Computer Communications 134 (2019) 83–95

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

# A Cartography of Web Tracking using DNS Records

Jingxiu Su [a,b], Zhenyu Li [a,b,*], Stephane Grumbach [c], Muhammad Ikram [d,e], Kave Salamatian [f], Gaogang Xie [a,b]

[a] *University of Chinese Academy of Sciences, China*
[b] *Institute of Computing Technology, Chinese Academy of Sciences, China*
[c] *Inria, France*
[d] *University of Michigan, USA*
[e] *Macquarie University, Australia*
[f] *LISTIC, University of Savoie-Mont Blanc, France*

## ABSTRACT

Web tracking plays a crucial role in the Web ecosystem. It relies on third-party tracking actors collecting user information that are used for various applications such as advertisement and analytics, *etc.* With the massive growth of the Internet, understanding the geography of tracking 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 strong Pareto principle applies on the traffic toward trackers, with only 26 trackers, 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 advertisement 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.

## 1. Introduction

Web tracking is used to collect and correlate user web browsing behavior [1]. Such information are 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, as the browsing activity leaks highly private information. 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 the presence of third-party trackers activities grew from 2005 to 2008 from 10% to 60% of web sessions. Studies show a continuous increase of the activity of third-party trackers in the following years both in term of volume and diversity of tracking techniques [6–11]. Previous studies, have also shown the domination of the tracking market by a small number of mainly US based corporations in almost all countries [5,7,12,13].

Nonetheless, the activities of web trackers in China have been much less studied than the US or European countries, despite it being 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 a very expansive content filtering architecture, 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 Internet market and its specifics, 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. What is observed over China regarding the dominance of the main tracking *actors* might be considered as a minimal view of what might be seen in other countries with less stringent protectionist policy on the Internet.

For this purpose we used a very unique dataset that is very rarely available for this type of research: two days of logs coming from the servers of the Domain Name System (DNS) of a large scale ISP with countrywide presence in China, containing 150 billions records covering all regions of China. DNS is a decentralized system in charge of translating given domain names into IP addresses. Almost all network

[*] Correspondence to: No.6, Kexueyuan South Road, Haidian Distinct, Beijing, China.
*E-mail addresses:* sujingxiu@ict.ac.cn (J. Su), zyli@ict.ac.cn (Z. Li).

https://doi.org/10.1016/j.comcom.2018.11.008
Received 22 September 2018; Accepted 28 November 2018
Available online 5 December 2018

services depend on DNS and leverage on its infrastructure [17]. While all previous research on trackers have mainly used sampling, Alexa [18] is sampling from a pool of users that have instrumented their browsers, or by sampling destination websites and looking for trackers on them, the DNS records give a direct and comprehensive view of the web activities, in particular tracking activity. For these reasons DNS is a major source of information to observe *in vivo* global network usages.

Nevertheless, dealing with DNS traces presents some technical challenges that have to be addressed. First of all, there is challenge of clustering DNS requests into sessions, *i.e.*, regrouping DNS requests that are directly related to the activity of a single user's session, *e.g.* browsing a single website. Addressing this challenge needs to overcome the impact of the use of Network Address Translation (NAT) that enable several users to share a single IP address and mix therefore the activity of several users. The second challenge is relative to the impact of DNS caching mechanisms, that ensure that a DNS request is not made if the answer is already available in local caches. This means that an operator DNS servers will not have an exhaustive view of all DNS requests made, and will only see the ones that are not filtered by the caches. Indeed, we have also to address some privacy and ethical issues that arise with the handling of such rich and sensitive datasets. Last but not least challenge is related to the volume of data. In the paper we will deal with a trace with 150 billions records. This means that the solution that will be applied to solve the above mentioned challenges should have a reasonable complexity in order to be applied on the data in a reasonable time.

This paper deals on the state of web tracking in China and makes the following contributions:

(1) We propose methodologies to cluster DNS requests into sessions and to alleviate the impact of DNS caching.
(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 sessions going to Chinese services, yet around 87% of tracking activity is ensured by US trackers. This share is almost alike in 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 processing DNS data and present a methodology for alleviating the impact of DNS caching in Section 2. We identify and characterize activities of main tracking actors in Section 3. We look at the typology of information collected by trackers in Section 4. We further investigate the geolocations of tracking activities in Section 5. We survey the related works in Section 6. Finally, we conclude our work and discuss some implications of our major findings in Section 7.

## 2. DNS processing challenges

In this section we will first describe the DNS dataset used in this study and propose methodologies to deal with the challenges of processing DNS data. It is noteworthy that the methodologies developed in this paper can be extended to other relevant problems besides DNS traces, like passive HTTP traces from ISPs' gateway that have similar issues.

### 2.1. DNS dataset and advertisers/trackers labeling

The dataset consists of all the DNS requests and their resolution information received during two days (in July 2015) 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 (see Table 1), each having five fields: a timestamp (at second level precision), the "**anonymized**" 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.

**Table 1**
Detail of dataset.

| Num records | Num IP | Num destinations |
|---|---|---|
| 149,619,580,908 | 18,507,392 | 711,660,375 |

The paper aims into studying the tracking and advertisement market, we have therefore to identify if a requested resource belongs to a tracker. For this purpose, we use a similar approach to the one commonly implemented by widely used Ad blocking utilities [19], that consists of using a blacklist of suspicious URLs that are matched with requested domain by exact or wildcards matching. In order to build the blacklists, we have combined lists obtained from Adblock Plus [20], Ghostery [21,22] and Disconnect [23] applications. To ensure identification of all Chinese trackers, we have used the specific China targeted blacklist from Adblock Plus [24]. All these blacklists are widely used in practice and they are maintained up to date by their providers. These blacklists are partially overlapping and we finally end up with a blacklist of 74855 domains relative to advertisement and tracker companies.

### 2.2. Extraction of user-sessions

The first step into the analysis of our DNS traces is to extract DNS requests relative to a user session. The typical user behavior on Internet consists in activity period, where the user browses the Internet, alternating with silent period over which the user is not active. The active period will be coined through the paper as *user-session*. A user-session might consist 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 the web page. The concept of user-session is also relevant to other applications beside web [25].

There exists a relatively rich literature on web user-session extraction [26–28]. Most of the existing works rely on exhaustive packet or connection level traces and assume that only a single user is behind an IP address. In our case we have only DNS traces and we need to leverage on them. We will assume in the forthcoming that we have ensured that there is a single user behind each analyzed IP address. We will describe later in Section 2.3 how we do this. Moreover, we need a method with low complexity so that it can be applied in a reasonable time.

