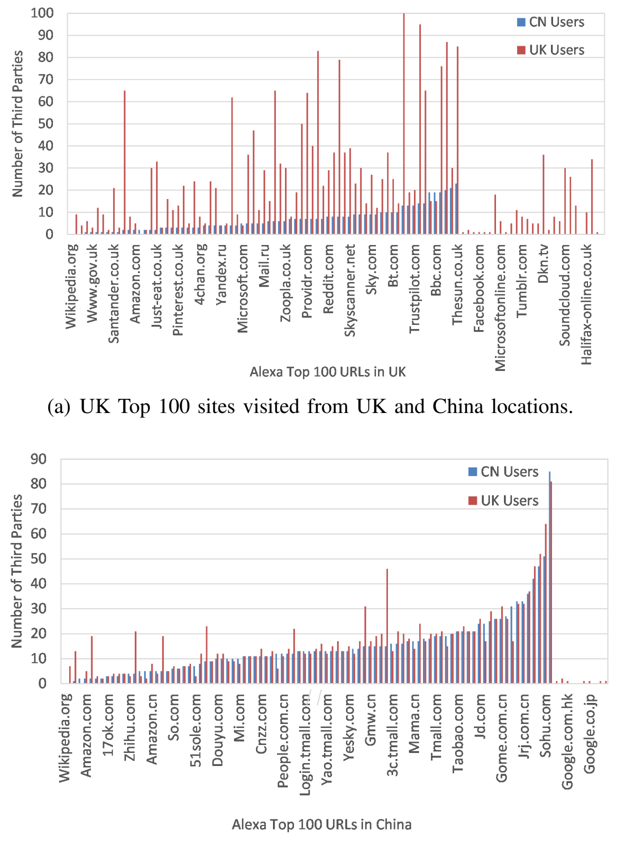
## 4. Country-level Differences

Due to the differences observed between UK and China, in this section we look at third-party technology that operates at a country level. First, we contextualize our study by comparing global third parties with local (CN and UK) third parties in §4.1. We then look at the cumulative growth on third parties targeting users throughout one year (§4.2). And lastly, we present the top actors across sectors and per country in §4.3.

### 4.1. Number of Third Parties in CN & UK

It is common to find websites that own the same second-level domain name in different countries (i.e., under several top-level domain like *.com* or *.uk*). Other sites, like Taobao, one of the highest traffic websites in China, operates two versions of their website homepage in different third-level domains (*'www.taobao.com'* for local users and *'world.taobao.com'* for global users). This is typically used to customize the homepage based on the origin of the users. During the course of our study, we have observed how some of these sites insert different

[11] Github: https://github.com/uBlockOrigin/uAssets



(a) UK Top 100 sites visited from UK and China locations.

(b) China Top 100 sites visited from UK and China locations.

Figure 9. Number of third parties targeting Chinese (CN) and UK users respectively on Alexa UK (top) and CN (bottom) most popular 100 sites. Top UK sites appear to target UK users more, but top Chinese sites target users from both CN and UK locations equally. (Each figure has 100 sites, but only a selection are labeled to ensure legibility).

TABLE 6. TOP 3 HOSTING LOCATIONS OF THIRD PARTY DOMAINS ENCOUNTERED BY REAL USERS FROM CN AND UK IN OUR USER STUDY. CHINESE USERS ARE MOSTLY SERVED BY LOCALISED THIRD PARTIES IN CHINA WHEREAS UK USERS ARE TRACKED BY US-BASED THIRD PARTY PROVIDERS.

| **Hosting Loc. (CN users)** | **x% at loc.** | **Hosting Loc. (UK users)** | **x% at loc.** |
|---|---|---|---|
| China | 66.3% | United States | 76.6% |
| United States | 24.5% | United Kingdom | 7.7% |
| South Korea | 1.8% | Ireland | 5.2% |

third-party technology on the sites they own. This suggests that some sites might tailor the number and type of third parties based on the location of the user. To verify this, we first study the interplay between the location of the user and the overall number of third parties observed.

We start by connecting from locations in China (CN) and UK to to Alexa top 100 websites (first parties) of each country. It is important to note that all sites are loaded in a controlled experiment, where we connect from the CN/UK locations to the same sites simultaneously. We also ensure that a clean browser profile with no previous history of cookies is used when visiting each website.

Fig. 9(a) shows the number of third parties observed when connecting to Alexa **UK** top 100 websites from China (CN) and UK. Observe that users located in the UK see significantly more third-party technology than users in CN. Interestingly, when repeating the experiment with Alexa **China** top 100 websites (Fig 9(b)), we found that the number of third parties does not vary as much with the location of the user. This suggests that *trackers in UK websites are more location sensitive than in China*.

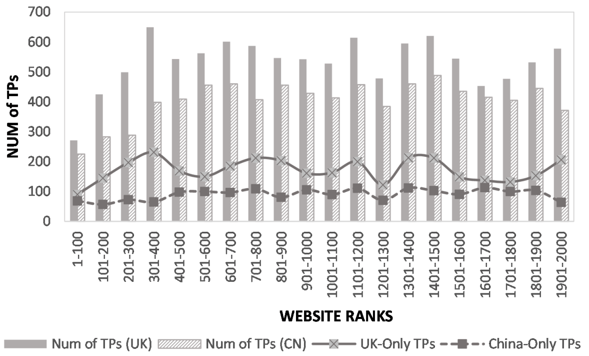


Figure 10. Numbers of third parties seen on Alexa top2000 (global ranking) websites, when accessed from UK and China (CN) locations; and the number of country-specific third parties found only when accessing from one location (UK or CN).

We explore this systematically in Fig. 10 by loading each website from the Alexa **global** top 2000 websites from locations in UK and CN, and counting the numbers of third parties (TPs) observed. We find that *across websites of different levels of popularity, UK-based users see more trackers than CN-based users*.

This is an unexpected result. We conjecture that this may partly be because users' locations play an important role in advertisers deciding whether to place an ad or not. Other third parties may also have similar reasons. However, an important reason may be that certain third parties are being thwarted by the Great Firewall (GFW) of China, which blocks services such as Facebook, Twitter and Google. None of these domains are seen from our China locations. In summary, *whether because of user demographic characteristics inferred based on locations, or because of GFW, users in China are subject to lesser tracking than users in the UK*.

