Web Tracking Cartography with DNS Records

Jingxiu Su*†, Zhenyu Li *, Stéphane Grumbach‡, Muhammad Ikram§, Kave Salamatian¶, Gaogang Xie*
*Institute of Computing Technology, China
†University of Chinese Academy of Sciences, China
‡Inria, France
§ University of Michigan, USA and Macquarie University, Australia
¶LISTIC, University of Savoie, France

Abstract—**Web tracking plays a crucial role in the Web ecosystem. It relies on third-party tracking domains collecting user information for various applications such as advertisement and analytics. With the massive growth of the Internet, understanding tracking and its geographical roots is of strategic importance. The goal of this paper is to propose a thorough investigation of web tracking inside China taking advantage of a large dataset (10^{11} records) containing two days of full DNS access from a major ISP providing both mobile and landline ADSL. Our results show that a power law applies on the traffic of both sites and trackers with a handful of trackers, 26, representing 90% of tracking activity. We then show that although most first-party sites accessed from China are owned by Chinese corporations, large proportion of trackers belong to US ones. This raises concerns about the analytics industry in China, and more generally shed new lights on the international data flows, the interdependency of the main actors, and the complexity of the threats for both people and states.**

I. Introduction

Web tracking is used to collect and correlate user web browsing behavior [1]. Such information is of interest to various parties: advertisement companies like Google AdSense [2] actively collect information about users to tailor personalized advertisement; web applications might benefit from tracking information to foster better design [3]; web analytics, like Google Analytics [4], also leverage tracking information to provide usage information. Tracking information might also be used by authorities to implement surveillance of targeted persons. More generally, data gathered on Internet users browsing behavior represent a source of strategic information that have both economic and political value.

Krishnarmurthy and Wills [5] provided an early insight into web tracking and showed that third-party trackers activities grew from 2005 to 2008 from 10% to 60% of web sessions. Studies show a continuous increase in the following years of third-party trackers both in activity volume and in the diversity of tracking techniques [6-11]. Previous studies showed the domination of the tracking market by a small number of mainly US based corporations in almost all countries [5, 7, 12, 13].

Although much less studied than the US or European countries, China is a particularly important example [14], with the largest internet market in the world, with more than 731 million Internet users, accounting for more than 25% of World Internet users [15]. China has a specific Internet market that is for part shaped by the implementation of very expansive content filtering, the Golden Shield or Great Firewall of China (GFW), and a protectionist policy privileging local Internet actors to international ones. Some preliminary results on China showed that while the traffic is mainly targeted towards local Chinese sites, there was a majority of US trackers [16]. Nonetheless, these observations were relying on shallow data sources and observations made from abroad.

Because of the importance of the Chinese market and its specificities, better measures are needed to assess it. In particular, it would be interesting to measure the impact of China protectionist policies together with the blocking strategy of the GFW of several major actors of online advertisement like Google or Facebook, on the tracking market. Moreover, evaluating the volume of information relative to Chinese users browsing behavior transferred oversea might reveal surprising figures.

We use a very unique dataset that is very rarely available for this type of research: a large scale Domain Name System (DNS) logs dataset from an ISP in China containing 150 billions records covering all regions of China. DNS is a decentralized system in charge of assigning IP adresses of authoritative servers to domain names. Almost all network services depend on DNS and leverage on its infrastructure [17]. This makes DNS a major source of information that enable to observe *in vivo* global network usages. While all previous research on trackers have mainly used sampling, such as sampling from a pool of users that have instrumented their browsers like for Alexa [18], or by sampling destination websites and looking for trackers on them, the DNS records give a direct and comprehensive view of the web tracking activity.

Nevertheless, dealing with DNS traces presents some technical challenges. First of all, it is necessary to analyse the data and overcome the impact of the use of Network Address Translation (NAT) middleboxes by the DNS. Then it is necessary to proceed to alleviate the impact of DNS caching. Finally, another challenge is relative to privacy and ethical issues when dealing with such traces.

The state of the tracking market in China is particularly interesting since China with its most stringent protectionist policy on the Internet, might be considered as a lower bound of the dominance of the main tracking *actors*. Therefore, by

978-1-5386-6808-5/18/$31.00 2018 IEEE

looking at the Chinese market, one might conjecture that the situation in other countries is much heavier. This paper deals on the state of web tracking in China and makes the following contributions:

1) We propose methodologies to alleviate the impact of DNS caching and NAT in processing DNS log records.
2) Our observations confirm the extreme concentration of the tracking market into a small number of companies. We present a detailed analysis of the tracking behavior of the main trackers.
3) We observe that while Chinese web activity is strongly concentrated inside China with more than 75% of web sessions going to Chinese web services, yet around 87% of tracking activity is ensured by US trackers. This share is almost alike on Chinese and US sites.
4) We also present an analysis of the information collected by trackers and categorize them into different platforms. We find that US and Chinese trackers actively collect information that are roughly of the same type.

The rest of the paper is organized as follows. We develop the challenges of dealing with DNS data and present a methodology for alleviating the impact of DNS caching in Section II. We identify and characterize activities of main tracking actors in Section III. We look at the typology of information collected by trackers in Section IV. We further investigate the geolocations of tracking activities in Section V. We survey the related works in Section VI. Finally, we discuss some implications of the major findings in Section VII.