In order to extract user-sessions, we first order for each observed IP address in the DNS trace a temporal sequence of DNS requests. In a second step, we have to split each temporal sequence into an alternation of user-sessions and inactivity periods. Several splitting approaches have been proposed in the literature. A simple approach proposed in [26,27] consists of choosing a time threshold $\theta$ used as the maximal time interval between two consecutive DNS requests belonging to the same session, *i.e.* whenever a DNS requests arrive later than $\theta$ sec after the previous one, the current user-session is finished and a new one is created. This approach has been shown to work well if the threshold value is correctly chosen. Other more sophisticated methods have been proposed based on hierarchical or agglomerative clustering [28]. However these complex methodologies are not applicable on the large volume of data we have to deal with in this paper (18,507,392 source IP addresses). Moreover, as we have the full trace in advance we can derive precisely the threshold $\theta$.

The choice of $\theta$ is done by looking at the distribution of *DNS request gaps*, *i.e.* the inter-arrival time between two consecutive DNS requests in the above defined temporal sequences. We will assume that the DNS request gaps are following a distribution that can be modeled as mixture of distributions. Because of the positive valued nature of the data, DNS requests gap are always positive, we will use a mixture of Gamma

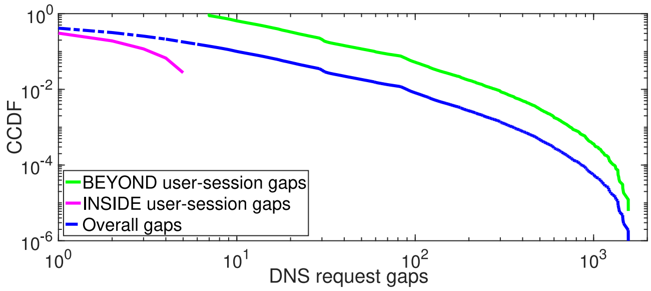


**Fig. 1.** Distribution of DNS request gaps.

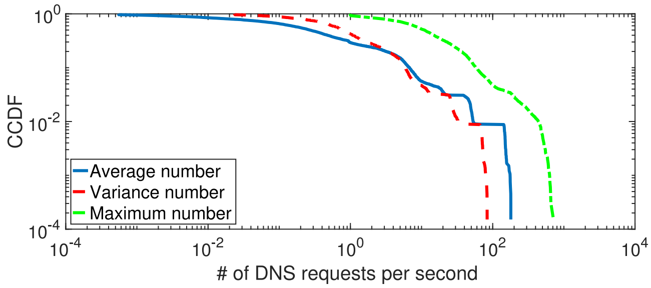


**Fig. 2.** CCDF of observed DNS requests per second.

**Table 2**
Cross-correlation values between average (Ave), variance (Var) and maximum (Max) values of number of DNS requests per second.

| | Ave | Var | Max |
|---|---|---|---|
| Ave | 1.0000 | 0.9857 | 0.8777 |
| Var | 0.9857 | 1.0000 | 0.9282 |
| Max | 0.8777 | 0.9282 | 1.0000 |

components to model the probability distribution of DNS request gaps. We have calibrated the parameter of the gamma distributions using an EM algorithm [29]. By using an Akaike information criteria [30] we have assessed that 3 classes are enough to model the empirical distribution. Fig. 1 shows the CCDF of the DNS request gaps measured over all DNS request sequences. We got three classes: one class with an average of 0.05 s, one with an average of 2.9 s, and the last with an average of 45 s. We thereafter assign each DNS request gap to one of the three classes using a Maximum Posterior probability criterion [29]. Gaps assigned to the first class are all 0 (not shown in the Figure), the second class (shown as INSIDE user-session gaps in Fig. 1) contains all DNS request gaps less than 5 s and the last class (shown as BEYOND user-session gaps) contains the remaining DNS request gaps.

Based on the above observation, we have chosen the threshold to be equal to 5 s, *i.e.*, all DNS requests that are distant by less than 5 s are assumed to belong to the same user-session, and when two DNS request are more than 5 s apart we finish a user-session. This simple mechanism is applied to split the temporal sequence of DNS requests relative to each source IP address into user-sessions. Each user-session contains therefore a sequence of DNS requests that are closely related to each other. However, in this paper we are interested in advertisement services and trackers. Generally, advertisement and tracker appear in user sessions using this pattern. A user access a web page or an online service that have some trackers or advertisers embedded in it. This results into a sequence of access to trackers services following closely the access to the initial page or service. In other term, we are interested into user-sessions that begin with a DNS request to a domain not identified as a tracker or an advertisement domain, and that contain later in the session access to at least one domain identified as tracker or advertiser. The identification of a domain as tracker or advertiser is done using a matching with the blacklist described earlier. As we are mainly interested into trackers/advertisers we will only keep into user-sessions the first non-tracker domain and the domain detected as tracker or advertiser, resulting into a user-session relative to a user $k$ and beginning at time $t$ being a set $S_i^k(t)$ containing the canonical name of the first non-tracker server followed by a sequence of tracker domain names accessed in the same user-sessions. The sets $S_i^k(t)$ are the main raw data we will use in the forthcoming. It is noteworthy, that as we use DNS records we do not know precisely to which URL the traffic is directed. This means that we see a DNS record with apple.com, it can go to anyone of the hosted websites on apple.com.

### 2.3. NAT issue

We explained before that we are assuming that there is only a single user-session at each time behind an IP addresses. This assumption is important as without it one would mix packet coming from different user-sessions into a single one. We therefore need to avoid using IP addresses that are shared between several users through the use of Network Address Translation (NAT) middleboxes [31]. In this section, we will present the methodology to identify the unique users for further analysis.

Averaging over all the DNS dataset, we observed for each source IP address an unexpected high value of 4.01 DNS requests per second on average. In order to investigate more thoroughly this value we derived for each source IP address three values: the average, the variance, and the maximum number of DNS requests per second. We show in Fig. 2 the Complementary Cumulative Distribution Function (CCDF) of these three values.

As we can see from Fig. 2, the average number of DNS request per second spans almost 6 orders of magnitude from $5 \times 10^{-4}$ to 181, exhibiting a very heavy tail. This shows clearly that the statistics of some IP addresses are largely contributing to the average. We can explain this by the fact that some IP addresses, in particular mobile IP addresses, are NATed and shared between several real users. We therefore need to detect these NATed addresses in order to not use them for user-sessions extraction and the derivation of user based statistics, *e.g.* the audience of online resource or co-occurrence. Nevertheless, some other aggregated statistics can benefit from these NATed IPs even if there are several users behind them.

We show in Table 2 the cross-correlation between the above defined three values. The average and the variance values are strongly correlated, while the correlation with maximum value is milder. Based on this observation, we use the average number along with the maximum number of DNS requests per second in order to classify IP addresses into NATed and non-NATed ones.