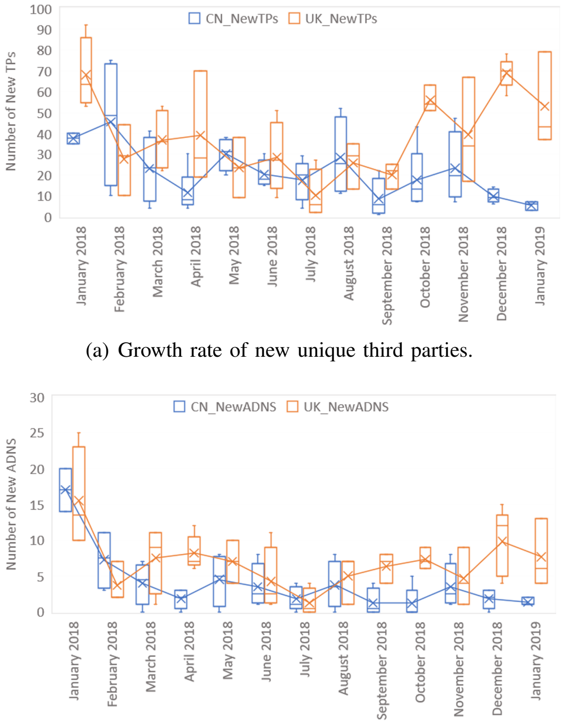
Fig. 10 also shows that there is a number of third parties which are only seen in the UK and not in CN. In total, approximately 46% of TPs seen in the UK are not seen in CN at all. This should be expected because of the above stated reasons of censorship at the GFW and demographic specialization. Interestingly, we also observe that there are several TPs which are only seen in the CN and not in the UK. 34% of TPs seen in China are endemic to users in that country. This indicates that *China also has a strong home-grown ecosystem of TPs*.

We explore this further in Table 6 by looking at all of the TPs encountered by our user groups in UK and China, and using a whois lookup to understand where those TPs are hosted. We find that most (66.3%) of third parties encountered by our Chinese users are located in China, although a significant minority (24.3%) are US-based. In stark contrast, nearly 77% of trackers for UK users are US-based, and only 7.7% are located in the UK. This provides further evidence of China's home-grown third party ecosystem. The globalized nature of third parties for UK-based users raises important questions about regulations and data management, especially in the wake of GDPR [7], [34].

## 4.2. Evolution over time

We next ask how third parties evolve over time through the lenses of our users. Fig. 11 shows the cumulative third-party growth over time for UK and CN users during one year. We observe a significant change on the number of new TPs at key times of the year, with spikes in February in China as well as December and April in the UK. These spikes may relate to Chinese New year, Easter, and Christmas respectively, where sites generally make promotional deals from new and different third-party advertisers. Besides, European Union's General Data Protection Regulation (GDPR) was also released during the collection period. Thus, the first data rebound of UK users in Fig. 11 might be also likely to be affected by new domains that process users' GDPR consents, with Trustarc and OneTrust offering their GDPR consent services starting from March/April 2018 [34].

We also do ADNS disambiguation for third-party collection from users, the growth in Fig. 11(b) is notably lower than in Fig. 11(a) — judging by the length of the box, we see that the difference between the users becomes smaller. In China, the most apparent ambiguity is in February, while the UK is in December. This means that although the growth of third parties is high during these two periods, most third parties come from the same tracking entity. However, the initial number of third parties in China exceeds those in UK after ADNS disambiguation. This means that the entities tracking Chinese users are *more dispersed*.



(a) Growth rate of new unique third parties.

(b) Growth rate of new third parties after ADNS.

Figure 11. Growth rate of the number of unique third-party domains/ADNS in our real-user study across one-year tracing.

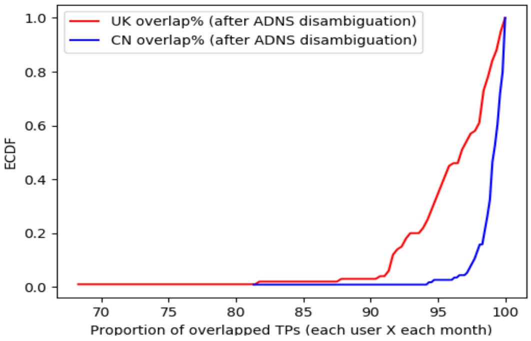


Figure 12. CDF of overlapped third-party domains and ADNS in our one-year online tracing against real users. China's overlap is higher than in the UK, indicating that the UK has a faster-developing third-party ecosystem. *New third parties are continuously growing faster in the UK* even the user visits the same number of websites, which is creepy.

TABLE 7. BROWSING HISTORY COVERAGE BASED ON THE PROPORTION OF FIRST PARTIES (FPS) OBSERVED BY TOP THIRD PARTIES (TPS) IN Alexa top2000 GLOBAL, AND IN OUR REAL-USER STUDY FROM UK, CHINA, AUSTRALIA AND US USERS.

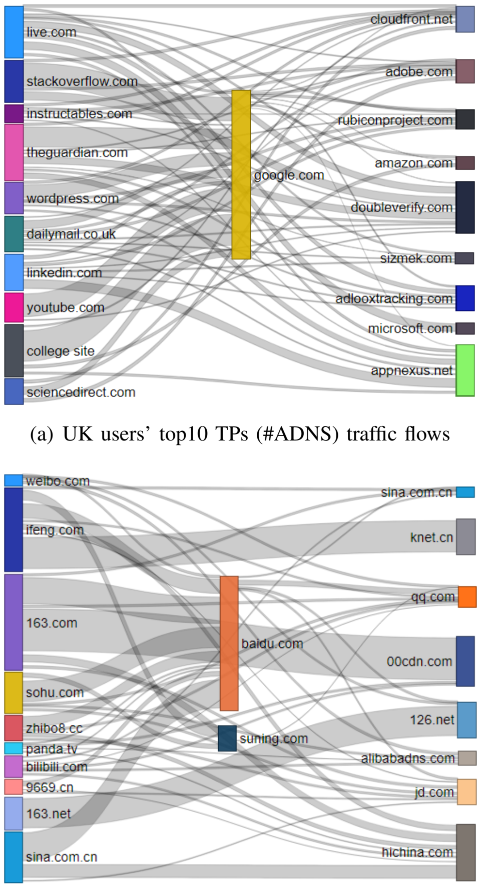
| | Proportion of FPs | | | | |
|---|---|---|---|---|---|
| **Top TPs** | **Alexa** | **UK** | **CN** | **AU** | **US** |
| **Google** | 79.6% | 64.2% | 11.4% | 35.71% | 83.85% |
| **Facebook** | 36.7% | 14.9% | - | 7.14% | 29.53% |
| **Scorecard** | 20.1% | 9.8% | - | - | 14.99% |
| **Twitter** | 13.9% | 10.5% | - | 2.38% | 19.49% |
| **Cloudfront** | 10.8% | 7.2% | 1.8% | 2.38% | 12.11% |
| **Quantserve** | 7.7% | 6.0% | - | - | 7.34% |
| **Bing** | 6.8% | 12.5% | - | - | 12.68% |
| **Baidu** | 5.6% | 20.9% | 46.4% | 7.14% | 11.06% |
| **Alibaba** | - | - | 18.2% | 11.90% | 6.53% |
| **Sina** | - | - | 13.7% | - | - |
| **QQ (Tenant)** | - | - | 4.6% | - | - |

## 4.3. Localisation and concentration of TPs

Motivated by the findings in the previous section, we look at how the network of third parties is structured globally, as well as in the UK and in China, and how much coverage a single third party can obtain of an individual user's browsing history, or of visits to a well specified set of websites. We first look at a well specified set of web sites – Alexa top2000 from the global ranking. Then, we use our dataset of real UK and Chinese participants. In both cases, we compute the browsing-history coverage that a single TP provider can obtain. Table 7 shows a summary ranked by TP provider for both Alexa top2000 and real user study.