II. DNS PROCESSING CHALLENGES

In this section, we describe the DNS dataset along with the challenges processing DNS data entails. First of all, it is necessary to analyse the data and overcome the impact of the use of Network Address Translation (NAT) middleboxes by the DNS. Then it is necessary to proceed to alleviate the impact of DNS caching. A detailed description of the methodology to deal with DNS processing challenges has been presented in our previous paper [19]. We will only present here a summary of the methodological steps.

A. DNS dataset and advertisers/trackers Labeling

The dataset consists of all the DNS requests and their resolution information received during two days by the DNS servers of a major mobile and ADSL ISP in China. The data are gathered from DNS servers located in different Chinese provinces and municipalities covering the whole country. The dataset contains about 150 billions DNS records in total, each having five fields: a timestamp (at second level precision), the "anomyzed" source IP sending the request, the domain name queried, the list of resolved IP addresses and a field indicating if the address resolution has been successful.

To study the tracking and advertisement market, we need to identify if a resource is relative to a tracker. For this purpose, we use an approach based on blacklists of filtering rules used for checking suspicious URLs by exact or wildcards matching. This approach is commonly implemented by largely used Ad suppressing utilities [20]. We combined blacklists obtained from Adblock Plus [21], Ghostery [22, 23] and Disconnect [24]. These blacklists are partly overlapping. We used in particular the specific China targeted blacklist from Adblock Plus [25] to ensure identification of all Chinese trackers. All these blacklists are used in practice and maintained up to date by their providers.

B. NAT issue

As we will see later, it is important for the analysis to ensure that at each instant of time only a unique user is using a given IP address. However, Network Address Translation (NAT) middleboxes enable sharing a single routable IP address between different users [26], which goes against our need. We observed that the distribution of the average DNS requests rate per IP address per second spans almost 6 orders of magnitude from 5×10^{-4} to 181, exhibiting a very heavy tail, *i.e.*, some IP addresses are largely contributing to the average. We also observed in the dataset an unexpectedly high value of average DNS requests rate per source IP address of 4.01 per second. We can explain this observation by NATed addresses used in particular for mobile network and shared by several users.

We use the average and maximum values of DNS requests rates to separate the IP addresses into NATed and nonNATed category. We fit the joint distribution of the average and maximum values of DNS requests rate to a mixture of correlated General Gamma (GG) Distributions [27]. We have further used a maximum Akaike information criterion [28] which gives that 2 mixture components are enough to classify the observations.

Our dataset contains mobile and ADSL users, the result shows that for both IP address categories, there is a strong differentiation between the two classes: one class generates a very low average DNS requests rate, while the other generates a much larger DNS requests rate (0.012 for mobile users and 0.046 for ADSL users). We observe similar differentiation for the maximum rates. We consider the first class to be compatible with a single user while the second class is definitely mixing multiple users. The above observations lead us to use the IP addresses detected in the first class for further analysis.

C. User-session's extraction

The typical user browsing behavior on Internet consists in activity periods, where the user browses the Internet, alternating with silent periods over which the user is not active. The active period will be coined through the paper as *user-session*. A user-session might consists of several TCP connections, opened by the same host toward possibly different servers. For example, a web session will contain all connections made to download objects embedded in a web page. The concept of user-session is also relevant to other applications beside web [29]. We will describe a generic user-session extraction method based on a threshold θ.

We first order the DNS dataset by source IP addresses and generate for each observed IP address a temporal sequence of DNS requests. In the second step, we split each IP source address sequence's into an alternance of user-sessions and inactivity periods. We therefore consider user-sessions with the following structure. An initial site request which is made to a service/content provider. This initial request is followed by a sequence of forthcoming requests to other servers. As we are interested in trackers, we will consider only connections made to servers detected as tracker/advertiser by the blacklists we describe in Section II-A.

A user-session begins with an access to a website and it is closed when the next DNS request arrives later than the threshold θ. In between, any DNS request arriving is aggregated into the same user-session. We derived the optimal value of θ as 5 secs through a mixture of distribution of *the DNS request gap*, *i.e.* the arrival time between DNS requests.

In summary, a user-session relative to a user k (more precisely relative to an IP address with a single user behind it) and beginning at time t is a set $S_i^k(t)$ containing the canonical name of the first non-tracker server contacted during the user-session followed by a sequence of tracker domain names, *e.g.* replacing `ads.flurry.com` and `data.flurry.com` with `flurry.com`, contacted during the same user-session.

D. Cache issue

The second challenge we have to address is relative to the impact of DNS cache. Generally, only the first DNS request of a client is answered by the DNS server and subsequent requests are answered from cache. This means that local cache filters out a relatively large proportion of DNS requests that never reach the ISP servers that we are monitoring, *i.e.*, any observation using DNS traces is a partial sampling of real Internet activity [30]. We proposed a rescaling methodology to solve the above cache issue based on user-session structure, which we will detail below.

1) Rescaling method: In [30], it is explained that no method can retrieve the rate of lost DNS requests. We can thus at best try to alleviate the effect of caches and to rescale the values. In Sec. II-C we described how we have split the DNS requests into sets $S_i^k(t)$. DNS caches make the set incomplete $S_i^k(t)$ as cached request will not appear in it, *i.e.*, different user-sessions going to the same content/service might contain different but still incomplete list of contacted trackers. We can leverage these differences and merge the different sets $S_i^k(t)$ relative to the same initial site `site1`. This will results into an inflated set of trackers that will contain all trackers that have been missed through the DNS caching. We can rescale the DNS observations, by replacing this merged set in place of any set $S_i^k(t)$ beginning with the `site1`.