NAT detection is a well studied area and several active and passive methods have been proposed to detect address translation boxes [32, 33]. However, most of these techniques leverage on header or payload contents. In our case, we attempt to detect NATed IP addresses using only DNS queries. Moreover, our goal is not *per se* to detect NAT boxes, but to detect cases where several users are sharing simultaneously a single IP address simultaneously. In order to achieve this, we leverage on the observation that NATed IP address generates a higher rate of DNS queries than non-NATed ones as they have several users behind them. It is noteworthy that a single web session might contain a large number of URLs resulting into a large influx of DNS requests, and a large maximum number of requests. This means that in practice, the DNS query flow comes from a mixture of NATed and non-NATed nodes and we have to classify incoming requests into these two classes.

Our dataset contains mobile and ADSL users. The practice of mobile operator is to use NAT. This means that mobile users using IPv4 addresses are very likely behind NAT. The ISP who provided the DNS dataset also provided us with IP address ranges used for ADSL and mobile networks. While mobile users are very likely behind a NAT, ADSL users might also decide to share their ADSL connectivity and become

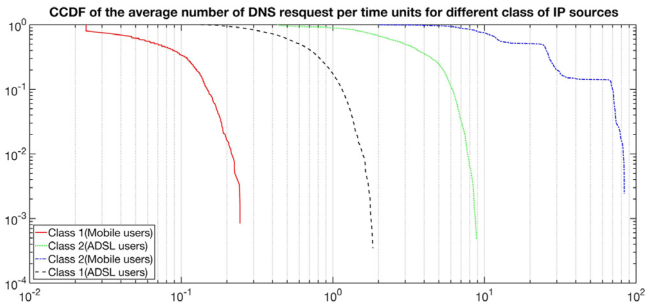


**Fig. 3.** The CCDF of the average number of DNS requests per second for IP flows assigned to each class.

**Table 3**
Mixture Classification results: Average and Maximum number of DNS requests per second for each class obtained from the mixture of correlated GG distribution.

| Category | Class | Average | Maximum |
|---|---|---|---|
| Mobile users | 1 | 0.012 | 2.45 |
| Mobile users | 2 | 5.03 | 29.21 |
| ADSL users | 1 | 0.046 | 3.83 |
| ADSL users | 2 | 0.78 | 18.60 |

NATed. Because of the difference between these two category of users we decided to analyze each one separately and compare the outcomes.

In order to separate NATed and non-NATed IP addresses we use a mixture based classification. We fit the joint distribution of the average and maximum values of DNS requests rate to a mixture of correlated General Gamma (GG) Distributions [34] using an EM algorithm. This GG distribution kernel is used in place of the classical Gaussian component because of the positivity constraint and its heavier tails. We further use a maximum Akaike information criterion in order to select the number of mixture components. This criterion gives that 2 components are enough to classify the observations. After classification we got two classes in each category, the total four classes are shown in Fig. 3.

We present in Table 3 the result of the classification for the two category of IP addresses. The table shows for each class, the mean average and mean maximum values of the number of DNS requests per second. In both IP address categories, there is a strong differentiation between the two resulting classes: one class generates on average a very low number of DNS requests per second (0.012 for mobile users and 0.046 for ADSL users) and the second class generates a much larger number of DNS requests per second. The values in the first class are compatible with a single user usage while the values in the second class can be assimilated to NATed IPs.

IP addresses can be assigned to classes using a maximum likelihood criterion. Based on the above observation, we decided to assume that IP addresses detected in class 2 are NATed and to filter them out, removing 29.2% of them. We use only IP addresses in class 1 for the extraction of user-sessions described in Section 2.2.

### 2.4. Alleviating cache effect

A network client accessing an online resource with a canonical name has to translate this name to an IP address. When a canonical name has been resolved it is stored into the local DNS cache with a time-to-live (TTL) property, specifying the maximum amount of time whether this record should stay in cache, the TTL continuously decreases and when it hits 0 the record is removed, ensuring the cache freshness. Generally the maximal value of the TTL is set in the DNS authoritative server, depending on the strategy of the content/service provider. Before sending the DNS request to the DNS server of its operator, the client first checks if this information is available in its local DNS cache. Therefore, if a DNS record is available in the local cache it will not be seen in the DNS trace. The issue of DNS cache is depicted clearly in Fig. 4. A requested DNS record not in the cache initiates a *miss* that results in querying the DNS record from higher level of the DNS cache hierarchy. When the record is retrieved, it is cached during a time defined by the TTL attached to the DNS request. A request arriving during the caching gets a *hit*. In the DNS server at the ISP level, we will only observe the first request and not the subsequent requests inside the caching duration. This means that local cache filters out a relatively large proportion of DNS requests that never reach the ISP caches that we are monitoring, *i.e.*, any observation using DNS traces is only a partial sampling of the real Internet activity. The variable value of TTL used by different content/service providers adds a level of complexity to the analysis of the effect of DNS cache. In [35] a detailed analysis of the cache effect is made and a general formula is derived that enable to precisely calculate the effect of a DNS cache as a function of the incoming DNS request inter-arrival distribution and the cache duration. This formula is inapplicable to our case as we only observe the outcoming DNS flow after the DNS cache filtering and we miss both incoming inter-arrival distribution and caching time duration distribution. In fact, [35] proved that it is impossible to retrieve precisely the statistics of incoming DNS requests from outcoming DNS flow. We propose here a rescaling methodology as a heuristic to solve the above described cache issue. The rescaling method leverages on the user-sessions, we derived in Section 2.2.

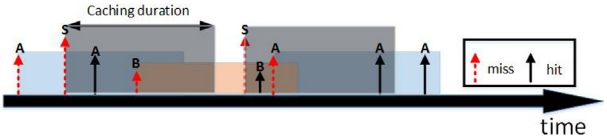


**Fig. 4.** DNS cache issue.

#### 2.4.1. Rescaling method

Let us assume that we have split the DNS requests into sets $S_i^k(t)$ as described in Section 2.2. The sets $S_i^k(t)$ contains an initial domain `site1` following by a list of trackers of advertisers domains. If there were no DNS cache, all trackers accessed by a user during the $i$th user-session of user $k$ would appear in $S_i^k(t)$. However, DNS caches make the set $S_i^k(t)$ incomplete as cached request will not appear in it. This is exacerbated for frequently asked tracker domains as they are more likely to be stored in DNS caches. However, different users, at different timestamps, and different locations, accessing the same destination service/content provider, will not observe the same DNS cache state, *i.e.*, different user-sessions going to the same content/service might contain different but still incomplete list of contacted trackers, giving different samples of the real user sessions sampled at different position in space. We can leverage these spatial samples and merge the different sets $S_i^k(t)$ relative to the same initial domain `site1`. This will results into an inflated set of trackers that will complete trackers that have been missed because of the DNS caching. We can rescale the DNS observations, by replacing the merged set in place of anyone of the sets $S_i^k(t)$ that are beginning with the domain `site1`.