### 4.3.1. Controlled experiment with Alexa top2000.

When looking at the Alexa column in Table 7, we observe that Google is the provider with largest coverage, and is present in nearly 80% of the Alexa top2000. Note again that we have grouped together all known networks owned by Google, like DoubleClick or Google Analytics. The second provider in terms of coverage is Facebook, which is shown to be capable of tracking users out of their site. Facebook has a visibility of 36.7% of Alexa top2000. Finally, we can see how other providers like Scorecard, Twitter, or CloudFront have a less dominant share although their presence is still significant. *This indicates a strong concentration of browsing history visibility in the hands of a few top third parties.*



(a) UK users' top10 TPs (#ADNS) traffic flows

(b) China users' top10 TPs (#ADNS) traffic flows

Figure 13. Sankey diagram of top 10 third-party websites (after ADNS disambiguation) over traffic of our real-user study (UK on the top, CN on the bottom) topsites. Left bar shows first parties (FPs) and right bars show the third parties (TPs) loaded by the FPs. Each flow represents a loading action, and the width of each flow is proportional to the number of times a TP is loaded by a given FP. (Left bars are top 10 first parties corresponding to the number of loads.)

**4.3.2. User study of multi-country participants.** We next look at the UK and CN columns in Table 7, which represent the third parties seen in our user group. First, we look at third-party networks monitoring UK users. We observe that Google is able to cover 65% of the browsing history of our users as shown in Table 7. This is slightly *lower* than what we observed in the controlled experiment with Alexa top2000 and corroborates our previous finding (in Sec. 3.2) that showed that Alexa-based studies may be over estimating the amount of tracking. Note that Baidu is able to observe about 20% of the browsing history of UK users. This corresponds to Chinese users based in UK (referred as CN-UK in §2.5).

Next, we look at users in China where Baidu is positioned as the top third-party provider, with a coverage of 46.4%. Interestingly, although Google.cn and Doubleclick ceased operations in China several years ago [35], we can see how Google still has access to the browsing history of 11% of the users in China, mainly through other domains owned by Google, which are not blocked. The remaining TP providers are fragmented. This fragmentation might be explained by the low cost of .cn first-year registration domains, which was set to only 1 RMB in 2007 to encourage the development of Chinese websites. This leads to a larger number of TP domains in CN, but with each of them having a smaller overview of the overall market. In the final column (US users), the proportion of TPs in each category shows a high concentration towards Google and Facebook, similar to the percentages found from Alexa top websites. This may be a result of Alexa rankings being influenced by a large extent by US websites.

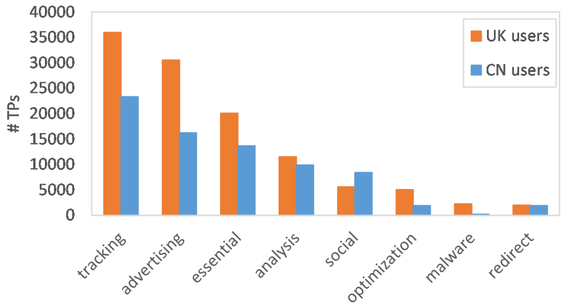


Figure 14. Number of third-party categories in actual user research: only *social* third-party vendors display a higher number in China.

Finally, we look at the provenance of the connections to the most prevalent third parties. Fig. 13 shows connections from the top 10 first-party websites (on the left side of the Sankey diagram) visited by our UK and CN users to the different third parties after ADNS disambiguation. Note that some Web domains, such as *google.com* and *bauidu.com*, can act as first and third party and they are thus presented in the middle of the Sankey diagram. Here, Google's third-party impact on UK users is *10% higher* than Baidu's third-party impact on our CN users. As a result, Google has more comprehensive user information with higher frequency and reach. Comparing the topology of the UK and CN Sankey diagrams, we can see that flows in UK are more complex and intertwined. The average first party in UK loads a wider-range of third parties as opposed to China, that are site- or entity-specific. This displays a lower cross-site data leakage in CN over UK and further evidences that Chinese third parties are site-specific decentralized structure (as discussed in §4.2).

**4.3.3. One-year third-party categorization.** We next present an overview of the third-party categories we have seen continuously tracking users for a year (i.e., between 2018 Jan-2019 Jan) in Fig. 14. We see that in almost all third-party categories, UK providers hold a relatively leading edge. However, in terms of *social* third parties, Chinese providers have a larger network of third parties than UK. Surprisingly, authors of [36] show that "the most widely used social media and content sharing application in the West are banned by the Chinese government", and it can be concluded that China's social media self-marketing has taken up a large space. Note that social media penetration in China mainland is at 71% (HongKong is 78%) and UK is only 67% [37] of the population. Therefore, frequent activities on Chinese social media attract the interests of relevant third-party providers.

## 4.4. Specialisation by category

We also see that, in general, third parties are specialized by sectors. Fig. 15 shows the type of websites in

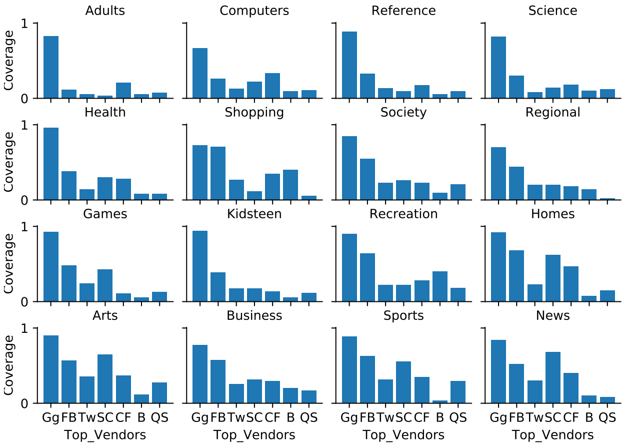


Figure 15. Coverage of third party providers among websites in the `Alexa top500` list for each of 16 categories. We only list a selection of the top TPs shown in Table 7. (Gg: Google, FB: Facebook, Tw:Twitter, CF: CloudFront, QS: Quantserve, B:Bing, SC: Scorecardresearch.)

which the main third party actors have larger coverage. While Google is generally present in all 16 first-party categories, we observe that the presence of other providers changes significantly from one category to another. For instance, Bing (denoted as 'B' in Fig. 15) is well positioned in the *shopping* category, but it holds poor coverage on websites that are part of the *news*, *games*, or *sports* category. Furthermore, we find that the *adults*, the *reference*, and the *science* category is primarily dominated by Google (labeled 'Gg') alone. In contrast, *arts*, *sports*, and *news* are competitive categories where providers like Facebook ('FB'), Scorecardresearch ('SC') and Twitter ('Tw') stand out in terms of coverage.