2) Validating the rescaling method: In order to validate the rescaling method described earlier, we have compared the set of trackers we obtain with the set predicted by the LightBeam tool [31], an add-on for browsers that displays third-party tracking cookies placed on the user's computer while visiting various websites. We visited a sample of 809 sites extracted from the DNS dataset, then used Lightbeam to extract the set of trackers for each site. The comparison between the two sets shows an overlap ratio (ratio of the size of the intersection of the two sets to the size of LightBeam obtained set) equal to 91.7%, validating our rescaling method.

III. Tracking the trackers

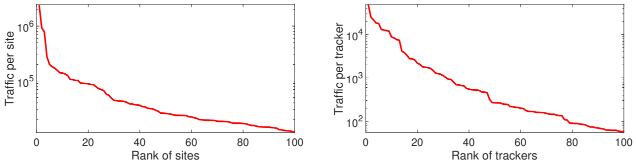
In this section, we first consider the tracking activities at a global level. We then identify the main sites and the main trackers, which represent a very large part of the global traffic. Finally, we analyse their tracking behaviors.

A. Global perspective on tracking

In the present investigation the measure of a site or tracker's traffic is defined as the number of its occurrences in the dataset after rescaling. 1 shows the relative traffic of sites and trackers according to this definition. Sites and trackers are sorted by descending order of traffic importance. Since there is a large number of sites and trackers in the data, the figure is limited for clarity to the top 100.

The traffic drops very fast. The top site or tracker has thousand times more traffic than the top 100. As expected the traffic follows a power law. A handful of sites represent the largest share of the global traffic. Such a phenomenon had already been observed on Alexa ranking for instance, and is fully confirmed by the DNS data.

We also consider for the top 100 sites, the number of distinct trackers per site. 2(a) shows that the sites which have more traffic seem to attract more trackers. Interestingly, the site ranked 96th, which is ”microsoft.com”, has much more trackers than sites with similar traffic. This may be due to the various related products, like online stores, softwares, mobile devices.

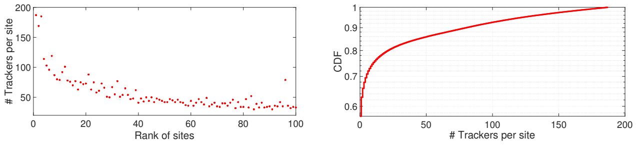


(a) Traffic per site.

(b) Traffic per tracker.

Fig. 1: Traffic per site/tracker (Top 100 are present).

The investigation for all sites is shown in 2(b). The number of trackers per site ranges from 0 to nearly 200. While more than half of the sites have no trackers, 90% of the sites have less than 100 trackers, about 80% of the sites have less than 20 trackers, and 70% of the sites have less than 5 trackers. If we consider again 1, it should be noted that the top 100 sites concentrate most of the tracking activity.



(a) Top 100 sites.

(b) All sites.

Fig. 2: Number of trackers per site.

B. Main actors

It follows immediately from the analysis above, that the traffic of both sites and trackers decreases very rapidly. Only a few actors have a strong influence. They are those we need to better understand.

We consider the sites and trackers whose individual traffic amounts to at least 0.5% of the global traffic. In the dataset, 28 sites and 26 trackers satisfy this requirement. They collectively represent 67% of the global traffic for sites and 90% of the global traffic for trackers respectively.

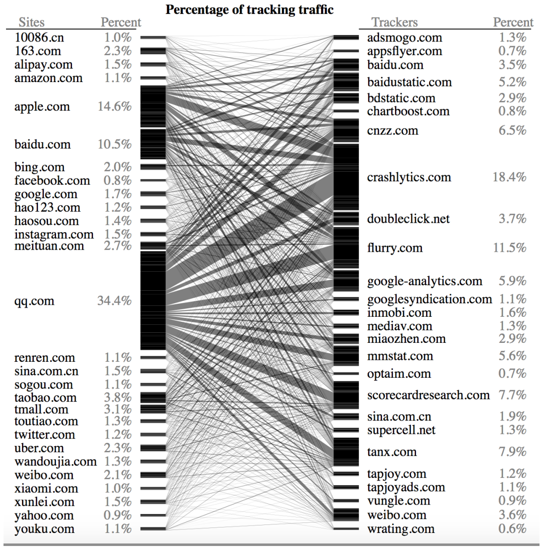


Fig. 3: Bipartite graph of tracking traffic between the 28 top sites and the 26 top trackers.

3 presents a static image[1] of a dynamic d3.js [32] bipartite graph between the 28 sites (on the left) and the 26 trackers (on the right). For each site, say `qq.com`, its tracking traffic equals the sum of the number of times each of the 26 trackers occurs in a session of qq.com. Note that this number can be much bigger than the number of occurrences of `qq.com` itself in the data, since in each session there are many trackers. The share, 34.4%, of the tracking traffic of this site among all 28 sites is shown in 3.