However, this method has some issues. First, the intuition of larger delay between two user session is not valid anymore, when several users are sharing the same IP address. This is the main reason why we need to ensure as much as possible that there is a single user behind an IP address and we enforce this through the approach described in Section 2.3. Even, if we ensure that there is a single user using an IP address, we can still have issues, *e.g.*, when a browser opens two web-pages simultaneously (because of reloading saved states after a reboot for example), the DNS requests relative to both web-pages become intertwined and they will be considered to belong to same user-session even if they are in fact two separate user-sessions. When such a mix happens between different user-sessions, trackers observed in other sessions are erroneously assigned to the first session. With the merging, this error propagates to all sessions sharing the same content/service provider. This can strongly impact the validity of the rescaling approach.

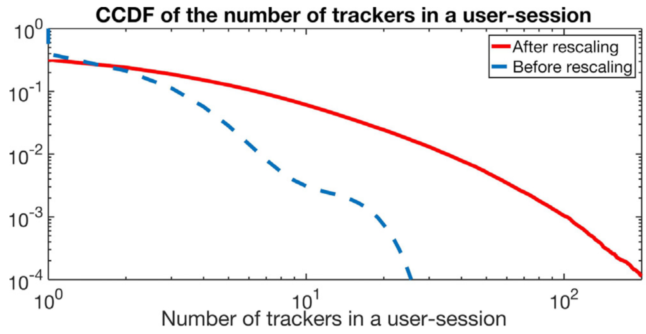


**Fig. 5.** CCDF of the number of trackers observed in a user session.

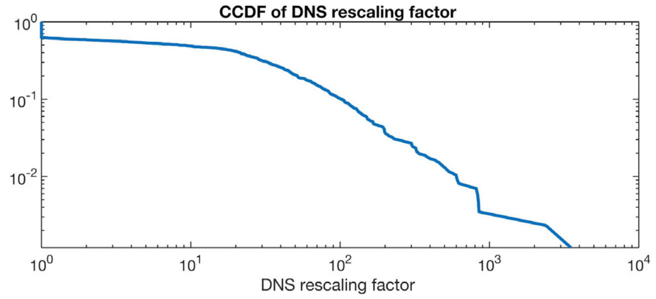


**Fig. 6.** CCDF of DNS rescaling factor applied to trackers.

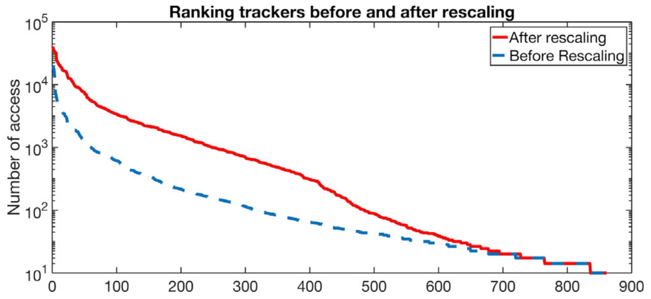


**Fig. 7.** Ranking plot of trackers before and after rescaling.

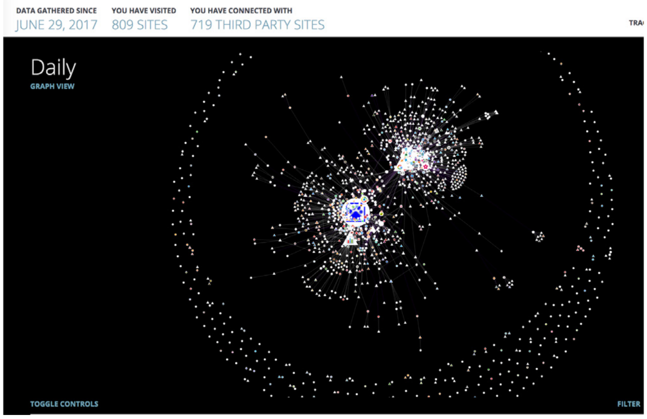


**Fig. 8.** Graph of the 800 sites and their trackers (Visualized by Lightbeam).

One approach to make robust the rescaling toward above described issues is to only merge a tracker to the list of trackers attached to a domain, *e.g.* `site1`, if this tracker have been observed at least in 2 sessions associated to domain `site1`. This precautionary step decreases the likelihood that a spurious tracker is added as it is very unlikely that the same set of sites are accessed at the same time several times in the dataset.

Second issues that we will have anyway is that will still be some domain DNS requests that will never be retrieved, *e.g.*, when the content/service provider canonical name is stored in the cache. In such case we might see some requests going to tracker servers, but unrelated to a content provider. This means that the absolute values of tracker related value we infer might be unreliable while the relative values are more robust as cache effect is spread over all incoming requests.

### 2.5. *Rescaling results*

We present here the results of applying the user-session splitting and the rescaling method to the DNS dataset. Applying the user-session splitting method describe before on the full dataset, we ended up with up to 160 millions user sessions containing at least one tracker with an average of 1.86 DNS requests per user-sessions. After rescaling this number increases to 2.73. We show in Fig. 5 the distribution of the number of tracker observed in a user session before and after the DNS rescaling. From Fig. 5 it can be seen the DNS rescaling fatten strongly the tail of the CCDF distribution. Before rescaling we had at maximum up to 40 trackers in a user session while after rescaling this number increases to 200. While the proportion of user sessions that witness such extreme situation is small, there very existence is interesting. We investigated the source of these case, and we observed that they are all linked to web pages or Internet services that aggregates several others, so that a contact to such web pages results in a cascade of DNS accesses. News aggregators or Internet portals are example of such situations. In these cases trackers in each contacted server add up with each other resulting into a large number of trackers in the session. As explained before, through DNS records we do not know precisely to which URL the traffic is directed. This means the number of trackers after and before rescaling will aggregate trackers seen of all web pages sharing the same domain name. This explain the relatively high number of trackers seen.