## 4.5. Key findings

Although the third-party ecosystem is concentrated in a handful of actors (e.g., Google Facebook and Baidu), we show that there is a degree of specialisation based primarily on: i) the type of sector of the first-party website, and ii) the location of the user. On the one hand, third-party providers that are specialized by sectors and small actors (in terms of overall coverage) can have access to most of the browsing-history of users interested in a given sector (e.g., Scorecardresearch with websites in the *Arts* category). On the other hand, we observe that third-party providers are also specialized by country. However, the vast majority of TP domains in the UK and a smaller portion of TP domains in China are hosted in the US. How user data is processed and where it is located has important regulatory implications. For this reason, large corporations like Google have in place frameworks[12] to protect cross-country data.

# 5. Related Work

Our work relates to two general areas: *(i)* online tracking behaviors, with emphasis in targeted advertising, and *(ii)* categorization of the different third parties. We also review other works that have looked at the location of the users when studying this ecosystem and that have used real-users, where there has not been as much work.

**Online tracking**. Related works such as [38]–[40] have been looking at better ways to detect online trackers, including anonymizing the *referer* field in HTTP requests [3]. Although our work does not directly aim at blocking trackers, or attacking them in other ways (e.g., [41]), identification of third parties is a paramount first step and a key concern for us. For this, we have referred to and used strategies, heuristics or third party lists from a number of efforts like ChromeDanger [42], Ghostery [8], Brave [43], AdReveal [44], Adblock [45], Plus [46], XRay [47], TrackAdvisor [3], and Disconnect [48]. Other works focus on advertising [3], [44] or on service media [9] alone as well as they do not consider country-specific trackers.

However, in our work we go one step further by measuring the current state of the tracking ecosystem in the Web. An overview of the evolution of the third-party tracking ecosystem is given in [40]. Authors show in 2012 that web measurements are an effective way to understand trackers. Later in 2015, authors in [31] provide an overview of the usage of cookies over 1 million sites. In another work from 2016, authors look at an advanced form of tracking that uses a cookie hijacking attack [49]. The setting proposed is adversarial and can therefore be considered less realistic than our study. More importantly, authors focus largely on DoubleClick, Google, and Amazon; TPs that hardly operate in China.

Our work looks at trackers today, but follows the design principles proposed in [40]. The size of our measurement is not as large as in [31]. However, we provide special attention to the most popular sites in different countries. We further consider the browsing habits of a user group that volunteer to our study.

**Categorization**. Similar to us, other authors such as [38] and [32] also categorize trackers and look at the prevalence of third parties across these categories. They both use the McAfee database [50]. One major problem of this dataset is that it only provides the category of domain and does not care about the popularity of the site. However, in our work we show that looking at datasets in bulk without considering the site's popularity can lead to over approximations in the amount of tracking. Other works such as [8], [51], [52] consider different categories of websites, but these are broad and they do not consider key fine-grained categories. Furthermore, we perform extensive manual validation and increase the size of publicly available lists for Western websites by 12.8% and Chinese websites by 23.4%.

**Country-specific analysis**. Related works have distinguished the location of the user when looking at trackers. First, [53] and [54] have looked at this in 2004 and in 2014 respectively. As opposed to our findings, they both conclude that the location of the user has limited influence in the amount and type of tracking. On the other hand, authors in [55] analyze which trackers are used in different countries. They conclude that the Chinese market is not dominated by the same trackers that are popular in other countries. However, their analysis does not provide deep insights and, more importantly, the type of tracking is not contextualized as we do in our work.

**Real-user contribution**. Most studies are based either

[12] EU-US Privacy Shield Framework: https://bit.ly/2XCgvYn

on visiting specific websites such as the Alexa most popular websites (e.g., [8], [56]–[59]), or they may at best artificially construct "personae" by initially visiting a number of websites that represent a particular persona or demographic. (e.g., [60]–[62]). In contrast, our results are based on 1 year of real user browsing behavior. Where our results coincide with previous studies, it offers a confirmation that real users in the wild are affected as researchers previously believed. We have highlighted differences where we observe them. For instance, we discover intensive *social-categorized third parties* tracking China users. Authors in [7] discuss the cross-region tracking flows based on the data collection from 350 real users. But their findings are only aimed at two categories of third parties (i.e, advertising and tracking) classified by AdBlockPlus list. Instead, our paper optimizes the categorization of third parties that improves the classification accuracy of a total of eight categories. With the broader classification and the one-year tracing of our UK and China participants, we figure out that a considerable number (over 30K) of trackers are following our UK users.

## 6. Discussion & Conclusion

In this paper, we presented a measurement study that sheds light into the magnitude of the tracking ecosystem in different countries. We built technology to capture the extent to which third-party trackers are profiling users. With a real-user study, we have highlighted the limitations of measurements that rely solely on Alexa. We also presented a categorization of third-party domains that improves the state of the art in terms of performance. All this, together with a set of experiments designed to understand the interplay between the location of the user and the strategies of trackers, showed that this ecosystem is quite complex.

**Takeaways.** Our analysis highlights that first-party web sites exhibit dynamic strategies that change over time, with respect to the location of the user and the *kind* of website the user connects to. In particular, we have observed an important wealth of country-specific trackers as well as trackers that are good at targeting segments of users' browsing histories (e.g., *Shopping*). We have also observed, for the first time, dynamic strategies whereby new third parties continue to be loaded by websites even several days after the initial loading of a website. All takeaways above stem from the Alexa dataset. Unexpectedly, we found that UK users see more trackers than China-based users judging by our real-user study. One of the reasons for this is the blocking in China of domains such as Google, Facebook and Twitter, which are not only major first parties but also some of the most important players in the UK third party ecosystem. Finally, we also observe a relatively larger number of social third parties in China. Finally, being able to study the real browsing habits of a panel of users, we are able to show that studies which rely solely on curated lists of popular websites such as from Alexa may be *over estimating* the level of tracking of real users.