For each tracker, such as `doubleclick.net`, its tracking traffic equals to the sum of the number of times each of the 28 sites occurs in a session which contains `doubleclick.net`. The share of this tracker among all 26 trackers, 3.7%, is shown in 3. The size of each bar, associated with sites and trackers, is logarithmically proportional to their corresponding traffic, to accommodate the image with the fact that the largest traffic can be hundred times larger than the smallest.

In the online dynamic d3.js visualization, it is possible to access the details for each site or tracker by simply clicking on it.

3 shows the high level of connection among top actors. It is essentially a complete bipartite graph. Each site has almost all top 26 trackers tracking on it, but with different levels of tracking traffic, and vice versa, each tracker can track almost all top 28 sites. This observation confirms previous results showing how the network is heavily dominated by only a small group of actors, which are highly connected [12, 33].

IV. Informations collected by trackers

In Section III, we identified the top trackers. In this section, we concentrate on the information they collect from users.

Trackers use various mechanisms such as third-party libraries and JavaScripts APIs to enable tracking both from mobile Apps as well as from web sites. We classify the trackers into categories depending upon their activities and the garnered attributes. Various studies [34-36] have been carried on that investigate and classify third-party ads/tracking libraries found in mobile apps and JavaScript code APIs embedded in web sites. We consider the trackers in our dataset together with the numerous garnered attributes — often protected by permissions [37, 38] to target users or profile users' activities.

Table I shows the list of the top 26 trackers. We found that respectively 16 (62%) and 15 (58%) of the top 26 trackers track users on mobile resp. web platforms, while 5 (19%) of them, including Google Analytics and Supercell, are performing "cross" platform tracking, i.e., tracking users on both web and mobile platforms.

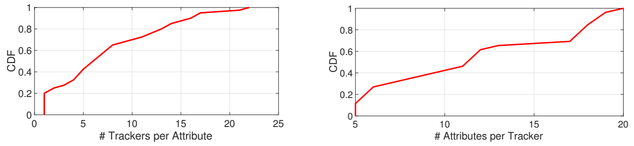
#	Trackers	Cat.	Mobile	Web
1	crashlytics.com	Utility	✓	
2	flurry.com	Analytics	✓	
3	tanx.com	Ads		✓
4	scorecardresearch.com	Analytics		✓
5	cnzz.com	Analytics	✓	✓
6	mmstat.com	Analytics		✓
7	weibo.com	Widget	✓	✓
8	baidustatic.com	Ads	✓	
9	google-analytics.com	Analytics	✓	✓
10	doubleclick.net	Analytics		✓
11	baidu.com	Search engine	✓	
12	bdstatic.com	Analytics		✓
13	miaozhen.com	Ads		✓
14	mediav.com	Ads		✓
15	sina.com.cn	Widget		✓
16	inmobi.com	Ads	✓	
17	supercell.net	Analytics	✓	✓
18	adsmogo.com	Ads	✓	
19	googlesyndication.com	Ads	✓	✓
20	tapjoy.com	Analytics		✓
21	tapjoyads.com	Ads	✓	
22	vungle.com	Targeted ads	✓	
23	wrating.com	Ads		✓
24	optaim.com	Ads	✓	
25	chartboost.com	Ads	✓	
26	appsflyer.com	Ads	✓	

TABLE I: List of mobile and web trackers with their category (Cat.) ordered by decreasing traffic in our data. Widget means social network widgets.

We then consider the attributes collected by the top 26 trackers. Following an approach pursued in [34, 36], we comprehensively survey three vantage points to extract attributes collected by trackers: (i) the Java API for the 16 mobile trackers, (ii) the JavaScript codes for the 15 web trackers, and finally (iii) the "privacy policies" of all 26 trackers. Since trackers may collect more attributes or enrich them by further

[1] The visualization is available online at: http://bl.ocks.org/WebTrackingCartography/raw/e59cfc5870d6ec8990a30e05fac72f74/

combination with other data, we obtain merely a lower bound on the garnered attributes per tracker.



(a) (b)

Fig. 4: CDFs of number of trackers per attributes, 4(a), and attributes per trackers, 4(b).

Fig. 4(a) shows the distribution of the number of trackers per attribute. We observe that 20% (8) of the attributes are collected by a unique tracker. For instance, only one tracker, cnzz.com, collects the attribute market_id, which provides information about users app marketplace, such as Google Play or iOS App store (cf. Table II). While 60% of the trackers are collecting at least 5 attributes. A closer look at the top-right of the curve reveals that the attribute u_time — representing date and time info — is the most collected attribute, garnered by 22 (85%) of the top 26 trackers. Similarly, user's device IP address and international mobile equipment identity (IMEI) number are respectively the second and third most collected attributes. In fact, as we could verify, all cross device/platform and mobile trackers access to device IMEI number.

Next we consider the number of attributes collected by each tracker. In Fig. 4(b), we observe that 61% (16) of the trackers collect at least 10 attributes. The top-right corner reveals that about 15% (4) of the trackers are collecting at least 19 attributes. They include: cnzz.com, analytics.google.com, doubleclick.net, and inmobi.com. While 12% (3) of trackers (sina.com.cn, mmstat.com, and weibo.com) are collecting at most five attributes.

V. Geography of tracking

In this section, we consider the trackers from the point of view of the geography. Our goal is to better understand the advertisement market, and at the same time increase the knowledge of the global data flows.