Fig. 6 shows the CCDF of the rescaling factor applied to different trackers count. It can be seen that the rescaling is not uniform. Around 40% of trackers are not rescaled at all, and 1% of them have a rescaling over 600. Overall, the rescaling increased the number of estimated access to tracker servers by a coefficient of 9.55, regions with higher population being rescaled stronger. we show in Fig. 7 another evaluation of the rescaling related to the ranking diagram of trackers as a function of the number of access observed before and after the rescaling. We can see that the rescaling pulled the central part of the ranking curve. Interestingly, when we estimate the zipf law exponent fitting the previous curves with $k^{-\alpha}$ using the method described in [36], we end up with an estimate of $\hat{\alpha} = 0.1329 \pm 0.036$ before rescaling with $R^2 = 0.78$ and $\hat{\alpha} = 0.1347 \pm 0.028$ after rescaling with $R^2 = 0.69$. The two value being very close means that the rescaling have not fundamentally changed the overall ranking structure.

### 2.6. *Validating the rescaling method*

In order to validate the rescaling method described earlier, we have compared our obtained set of trackers with the set predicted by the LightBeam tool [37], an add-on for browsers that displays connections to tracking and advertisement third-parties and their cookies that are placed on the user's computer.

We visited a sample of 800 sites extracted from the DNS dataset, as shown in Fig. 8, 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 two sets to the size of LightBeam obtained set) equal to 91.7%, showing that the rescaling method do a very good job on finding all trackers. We conjecture that the difference is mainly due to the two years of delay between gathering the DNS trace

**Table 4**
Comparison of two tracker lists related to "sohu.com" got from our methodology and from Lightbeam.

| | Trackers |
|---|---|
| In both | adnxs.com, rubiconproject.com, revsci.net, mathtag.com, doubleclick.net, baidustatic.com, scorecardresearch.com, tanx.com, miaozhen.com, optaim.com, googlesyndication.com, wrating.com, 2mdn.net, alicdn.com, focus.cn, rlcdn.com, sohucs.com, gentags.net, vamaker.com |
| DNS dataset | tapjoy.com, allyes.com, supercell.net |
| Lightbeam | mct01.com, mmtro.com, eulerian.net, adventori.com, irs01.com |

and the validation with LightBeam. During this time period trackers of web pages or services might have been changed.

We take site sohu.com as an example and show its details in Table 4. Comparing the two tracker lists obtained from our approach and from Lightbeam, 19 trackers are common, 3 trackers are only detect by our approach and they never occurred in Lightbeam. Five trackers are only detect by Lightbeam, among them only irs01.com is not correctly detected by our approach as the other four trackers are never accessed in our DNS dataset. These four trackers might be new trackers that have became operational after we collected our data.

### 2.7. Ethics and privacy issues

The DNS dataset we have been using is unique, and *per se*, it raises some ethical and privacy related questions. First, it is noteworthy that such datasets are routinely gathered by DNS servers in form of logs, and they are mainly used for security and operational purposes. The fact that such datasets are generally not shared with the research community does not mean that they do not exist. In no mean, we have asked to gather a specific DNS trace with some private information in it for the purpose of our study. The used dataset has been gathered under Chinese legal requirements, in particular this dataset was not directly accessed outside China. Indeed, having the real IP addresses of the user in the dataset was problematic. For this reason, we decided to use an anonymization technique similar to the one presented in [38] before accessing the data. We are aware that this technique is far from perfect but we never generated in this study any report with a granularity level that would enable the access to an individual activities. In particular, all statistics gathered in this study does not go below a province granularity. For this reason, we believe that, even if the dataset could be misused in the absolute, that we followed an ethical approach to the dataset and we did not misuse it. We had not gathered the information, we were not storing them, we were into the legal framework applicable in the country where the gathering and the processing happened, and no personal data has been accessed. This was the reasons, we felt that we do not need to go through an ethic board that was by the way not existing in some of the institutions the authors belong.

## 3. Tracking the trackers

After having rescaled the DNS trace and having generated the merged (rescaled) user-sessions we have now a reliable dataset that can be used to make an in-depth analysis of the trackers and advertisers environment and eco-system. 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 analyze their tracking behaviors.

### 3.1. Global perspective on tracking

In the present investigation we measure a site or tracker's traffic as the number of its occurrences in the rescaled dataset. It is noteworthy that this is not the traffic in term of bytes per second, as through the DNS traffic alone we have no idea about the volume of data transferred. However, one can clearly expect that this notion of traffic are clearly correlated.

Fig. 9 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. As can be seen the traffic volume drops very fast with decreasing rank, *e.g.*, the top element has thousand times more traffic than the 100th. We have therefore a Pareto like principle, similar to what observed already on Alexa ranking for instance, where a handful of sites represent the largest share of the global traffic.

We also consider for the top 100 sites, the number of distinct trackers observed per site. Fig. 10(a) shows that the sites which have more traffic seem to attract more trackers.

We plot in Fig. 10(b) the CDF of the number of trackers observed in the rescaled dataset per site. We can see that the number of trackers range from 0 to 200. While more than half of the sites have no trackers, 10% are witnessing more than 100 trackers. However, considering Fig. 9, the top 100 sites still concentrate most of the tracking activity.

### 3.2. Main actors

It follows immediately from the analysis above, that 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.

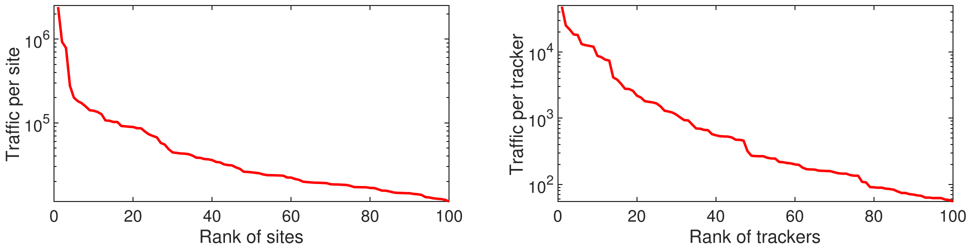
Fig. 11 shows the high level picture of connection among top actors. Fig. 11 presents a bipartite graph between the 28 sites and the 26 trackers that is generated using d3.js [39]. A dynamic visualization is available online at: http://bl.ocks.org/WebTrackingCartography/raw/e59cfc5870d6ec8990a30e05fac72f74/. It is possible on this visualization, to access the details for each site or tracker by simply clicking on it. For each site, say qq.com, the tracking traffic equals the sum of the number of times where each of the 26 trackers have occurred in a session of qq.com. This number can be several times larger than the number of occurrences of qq.com itself in the data, since there are many trackers in each session. The share of tracking traffic of qq.com among all 28 sites, *i.e.*, the percentage of tracking traffic generated from this site, 34.4%, is also shown in Fig. 11.