**Limitations.** Our panel of participants is small and does not represent the diversity of the world's population. In particular, all preliminary users sampled have computer-related jobs and are young adults. This means our cohort may be biased towards a specific demographic segmentation. We attempt to understand the representativeness of our study by comparing the third parties from our user study with the distribution of third parties in popular sites per country using Alexa. Although the distributions look similar (Fig. 1), we acknowledge that Alexa sites may also not be reflective of many users' browsing patterns. Extending the panel of participants is part of an on-going effort and reporting these results is precisely the scope of our future work. At the time of writing, we see 2384 installations of our browser plugin, of whom 566 users have consented to share their data with us, representing a coverage of 65 countries. Preliminary results from this cohort shows that similar results hold on the much larger population. As we collect real-user data from different countries, we are also becoming aware that cultural factors [63] and cultural influences on perceptions of privacy [64] would need to be factored in to achieve a complete understanding of how the third party ecosystem evolves in different areas of the world.

On another note, cookies are not the only way to track users. Trackers might use other technologies such as fingerprinting the browser of the user for canvas recognition. However, it possibly makes our study prone to false positives [65], [66]. Therefore, we have not considered more elaborate ways of profiling users when we computed the third-party overlaps. For this reason, our findings have to be seen as an under-approximation of the magnitude of the problem. Furthermore, although we have successfully categorized nearly 90% of third-party domains for UK users and 70% for Chinese users, still we have not been able to characterize all third parties. One of the strengths of this study has been the availability of browsing histories of real users. This has led to several new results (e.g., Tables 6 and 7), but these results are also limited by the fact that they have not been verified more generally.

**Future Work.** As the location matters, we would first like to extend our analysis to a wider range of countries. Also, we plan to study the effect that other demographic features (e.g., the language, age or gender of users) might have in the strategies used to track users. However, the main scope of our future work is to study additional solutions to protect the privacy of the users. In particular, we want to further explore the trade-off between usability and privacy (e.g., having an OAuth account in the right container), and the use of our categorization to cluster websites into containers in a private, but meaningful manner. An important source of variation is that Real Time Bidding (RTB) can lead to different TPs over different visits. We plan to explore this systematically in a multi-country context using our expanded Chrome plugin user base.

## Acknowledgements

We acknowledge support via EPSRC Grant Ref: EP/T001569/1 for "Artificial Intelligence for Science, Engineering, Health and Government", and particularly the "Tools, Practices and Systems" theme for "Detecting and Understanding Harmful Content Online: A Metatool Approach". Thanks also to the anonymous shepherd for the suggestions which greatly helped improve the paper.

# References

[1] G. Roels and K. Fridgeirsdottir, "Dynamic revenue management for online display advertising," *Journal of Revenue and Pricing Management*, vol. 8, no. 5, pp. 452–466, 2009.

[2] F. Roesner, T. Kohno, and D. Wetherall, "Detecting and defending against third-party tracking on the web," in *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. USENIX Association, 2012, pp. 12–12.

[3] T.-C. Li, H. Hang, M. Faloutsos, and P. Efstathopoulos, "Trackadvisor: Taking back browsing privacy from third-party trackers," in *International Conference on Passive and Active Network Measurement*. Springer, 2015, pp. 277–289.

[4] H. Metwalley, S. Traverso, and M. Mellia, "Using passive measurements to demystify online trackers," *Computer*, vol. 49, no. 3, pp. 50–55, 2016.

[5] M. Hanson, P. Lawler, and S. Macbeth, "The tracker tax: the impact of third-party trackers on website speed in the united states," 2018.

[6] A. Razaghpanah, R. Nithyanand, N. Vallina-Rodriguez, S. Sundaresan, M. Allman, C. Kreibich, and P. Gill, "Apps, trackers, privacy, and regulators: A global study of the mobile tracking ecosystem," in *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2018.

[7] C. Iordanou, G. Smaragdakis, I. Poese, and N. Laoutaris, "Tracing Cross Border Web Tracking," in *Proceedings of ACM IMC 2018*, Boston, MA, October 2018.

[8] S. Englehardt and A. Narayanan, "Online tracking: A 1-million-site measurement and analysis," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. ACM, 2016, pp. 1388–1401.

[9] J. K. Sørensen and H. Van den Bulck, "Public service media online, advertising and the third-party user data business: A trade versus trust dilemma?" *Convergence*, p. 1354856518790203, 2018.

[10] Mozilla, "Firefox lightbeam," Available at https://github.com/mozilla/lightbeam-we, 2012.

[11] Mozilla, "Openwpm," Available at https://github.com/mozilla/OpenWPM.

[12] Toolness, "Collusion," Available at http://www.toolness.com/wp/2011/07/collusion/, 2012.

[13] Disconnect, "Take back your privacy." Available at https://disconnect.me/, 2013.

[14] A. Gervais, A. Filios, V. Lenders, and S. Čapkun, "Quantifying web adblocker privacy," *Cryptology ePrint Archive*, vol. 2016, p. 900, 2016.

[15] M. E. Acer, E. Stark, A. P. Felt, S. Fahl, R. Bhargava, B. Dev, M. Braithwaite, R. Sleevi, and P. Tabriz, "Where the wild warnings are: Root causes of chrome https certificate errors," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2017, pp. 1407–1420.

[16] I. Sanchez-Rola, M. Dell'Amico, P. Kotzias, D. Balzarotti, L. Bilge, P.-A. Vervier, and I. Santos, "Can i opt out yet?: Gdpr and the global illusion of cookie control," in *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*. ACM, 2019, pp. 340–351.

[17] B. Krishnamurthy and C. Wills, "Privacy diffusion on the web: a longitudinal perspective," in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 541–550.

[18] P. Papadopoulos, N. Kourtellis, and E. Markatos, "Cookie synchronization: Everything you always wanted to know but were afraid to ask," in *The World Wide Web Conference*. ACM, 2019, pp. 1432–1442.

[19] T. L. Beauchamp, "The belmont report," *The Oxford textbook of clinical research ethics*, pp. 149–155, 2008.

[20] M. Falahrastegar, H. Haddadi, S. Uhlig, and R. Mortier, "The rise of panopticons: Examining region-specific third-party web tracking," in *International Workshop on Traffic Monitoring and Analysis*. Springer, 2014, pp. 104–114.

[21] Disconnect, "Disconnect third-party database," Available at https://github.com/disconnectme/disconnect-tracking-protection/blob/master/services.json.

[22] T. Daish, "The webpage toaster third-party database," Available at https://www.webpagetoaster.com/list3pdb.php.

[23] EasyList, "Easylist overview," Available at https://easylist.to/pages/other-supplementary-filter-lists-and-easylist-variants.html.

[24] P. Lowe, "pgl.yoyo.org blocklist," Available at http://pgl.yoyo.org/as/serverlist.php?hostformat=nohtml&showintro=1&mimetype=plaintext.

[25] Zeus, "Zeus tracker," Available at https://zeustracker.abuse.ch/blocklist.php.

[26] MalwareTips, "Malware removal guides," Available at https://malwaretips.com/forums/malware-removal-guides.11.

[27] Better, "Trackers collections," Available at https://better.fyi/trackers/.

[28] WhoTracks.me, "Trackers rank," Available at https://whotracks.me/trackers.html.