Determining the country of origin of a tracker is not completely easy. An advertisement server owned by a French subsidiary of a Chinese company, running over a physical infrastructure located in a data-center in the Netherlands and managing advertisement traffic sent to Russia is a plausible scenario that one might be confronted to in the area of trackers.

For simplicity but without lack of relevance, we assign a tracker service to the country that is registered in the WHOis database along with the corresponding canonical domain name. The WHOis database contains contact information for administrative and technical contact points along with the country. We therefore assign to each tracker the country reported in the WHOis database to the corresponding DNS canonical domain name entry.

#	Attribute	Count	Description
1	u_time	22 (85%)	date and time
2	ip	21 (81%)	IP address
3	os_info	17 (65%)	OS info, version, type
4	dev_info	17 (65%)	device or hardware type, model
5	imei	16 (62%)	IMEI number
6	loc	15 (58%)	geo-location i.e., GPS info
7	cookie_info	14 (54%)	cookie info
8	lang_id	14 (54%)	language
9	browser_info	13 (50%)	browser (agent) info
10	ad_view	13 (50%)	ads view and interaction with ads
11	interaction_data	13 (50%)	post-click activity, start/boot-up info
12	brow_hist	11 (42%)	browsing history and analytics
13	isp	10 (38%)	internet service provider
14	apps_list	9 (35%)	list of user installed and running apps ids
15	email_id	8 (31%)	email id
16	aaid	8 (31%)	amount played/session length information
17	session_info	8 (31%)	Android advertising identifier
18	idfa	7 (27%)	iOS advertising identifier
19	mac_id	7 (27%)	mac address
20	time_zone	7 (27%)	time zone
21	dev_stats	6 (23%)	devie stats e.g., CPU and battery usage
22	p_view	6 (23%)	demographic info e.g., gender, age
23	search_hist	6 (23%)	Errors or Page Views
24	demo_info	5 (19%)	search queries history
25	p_address	5 (19%)	post address or zip code
26	wifi	5 (19%)	wifi network and its status
27	friendlist	5 (19%)	contacts phone or email ids
28	phone_number	4 (15%)	phone number
29	user_id	4 (15%)	user id
30	c_domain	3 (12%)	current serving domain
31	wifi_info	2 (8 %)	wifi network and its status
32	crash_info	2 (8%)	crash event
33	cd_hist	1 (4%)	cross_device tracking
34	scookie_info	1 (4%)	persistent cookie id and data
35	action_info	1 (4%)	session cookie id and data
36	pcookie_info	1 (4%)	persistent cookie
37	apps_versions	1 (4%)	version of applications,
38	bluetooth_info	1 (4%)	Bluetooth stats
39	cr_hist	1 (4%)	bluetooth network and its status
40	market_id	1 (4%)	GPlay or iOS marketplace ID

TABLE II: List of tracker attributes, with their frequency and description.

A. Inequality between countries

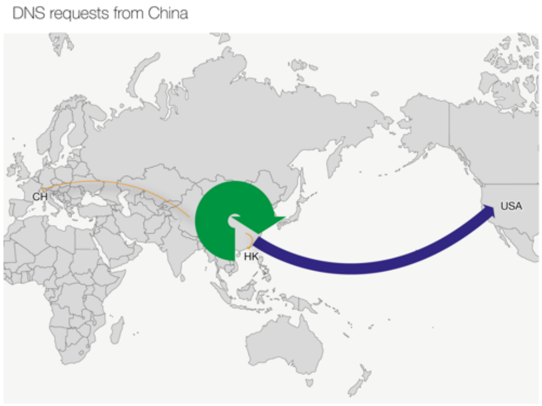


Fig. 5: Traffic share between countries from China.

The traffic resulting from the DNS traces we have analyzed can thus be attached to destination countries. More precisely, the traffic load of a country is measured as the number of DNS requests that resolve to an IP address in this country, or more precisely to an IP address that belongs to a corporation based in this country. As can be seen in Fig. 5, China is the destination of more than 73% percent of the traffic from China; while the US account for 24%. All other countries account for less than 3% of the whole traffic.

When we consider instead the tracker traffic, we observe a rather different trend. We show in Fig. 6 the distribution of the traffic to tracker services worldwide. It can be observed that the US is attracting more than 87% of all the tracker traffic from China. The second rank is occupied by the UK with 7.2% of tracking traffic, while China itself only occupies the third rank with 3.2% of the tracker traffic on its own territory. These results confirm trends observed previously on a different dataset in [16], where it was shown that China dominates its local Web with more than 80% of local sites, while these sites contain a majority of US trackers.

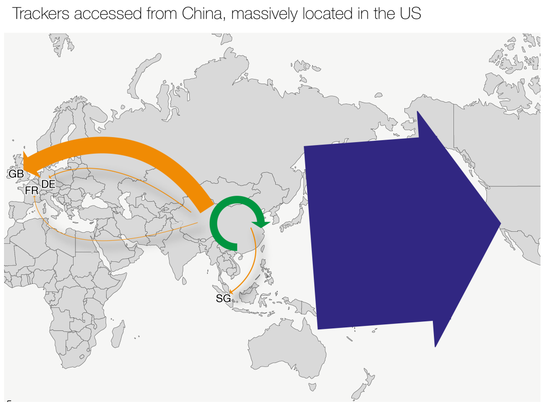


Fig. 6: Traffic share of tracking services from China.