For each tracker, such as doubleclick.net, the tracking traffic equals the sum of the number of times where each of the 28 top sites have occurred in a session which contains doubleclick.net. The share of this tracker among all 26 trackers, *i.e.*, the percentage of sites that have this tracker, 3.7%, is shown in Fig. 11. To accommodate the image with the fact that the largest traffic can be hundred times larger than the smallest, the size of each bar, associated with sites and trackers, is logarithmically proportional to their corresponding traffic.

As can be seen from Fig. 11 the graph is almost a complete bipartite graph where each one of the 28 sites (on the left) has almost all top 26 trackers (on the right) tracking on it, but with different levels of tracking traffic. *Vice versa*, each tracker will track almost all top 28 sites. This observation confirms previous results showing that the tracking and advertisement ecosystem is heavily dominated by only a small group of actors, which are highly connected [12,40]. Nonetheless, this graph shows also the relative strength of different tracking and advertisement actors. We will elaborate more on this later.

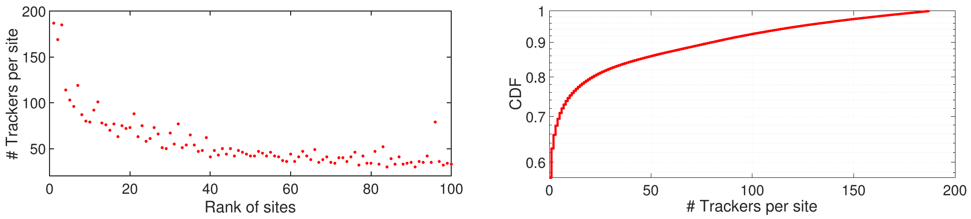
## 4. Information collected by trackers

In Section 3, we identified the top trackers. Although, knowing which tracker are the prominent one is of interest, however the DNS trace does not give us a view about which kind of data are collected and gathered by these trackers. In this section, we will focus on the information collected by trackers from users.



(a) Traffic per site.
(b) Traffic per tracker.

**Fig. 9.** Traffic per site/tracker (Top 100 are present).



(a) Top 100 sites.
(b) All sites.

**Fig. 10.** Number of trackers per site.

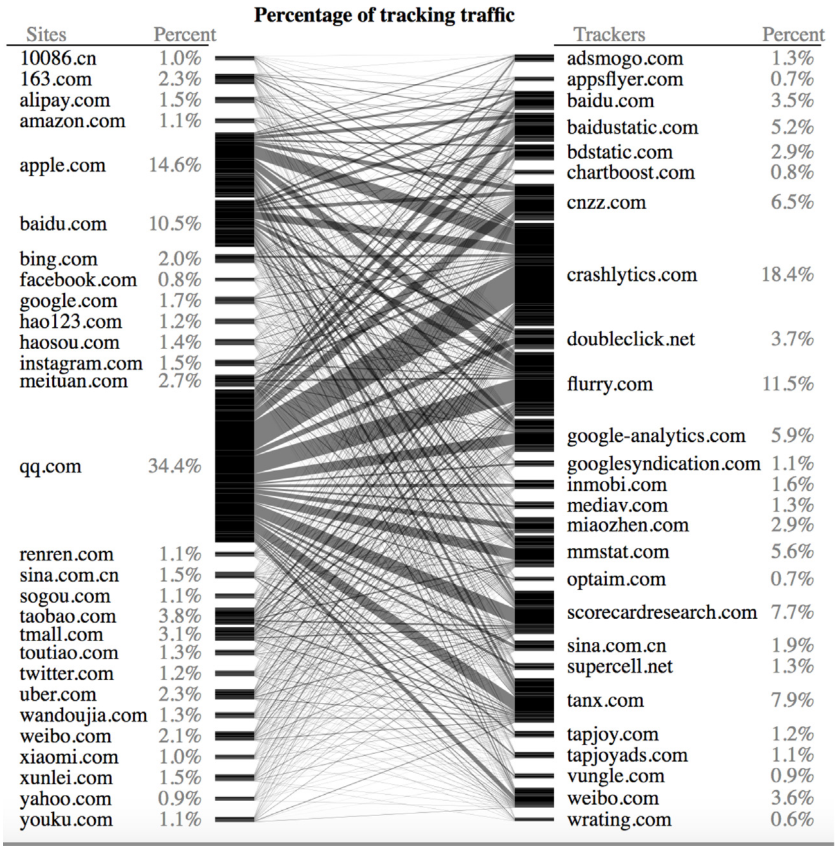


**Fig. 11.** Bipartite graph of tracking traffic between the 28 top sites and the 26 top trackers.

Trackers use various mechanisms, *e.g.*, third-party libraries or JavaScripts APIs, to harvest information both from mobile Applications and web sites. Several studies [41–43] targeted the investigation and classification of third-party ads/tracking libraries found in mobile apps and JavaScript code APIs embedded in web sites, in term of their activities and the garnered attributes. We consider the trackers in our dataset together with the numerous garnered attributes — often protected by permissions [44,45] to target users or profile users' activities.

Table 5 shows the list of the top 26 trackers. We found that respectively 16 (62%), resp. 15 (58%) of the top 26 trackers track users only on

**Table 5**
List of mobile and web trackers with their category (Cat.) ordered by decreasing traffic in our data. Widget means social network widgets.

| # | 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 6**
List of tracker attributes, with their frequency and description.

| # | 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%) | locale |
| 9 | browser_info | 13 (50%) | browser (agent) info |
| 10 | ad_view | 13 (50%) | ads veiw 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 |

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.

We then consider the attributes collected by the top 26 trackers. Following an approach pursued in [41,43], 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 combination with other data, we obtain merely a lower bound on the garnered attributes per tracker.

Fig. 12(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 6). We observe that 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 data 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. 12(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.

## 5. Geography of tracking

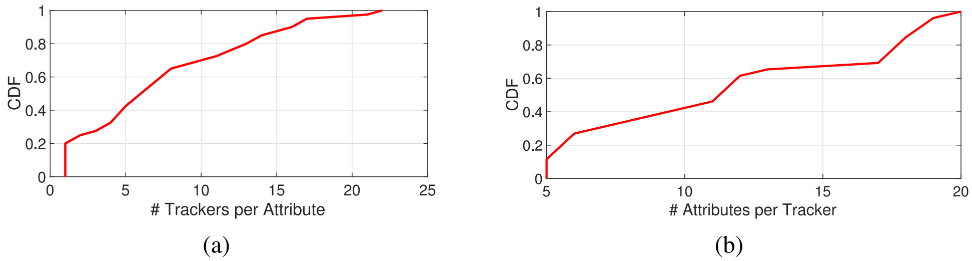
In this section, we consider the trackers from the point of view of geography. Our goal is to better understand the advertisement market, and at to gather insight about the global distribution of data flows.