[29] J. Drew and T. Moore, "Automatic identification of replicated criminal websites using combined clustering," in *2014 IEEE Security and Privacy Workshops*. IEEE, 2014, pp. 116–123.

[30] I. Altaweel, M. Hils, and C. J. Hoofnagle, "Privacy on adult websites," in *Altaweel et al., Privacy on Adult Websites, Workshop on Technology and Consumer Protection (ConPro'17), co-located with the 38th IEEE Symposium on Security and Privacy, San Jose, CA (2017)*, 2016.

[31] T. Libert, "Exposing the hidden web: An analysis of third-party http requests on 1 million websites," *arXiv preprint arXiv:1511.00619*, 2015.

[32] A. Chaabane, M. A. Kaafar, and R. Boreli, "Big friend is watching you: Analyzing online social networks tracking capabilities," in *Proceedings of the 2012 ACM workshop on Workshop on online social networks*. ACM, 2012, pp. 7–12.

[33] T. Libert and R. K. Nielsen, "Third-party web content on eu news sites: Potential challenges and paths to privacy improvement," 2018.

[34] X. Hu and N. Sastry, "Characterising third party cookie usage in the eu after gdpr," in *Proceedings of the 10th ACM Conference on Web Science*, ser. WebSci '19. New York, NY, USA: ACM, 2019, pp. 137–141.

[35] T. Libert and M. Repnikova, "Google is returning to china? it never really left," Available at https://www.theguardian.com/technology/2015/sep/21/google-is-returning-to-china-it-never-really-left, 2015.

[36] "China social media statistics facts in 2018 to shape your marketing strategy," Available at https://www.marketingexpertus.co.uk/blog/china-social-media-statistics-2018, 2018.

[37] "Chaffey, dave," Available at https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/, 2019.

[38] B. Krishnamurthy, K. Naryshkin, and C. Wills, "Privacy leakage vs. protection measures: the growing disconnect," in *Proceedings of the Web*, vol. 2, no. 2011, 2011, pp. 1–10.

[39] A. Soltani, S. Canty, Q. Mayo, L. Thomas, and C. J. Hoofnagle, "Flash cookies and privacy," in *2010 AAAI Spring Symposium Series*, 2010.

[40] J. R. Mayer and J. C. Mitchell, "Third-party web tracking: Policy and technology," in *2012 IEEE symposium on security and privacy*. IEEE, 2012, pp. 413–427.

[41] I. L. Kim, W. Wang, Y. Kwon, Y. Zheng, Y. Aafer, W. Meng, and X. Zhang, "Adbudgetkiller: Online advertising budget draining attack," in *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 2018, pp. 297–307.

[42] L. Bauer, S. Cai, L. Jia, T. Passaro, and Y. Tian, "Analyzing the dangers posed by chrome extensions," in *2014 IEEE Conference on Communications and Network Security*. IEEE, 2014, pp. 184–192.

[43] G. Franken, T. Van Goethem, and W. I. Joosen, "Exposing cookie policy flaws through an extensive evaluation of browsers and their extensions," *IEEE Security & Privacy*, 2019.

[44] B. Liu, A. Sheth, U. Weinsberg, J. Chandrashekar, and R. Govindan, "Adreveal: improving transparency into online targeted advertising," in *Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks*. ACM, 2013, p. 12.

[45] P. Papadopoulos, "Analyzing the impact of digitaladvertising on user privacy," Available at http://users.ics.forth.gr/~panpap/thesis/panpap_phd_dissertation.pdf.

[46] A. Gervais, A. Filios, V. Lenders, and S. Capkun, "Quantifying web adblocker privacy," in *European Symposium on Research in Computer Security*. Springer, 2017, pp. 21–42.

[47] M. Lécuyer, G. Ducoffe, F. Lan, A. Papancea, T. Petsios, R. Spahn, A. Chaintreau, and R. Geambasu, "Xray: Enhancing the web's transparency with differential correlation," in *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, 2014, pp. 49–64.

[48] W. Boumans and I. E. Poll, "Web tracking and current countermeasures," 2017.

[49] S. Sivakorn, I. Polakis, and A. D. Keromytis, "The cracked cookie jar: Http cookie hijacking and the exposure of private information," in *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2016, pp. 724–742.

[50] TrustedSource, "Customer url ticketing system," Available at www.trustedsource.org/en/feedback/url.

[51] P. Barford, I. Canadi, D. Krushevskaja, Q. Ma, and S. Muthukrishnan, "Adscape: Harvesting and analyzing online display ads," in *Proceedings of the 23rd international conference on World wide web*. ACM, 2014, pp. 597–608.

[52] P. G. Leon, B. Ur, Y. Wang, M. Sleeper, R. Balebako, R. Shay, L. Bauer, M. Christodorescu, and L. F. Cranor, "What matters to users?: factors that affect users' willingness to share information with online advertisers," in *Proceedings of the ninth symposium on usable privacy and security*. ACM, 2013, p. 7.

[53] S. Bellman, E. J. Johnson, S. J. Kobrin, and G. L. Lohse, "International differences in information privacy concerns: A global survey of consumers," *The Information Society*, vol. 20, no. 5, pp. 313–324, 2004.

[54] M. Falahrastegar, H. Haddadi, S. Uhlig, and R. Mortier, "Anatomy of the third-party web tracking ecosystem," *arXiv preprint arXiv:1409.1066*, 2014.

[55] S. Schelter and J. Kunegis, "Tracking the trackers: A large-scale analysis of embedded web trackers," in *Tenth International AAAI Conference on Web and Social Media*, 2016.

[56] P. Peng, C. Xu, L. Quinn, H. Hu, B. Viswanath, and G. Wang, "What happens after you leak your password: Understanding credential sharing on phishing sites," 2019.

[57] J. Brookman, P. Rouge, A. Alva, and C. Yeung, "Cross-device tracking: Measurement and disclosures," *Proceedings on Privacy Enhancing Technologies*, vol. 2017, no. 2, pp. 133–148, 2017.

[58] P. Papadopoulos, P. Snyder, and B. Livshits, "Another brick in the paywall: The popularity and privacy implications of paywalls," *arXiv preprint arXiv:1903.01406*, 2019.

[59] M. H. Mughees, Z. Qian, and Z. Shafiq, "Detecting anti ad-blockers in the wild," *Proceedings on Privacy Enhancing Technologies*, vol. 2017, no. 3, pp. 130–146, 2017.

[60] M. A. Bashir and C. Wilson, "Diffusion of user tracking data in the online advertising ecosystem," *Proceedings on Privacy Enhancing Technologies*, vol. 2018, no. 4, pp. 85–103, 2018.

[61] K. Solomos, P. Ilia, S. Ioannidis, and N. Kourtellis, "Clash of the trackers: Measuring the evolution of the online tracking ecosystem," *arXiv preprint arXiv:1907.12860*, 2019.