It should be noted that this surprising situation holds despite the fact that China has a rich advertisement and e-commerce ecosystem.

B. Analysis of main sites and their trackers

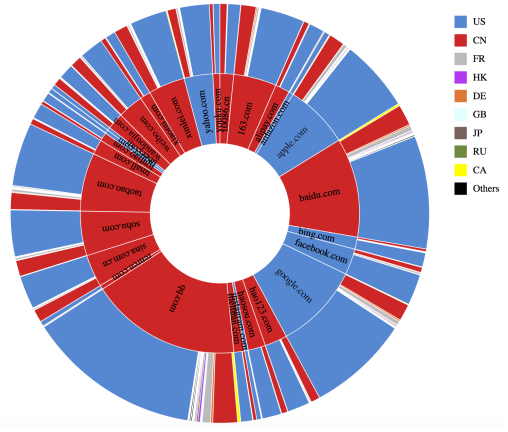


Fig. 7: Geography of the top 28 sites and their trackers.

We consider in Fig. 7 the top 28 sites identified in Section III, displayed in a bilevel partition, obtained using d3.js [32]. The inner ring shows the share of the sites in China. Among these sites, 19, in red, belong to Chinese corporations, and 9, in blue, to US ones. The outer ring, associates to each domain of the inner circle, the trackers related to this site, classified by country following the same convention.

The bilevel partition allows to navigate dynamically in the image to obtain more detailed information on each actor. The demo allows, by a simple click, to expand each sector to visualize detailed information, for all the sites and corresponding trackers. It is accessible online[2].

The similarity between the distribution of trackers on sites of both countries is striking. For example, `qq.com` and `google.com`, while they carry on different activities, `qq.com` is a large social platform in China, and `google.com`, hosts the largest global search engine in the world, have similar tracking patterns. In fact, such a distribution is not uncommon, and is shared by most sites.

C. Attributes collected in each region

We next carry on the analysis by considering the number of collected attributes by trackers in relation with the different geographic regions they belong to.

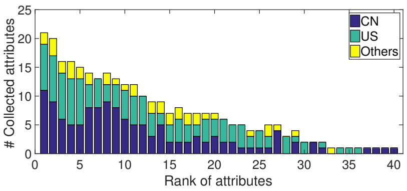


Fig. 8: Proportion of collected attributes in each region.

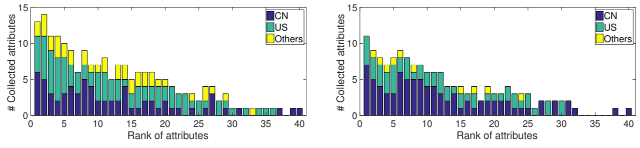
Fig. 8 shows the number of times each attribute has been collected, grouped by the different countries, to which the tracker collecting the attribute is affiliated. The rank of each attribute corresponds to the one of Table II. Note that a similar rank will be used in the sequel for Fig. 9 and Fig. 10. The occurrence count for each attribute is displayed on each bar. Since the top 26 trackers are mostly held by China and US, for simplicity, we only distinguish between three regions: CN, US and Others.

Trackers from China are extracting more attributes such as `u_time`, `ip`, `lang_id` than `scookie_info`, `action_info`, `pcookie_info` for instance. Trackers from US and countries other than China, collect mostly attributes among the most frequent, that is ranked high in Table II. Chinese trackers are also extracting more `loc`, `cookie_info`, `lang_id`, `browser_info` and other similar informations than trackers from the US, which are relatively more active on attributes such as in `dev_info`, `imei`, `search_hist`, compared to trackers from China.

Fig. 9 presents a refinement of the analysis by distinguishing between the trackers in mobile and Web platforms, while the results split by the categories of activity presented in Table I are shown in Fig. 10.

Trackers from outside China and US are mostly mobile trackers. They collect mostly attributes with high ranks. US

[2] The visualization is available online at: http://bl.ocks.org/WebTrackingCartography/raw/f2bca61a0780f47dca5f618700d76065/



(a) Mobile

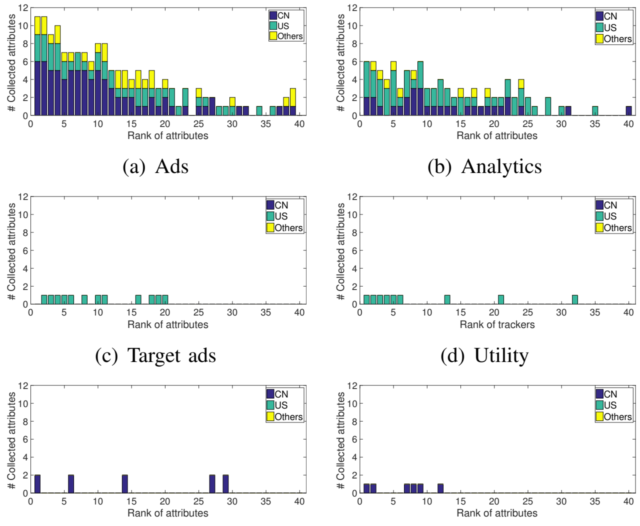
(b) Web

Fig. 9: Proportion of collected attributes for mobile or Web.

trackers are more active on mobile platforms as well. Moreover, they collect more attributes than their Chinese counterparts on mobile. The situation is more balanced for Web platforms.