Determining the country of origin of a tracker company is not a straightforward task. It is plausible to see 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. In this study we have assigned a tracker/ads service to the country that is registered in the WHOis database along with its corresponding canonical domain name. The WHOis database contains contact information for administrative and technical contact points along with the country. While the WHOis database is known to not be fully reliable and up to date in general, the DNS service being a critical service for trackers, we have assumed that information relative to trackers/ads are globally up to date and reliable.

### 5.1. Inequality between countries

The traffic derived from the DNS traces can thus be attached to destination countries. More precisely, we measure the traffic load of a country 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 corporate based in this country. It is important to remind here, that the traffic we talk about is not in term of bytes per second but rather in term of number of connection as assessed through DNS traces. As can be seen in Fig. 13, 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. 14 the worldwide distribution of the traffic to tracker and ads services. 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



(a)

(b)

**Fig. 12.** CDFs of number of trackers per attributes, 12(a), and attributes per trackers, 12(b).

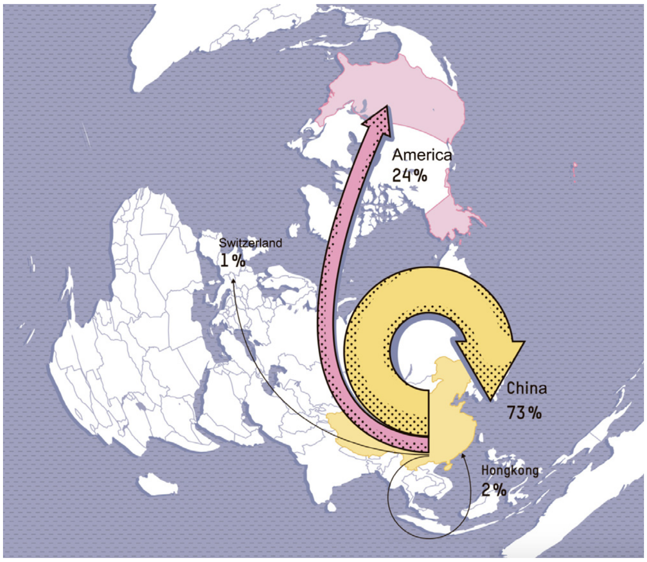


**Fig. 13.** Traffic share between countries from China.

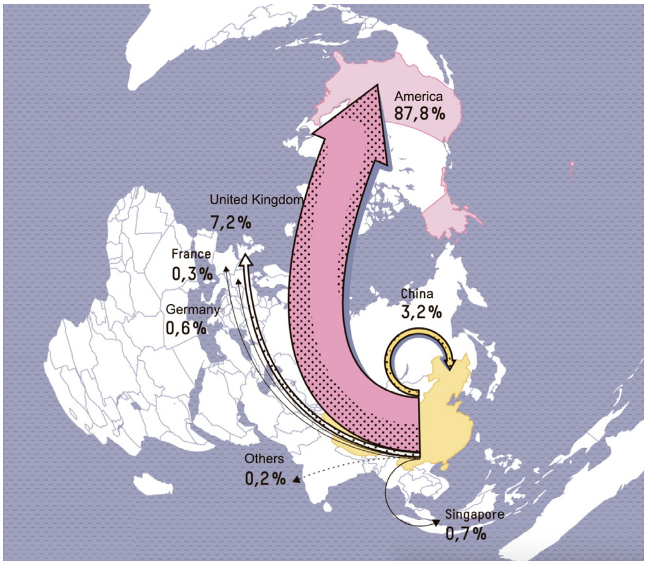


**Fig. 14.** Traffic share of tracking services from China.

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. It should be noted that this surprising situation holds despite the fact that China has a rich advertisement and e-commerce ecosystem and despite the protectionist stand of China toward Internet that is privileging local actors to international ones.

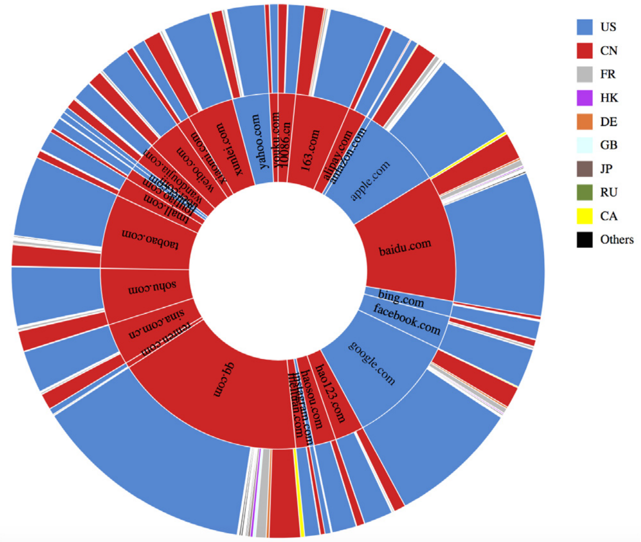
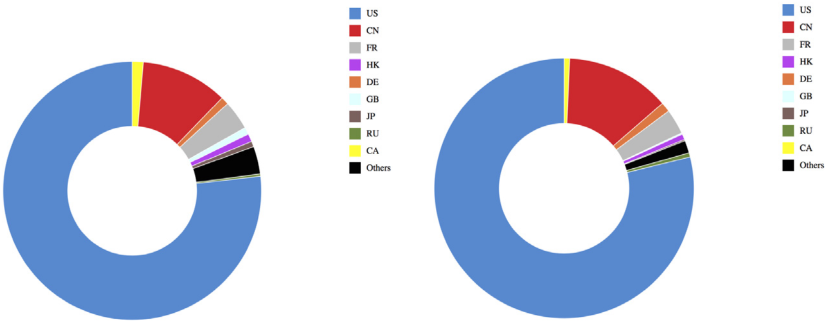


**Fig. 15.** Geography of the top 28 sites and their trackers . (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

### 5.2. Analysis of main sites and their trackers

We consider  in Fig. 15 the top 28 sites identified in Section 3, displayed as a bi-level rings partition, obtained using d3.js [39]. 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. A dynamic of version of this figure is available online at: http://bl.ocks.org/WebTrackingCartography/raw/f2bca61a0780f47dca5f618700d76065/ and allows to navigate dynamically in the image to obtain more detailed information on each actor, by a simple click. The similarity between the distribution of trackers on Chinese an US sites are striking. For example, as shown in Fig. 16, while qq.com and google.com carry on different activities, the have similar tracking patterns; qq.com being a large social platform in China with messaging application as its main products, and google.com hosts a global search engine.