[62] J. Cook, R. Nithyanand, and Z. Shafiq, "Inferring tracker-advertiser relationships in the online advertising ecosystem using header bidding," *arXiv preprint arXiv:1907.07275*, 2019.

[63] E. M. Redmiles, """ should i worry?" a cross-cultural examination of account security incident response," in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019, pp. 920–934.

[64] L. Zheng, "Does online perceived risk depend on culture? individualistic versus collectivistic culture," *Journal of Decision Systems*, vol. 26, no. 3, pp. 256–274, 2017. [Online]. Available: https://doi.org/10.1080/12460125.2017.1351861

[65] N. M. Al-Fannah and C. Mitchell, "Too little too late: can we control browser fingerprinting?" *Journal of Intellectual Capital*, 2020.

[66] S. Luangmaneerote, E. Zaluska, and L. Carr, "Survey of existing fingerprint countermeasures," in *2016 International Conference on Information Society (i-Society)*. IEEE, 2016, pp. 137–141.

# Appendix A. Cookies Synchronization

We reset the length range of extracting userIDs and reposition the length reduced process to speed the detection. We adopt the twofold restriction on the length of the userID (red dashed line in Appendix Fig. 17): while decoding cookies and splitting userIDs, which increases the time complexity and saves about 20% of the processing time. Fig. 17 in the Appendix shows the workflows/model of how we detect the relationship among different third parties.

TABLE 8. Example of the tracker's cookie synchronization (tracked userIDs in *smartadserver.com*), and the Cookie Synchronization process is in Fig. 16

<table>
<tr><th>Name: csync</th></tr>
<tr><td>Content:<br>76672661796328007079<br>| 25:**33ae5d59-b0db-4e00-86c7-c8556a3de0d5**<br>| 86:**2333759400305871849**<br>| 100:cfda683f-2263-4a8d-9888-06e322420770<br>| 49:**6726617963280070796**<br>| 124:**0c2a3972-f2da-41ad-b62c-b16841e4988d**<br>| 127:**AAW1Bk66s3wAACEmJLl5fg**<br>| 107:**aa64d0ec-9964-4d8c-b670-ea7e9ca10bed-tuct45343bd**<br>| 101:**Gz9uGy6TFeJbytdtnGOIz-y59NhuPkC1jweVBZPfYKw=**<br>| 111:3737669618283368<br>| 31:**9e29c053-db56-499a-a86a-a7ad25b3b09a**<br>| 91:**F1942AE1-2F65-47E2-B8D9-7EE8F3A62F9E**<br>| 116:**IcRHiNKwF5EE3A6-flJ3**<br>| 79:**ca2cfef3-074d-42be-a5d4-a4573b3a9e29**</td></tr>
<tr><td>Domain: .smartadserver.com</td></tr>
</table>

Based on the example in Table 8 and Fig. 16 in Appendix, we could understand how the domain of *smartadserver.com* aggregates userIDs from other connected third-party domains, which is the process of sharing user information between vendors that through the same userID. So *smartadserver.com* linked to at least 25 $2^{nd}$-level domains through the cookie synchronization.