Fig. 10 displays the number of collected attributes for six different categories corresponding to the categories of the trackers. The figures show that a majority of the attributes are used by trackers for ads and analytics activity. Our results show indeed that most attributes are collected by these trackers. Chinese trackers are more present for ads, while US trackers have more interest on analytics. Trackers used for target ads and utility are US trackers, as shown in Fig. 10(c) and Fig. 10(d). They collect `ip`, `os_info`, `dev_info`, `imei` and `loc` from the users.

Trackers used for resp. widgets and search engine are shown resp. in Fig. 10(e) and Fig. 10(f). They collect different kinds of attributes, widgets target attributes such as in `u_time`, `loc`, `apps_list`, `friendlist` and `user_id`, because they make great use of contact information (`friendlist`), location (`loc`), etc. to help share content on social platforms. While search engines collect attributes such as `u_time`, `ip`, `cookie_info`, `lang_id`, `browser_info` and `brow_hist`, to track users' browsing behavior. It is no surprise that all these attributes are collected by Chinese trackers, since US social platforms and search engine have little penetration in the Chinese market.



(a) Ads

(b) Analytics

(c) Target ads

(d) Utility

(e) Widget

(f) Search engine

Fig. 10: Proportion of collected attributes by category.

VI. Related works

Web tracking ecosystems have been well analyzed by a number of studies, focusing on behavioral and privacy aspects. Krishnamurthy et al. [5] present the results of a longitudinal measurement of web tracking and prevalence of trackers, then they examined the access of third-parties to personal information, and they found leakage in sites for every categories they have obtained [39]. Roesner et al. [7] made a classification between different types of web trackers and measured the prevalence of these tracker classes amongst the top 500 websites in the world. In paper [16], Castelluccia et al. analyze the provenance of most important third party tracking services using two popular browser extensions, for the geographical classification. They focused on measuring the penetration of US-based services in different countries. Gomer et al. [12] show a consistent network structure across different markets as well as high efficiency in exchanging information among third-parties. Mayer et al. [6] surveyed different techniques which are used by web trackers to collect user information. In [13], Falahrastegar et al. crawl the top websites in Alexa rank for different countries, and measure the per-country pervasiveness of third party trackers. The above studies either focused on the analysis of specific types of third-party trackers or on a geographical classification but without much consideration of China.

VII. Conclusion and Discussion

This work has been made possible thanks to the access to an exceptionally large DNS trace of 10^{11} records of two days of activity in China. The first issue has been to develop new methods to accurately analyze the data, and get around the optimized behavior of the DNS system to recover a realistic view of the real traffic, and recollect sessions.

Our results confirm the extreme concentration of online services into a small number of corporations. We focus on the actors that have at least 0.5% of the traffic, that is only 28 sites representing 67% of the whole sites traffic, and 26 trackers representing 90% of the whole trackers traffic. Interestingly, while China dominates its web of services, with around 3/4 of national services, around 87% of tracking activity is ensured by US trackers over Chinese and US sites alike.

A first question concerns the reasons of such an imbalance. Several causes can be considered. The first one could be related to the open-source development frames, such as Android or WordPress, which offer sometimes by default US based advertisers/trackers. A second reason might be related to economic arguments. The pay-per-click model used by some major advertisement actors is very attractive and the offer proposed by Chinese actors are not matching it. Another cause might be related to the focalisation on other platforms, like WeChat, that is controlled by Tencent, which is also the main actor for online advertisements in China.

The observation that both trackers and tracking traffic in China are dominated by US corporations have two very important implications. Firstly, as US actors are mostly not installed into the mainland Chinese network, this means crossing the GFW for access to tracker and having long-distance interactions when loading webpages or mobile apps. This may negatively affect the loading performance. Deploying replicas of the trackers within China will be helpful to alleviate this problem, however this goes against some Chinese regulations. Secondly, this observation indicates that US corporations may have a better view of Chinese users behavior than China does because US and Chinese trackers collect information that are roughly of the same type, but not to the same extent.

This raises huge concerns on advertising market, user privacy and cyber security. Enforcement of data protection law similar to Western countries [13] to the Chinese Internet could be an option to address these issues, as in this case, deploying local replicas of trackers could become mandatory. This is indeed what just happened, with the Chinese cybersecurity law enforced since June 2017, which requires mandatory in-country data storage. We will investigate the impact of this law on web tracking in future work.

Another overlooked issue is that currently the economy of Internet is heavily based on advertisement revenue to provide free access to services. The fact that a non-negiglible part of this revenue is diverted to a foreign country might have an impact on the economic eco-system of Internet in China. On the other hand, the present situation offers great opportunities for domestic online advertising companies in China. They can first exploit the niche market because, as we found, there are huge numbers of moderately popular sites, that have none or very few embedded trackers. Another opportunity could be to compete with the major US actors with the advantage of keeping the tracked data locally.

For individual users, the existence of pervasive trackers imposes a privacy concern. But as far as we know, tracker blocking tools in China are not as common as in Western countries, despite that some tools, like Adblock, provide specific lists targeted for Chinese trackers. Yet, we have limited knowledge of the usage of trackers blocking tools in China, which is left for future work.

VIII. Acknowledgement

This work is supported by National Key R&D Program of China (Grant No. 2016YFE0133000): EU-China study on IoT and 5G(EXCITING).

References

[1] N Schmucker. Web tracking. In *SNET2 Seminar Paper-Summer Term*, 2011.

[2] Google adsense. https://www.google.com/adsense.

[3] Richard Atterer et al. Knowing the user's every move: User activity tracking for website usability evaluation and implicit interaction. In *WWW '06*.

[4] Google analytic. http://google.com/analytic.

[5] Balachander Krishnamurthy and Craig Wills. Privacy diffusion on the web: a longitudinal perspective. In *International Conference on World Wide Web*, pages 541–550, 2009.

[6] Jonathan R Mayer and John C Mitchell. Third-party web tracking: Policy and technology. In *SP'12*, pages 413–427, 2012.

[7] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. Detecting and defending against third-party tracking on the web. In *NSDI'12*.

[8] Chris Jay Hoofnagle and Nathan Good. Web privacy census. 2012.

[9] Gunes Acar et al. Fpdetective: dusting the web for fingerprinters. In *CCS'13*. ACM, 2013.

[10] Gunes Acar et al. The web never forgets: Persistent tracking mechanisms in the wild. In *CCS'14*, pages 674–689. ACM, 2014.

[11] Ibrahim Altaweel, Nathan Good, and Chris Jay Hoofnagle. Web privacy census. 2015.

[12] Richard Gomer et al. Network analysis of third party tracking: User exposure to tracking cookies through search. In *WI-IAT '13*, pages 549–556.

[13] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. Anatomy of the third-party web tracking ecosystem. *arXiv preprint. arXiv:1409.1066*, 2014.

[14] Stéphane Grumbach. The stakes of big data in the it industry: China as the next global challenger? In *The 18th International Euro-Asia Research Conference, The Globalisation of Asian Markets: im-plications for Multinational Investors, Venezia, 2013*.

[15] https://hostingfacts.com/internet-facts-stats-2016/.

[16] Claude Castelluccia, Stéphane Grumbach, and Lukasz Olejnik. Data harvesting 2.0: from the visible to the invisible web. In *The Twelfth Workshop on the Economics of Information Security*, 2013.

[17] David Pariag and Tim Brecht. Application bandwidth and flow rates from 3 trillion flows across 45 carrier networks. In *PAM'17*.

[18] Amazon company. The top sites on the web. http://www.alexa.com/topsites.

[19] Jingxiu Su, Zhenyu Li, et al. Toward accurate inference of web activities from passive dns data. In *IWQoS'18*, 2018.

[20] Akira Yamada, Hara Masanori, et al. Web tracking site detection based on temporal link analysis. In *WAINA'10*, pages 626–631.

[21] Wladimir Palant. Adblock plus: Save your time and traffic. https://easylist.adblockplus.org/en/, 2017.

[22] Ricardo Bilton. Ghostery: A web tracking blocker that actually helps the ad industry. 31:2012, 2012.

[23] Ricardo Bilton. Ghostery. https://www.ghostery.com, 2017.

[24] Disconnect. malvertising list. https://disconnect.me/lists/malvertising.

[25] Adblock plus easylist china. https://easylist-downloads.adblockplus.org/easylistchina+easylist.txt, 2017.

[26] Franois Audet and Cullen Jennings. Network Address Translation (NAT) Behavioral Requirements for Unicast UDP. RFC 4787, RFC Editor, January 2007.

[27] Petros S Bithas, Nikos C Sagias, Theodoros A Tsiftsis, and George K Karagiannidis. Distributions involving correlated generalized gamma variables. In *Proc. Int. Conf. on Applied Stochastic Models and Data Analysis*, volume 12, 2007.

[28] Wikipedia. Akaike information criterion. https://en.wikipedia.org/wiki/Akaike_information_criterion, 2017.

[29] Aaron Halfaker, Oliver Keyes, et al. User session identification based on strong regularities in inter-activity time. In *WWW '15*, 2015.

[30] Nicaise Choungmo Fofack and Sara Alouf. Modeling modern dns caches. In *ValueTools '13*, pages 184–193, ICST, Brussels, Belgium, Belgium, 2013. ICST.

[31] Lightbeam. Shine a light on who is watching you. https://www.mozilla.org/en-US/lightbeam/.

[32] ds.js. d3.js. https://bl.ocks.org/mbostock/5944371.

[33] Vito Latora and Massimo Marchiori. Efficient behavior of small-world networks. *Physical review letters*, 87(19):198701, 2001.

[34] Muhammad Ikram and Mohamed Ali Kaafar. A first look at ad-blocking apps. In *IEEE NCA*, 2017.

[35] Muhammad Ikram, Narseo Vallina-Rodriguez, et al. An analysis of the privacy and security risks of android vpn permission-enabled apps. In *IMC '16*.

[36] Muhammad Ikram et al. Towards seamless tracking-free web: Improved detection of trackers via one-class learning. *Proceedings on Privacy Enhancing Technologies*, 2017.

[37] Normal permissions — android developers. https://developer.android.com/guide/topics/permissions/normal-permissions.html.

[38] Chrome permissions. https://developer.chrome.com/apps/declare_permissions.

[39] Balachander Krishnamurthy et al. Privacy leakage vs. protection measures: the growing disconnect. In *Proceedings of the Web*, volume 2, pages 1–10, 2011.