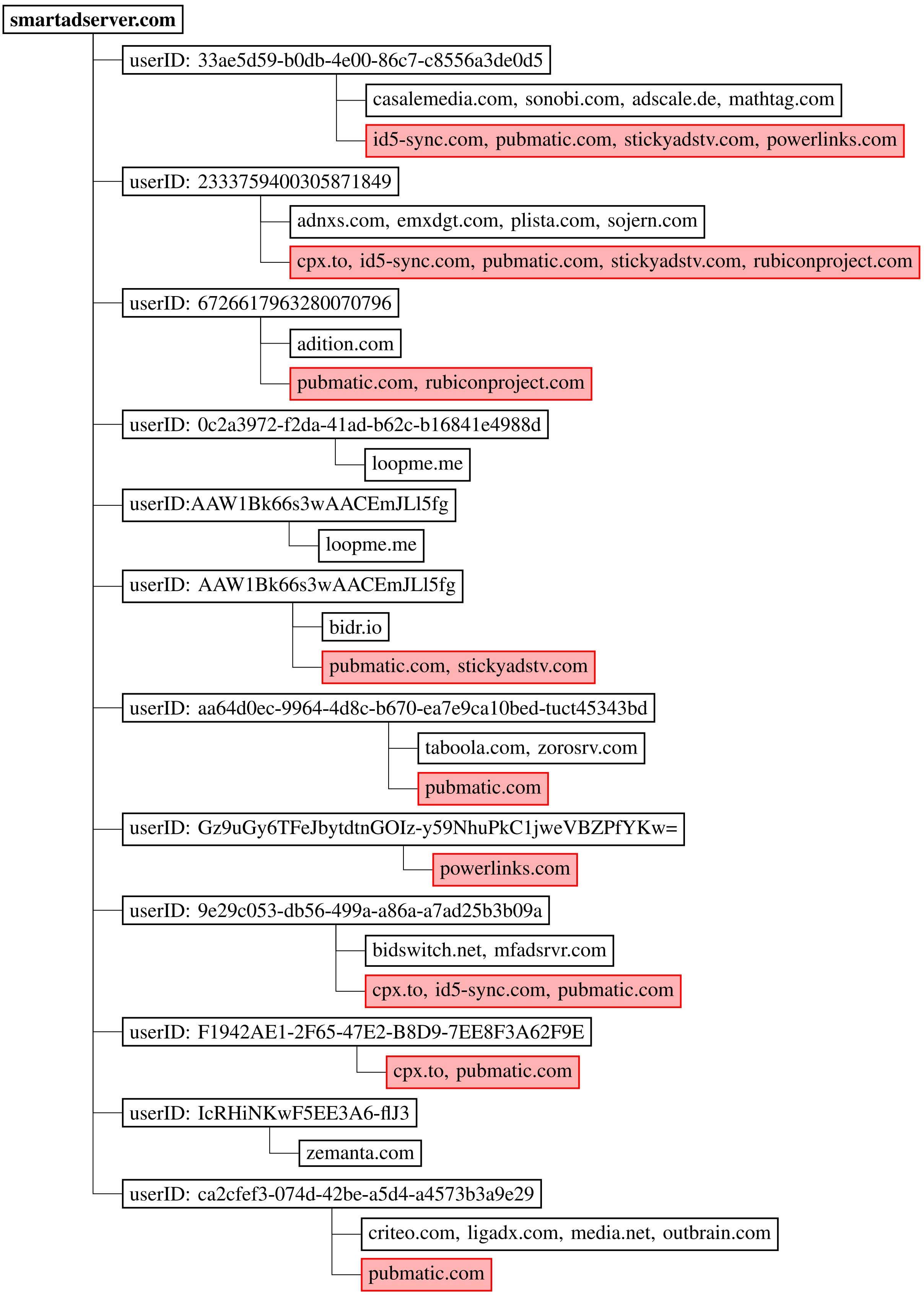

smartadserver.com
userID: 33ae5d59-b0db-4e00-86c7-c8556a3de0d5
casalemedia.com, sonobi.com, adscale.de, mathtag.com
id5-sync.com, pubmatic.com, stickyadstv.com, powerlinks.com
userID: 2333759400305871849
adnxs.com, emxdgt.com, plista.com, sojern.com
cpx.to, id5-sync.com, pubmatic.com, stickyadstv.com, rubiconproject.com
userID: 6726617963280070796
adition.com
pubmatic.com, rubiconproject.com
userID: 0c2a3972-f2da-41ad-b62c-b16841e4988d
loopme.me
userID:AAW1Bk66s3wAACEmJLl5fg
loopme.me
userID: AAW1Bk66s3wAACEmJLl5fg
bidr.io
pubmatic.com, stickyadstv.com
userID: aa64d0ec-9964-4d8c-b670-ea7e9ca10bed-tuct45343bd
taboola.com, zorosrv.com
pubmatic.com
userID: Gz9uGy6TFeJbytdtnGOIz-y59NhuPkC1jweVBZPfYKw=
powerlinks.com
userID: 9e29c053-db56-499a-a86a-a7ad25b3b09a
bidswitch.net, mfadsrvr.com
cpx.to, id5-sync.com, pubmatic.com
userID: F1942AE1-2F65-47E2-B8D9-7EE8F3A62F9E
cpx.to, pubmatic.com
userID: IcRHiNKwF5EE3A6-flJ3
zemanta.com
userID: ca2cfef3-074d-42be-a5d4-a4573b3a9e29
criteo.com, ligadx.com, media.net, outbrain.com
pubmatic.com

Figure 16. Cookies of *smartadserver.com* are synchronized with 25 unique third parties, while those third-party websites in the red bar are the new sync center

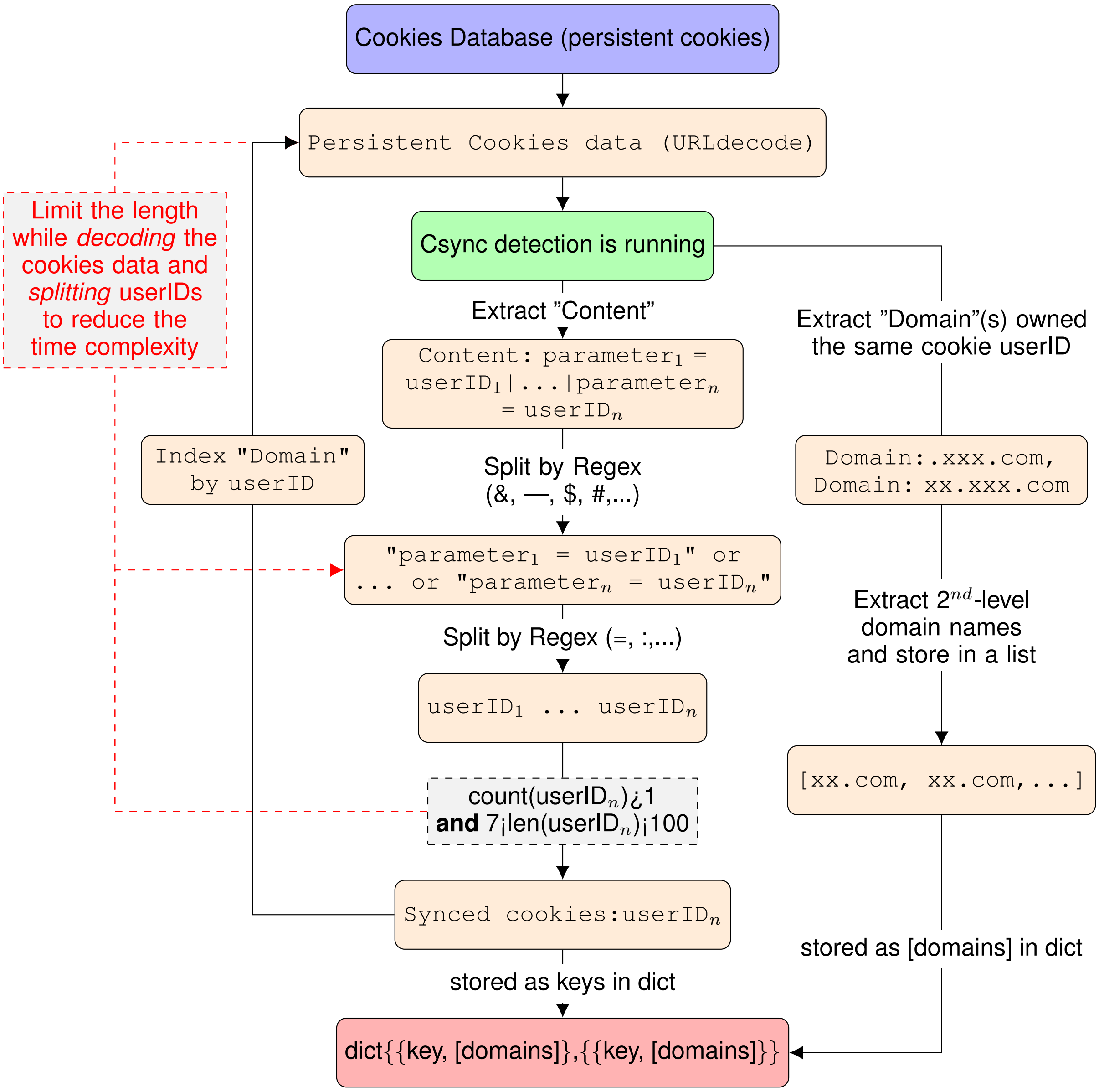

Cookies Database (persistent cookies)
Persistent Cookies data (URLdecode)
Limit the length while *decoding* the cookies data and *splitting* userIDs to reduce the time complexity
Csync detection is running
Extract "Content"
Content: parameter$_1$ = userID$_1$|...|parameter$_n$ = userID$_n$
Extract "Domain"(s) owned the same cookie userID
Index "Domain" by userID
Split by Regex (&, —, $, #,...)
Domain:.xxx.com, Domain: xx.xxx.com
"parameter$_1$ = userID$_1$" or ... or "parameter$_n$ = userID$_n$"
Split by Regex (=, :,...)
Extract 2$^{nd}$-level domain names and store in a list
userID$_1$ ... userID$_n$
[xx.com, xx.com,...]
count(userID$_n$)¿1 **and** 7¡len(userID$_n$)¡100
Synced cookies:userID$_n$
stored as [domains] in dict
stored as keys in dict
dict{{key, [domains]},{{key, [domains]}}

Figure 17. Detection of synchronized cookies