To go deeper into this analysis, we show in Fig. 17 the cosine similarity [46] of the trackers share between pairs of sites, the first 19 sites being Chinese, while the last 9 ones being US sites. All these sites show very similar patterns of tracking, which cosine similarity ranging from 0.87 to nearly 1. Some sites though, such as twitter.com , instagram.com, uber.com, renren.com and wandoujia.com, exhibit somehow a different pattern. When we inspect their trackers, they present less diverse trackers compared to other sites and can rely exclusively on US trackers for example.



(a) qq.com

(b) google.com

**Fig. 16.** Country wise distribution of trackers related to qq.com and google.com.

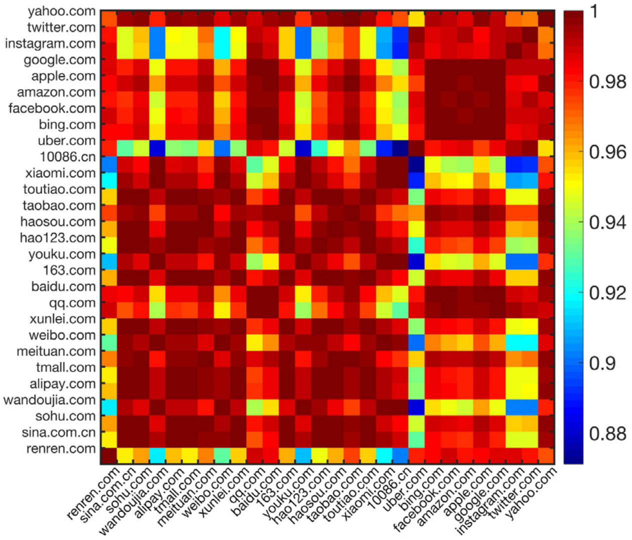


**Fig. 17.** Cosine similarity of the geographic distribution of top sites.

### 5.3. Attributes collected in each region

As we have categorized the type of information gathered by the trackers, we can carry on an analysis considering the number of collected attributes by trackers in relation with the different geographic regions they belong to.

Fig. 18 shows the number of times each attribute has been collected, grouped by the trackers affiliated countries (since the top 26 trackers are coming from China and US, for simplicity, we only distinguish between three regions: CN, US and Others). The rank of each attribute corresponds to the one derived from Table 6. Note that a similar ranking will be used in the sequel for Figs. 19 and 20.

We observed that China based trackers are extracting more specific attributes, `u_time`, `ip`, `lang_id` `scookie_info`, `action_info`, `pcookie_info` for instance. While trackers from US and other countries, collect mostly attributes that among the most frequently garnered, as can be seen in Table 6. Chinese trackers are also extracting more frequently `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`.

In Fig. 19 we presents a more refined analysis by distinguishing between the mobile and Web platforms. We observe that trackers from other countries, beside China and US are mainly mobile trackers. They collect mainly high ranked attributes in Table 6. We observe also that US trackers are more active on mobile platforms. Moreover, on mobile they collect a larger set of attributes than their Chinese counterparts. The situation is more balanced for web platforms. We also show in Fig. 20 the results split as six different categories of activity as defined in Table 5. The figures show that a majority of the attributes are used by trackers for ads and analytics activity. Chinese trackers are more present for ads, while US trackers are more interested on analytics information. Trackers used for targeted ads and utility are essentially US based, as shown in Figs. 20(c) and 20(d). They collect `ip`, `os_info`, `dev_info`, `imei` and `loc` from the users.

Trackers used for widgets and search engine are shown in Figs. 20(c) and 20(d). Widgets trackers target attributes such as in `u_time`, `loc`, `apps_list`, `friendlist` and `user_id`, to help share content on social platforms. While search engines related trackers collect attributes such as `u_time`, `ip`, `cookie_info`, `lang_id`, `browser_info` and `brow_hist`, to track users' browsing behavior. It is no surprising that all these attributes are collected by Chinese trackers, since US social platforms and search engine have lower penetration in the Chinese market.

## 6. Related works

Web tracking ecosystem has attracted a rich literature, focusing on behavioral and privacy aspects.

In [5], Krishnamurthy et al. present the results of a longitudinal measurement of web tracking and prevalence of trackers. They had previously analyzed the growing association between first-party and third-parties in [47], then in [48], the access of third-parties to personal information was analyzed and leakage were found for every categories of first-party websites.

Roesner et al. [7] made a classification between different types of web trackers and measured the prevalence of these classes among the world's top 500 websites. In [16], Castelluccia et al. using two popular browser extensions to analyze the geographical provenance of major third party tracking services. They focused on measuring the penetration of US-based trackers in different countries. Gomer et al. [12] focused on the networking aspects of third-party trackers in three search markets. They show a consistent network structure across different markets as well as a high level of efficiency in information exchanges between third-parties. Mayer et al. [6] surveyed different techniques which are used by web trackers to collect user information. In paper [13], Falahrastegar et al. crawl the top Alexa ranked websites in 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 the worldwide tracking eco-system. Moreover, they all used partial samples websites or services from Alexa that is itself using a relatively small number of volunteer users installing the measurement plugin. Our study while being focused on the Chinese tracking and ads ecosystem is using an exhaustive DNS trace over all regions of China. While we have confirmed previous observations made in the literature, our work present new insights that were not reported before.

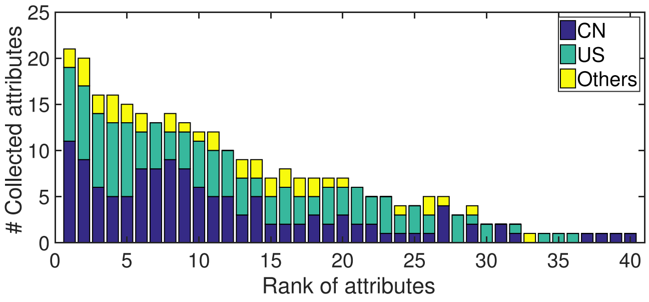
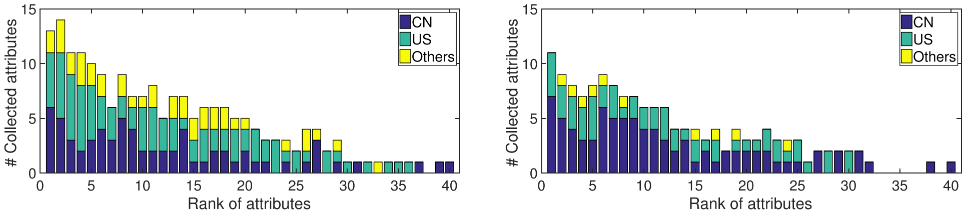


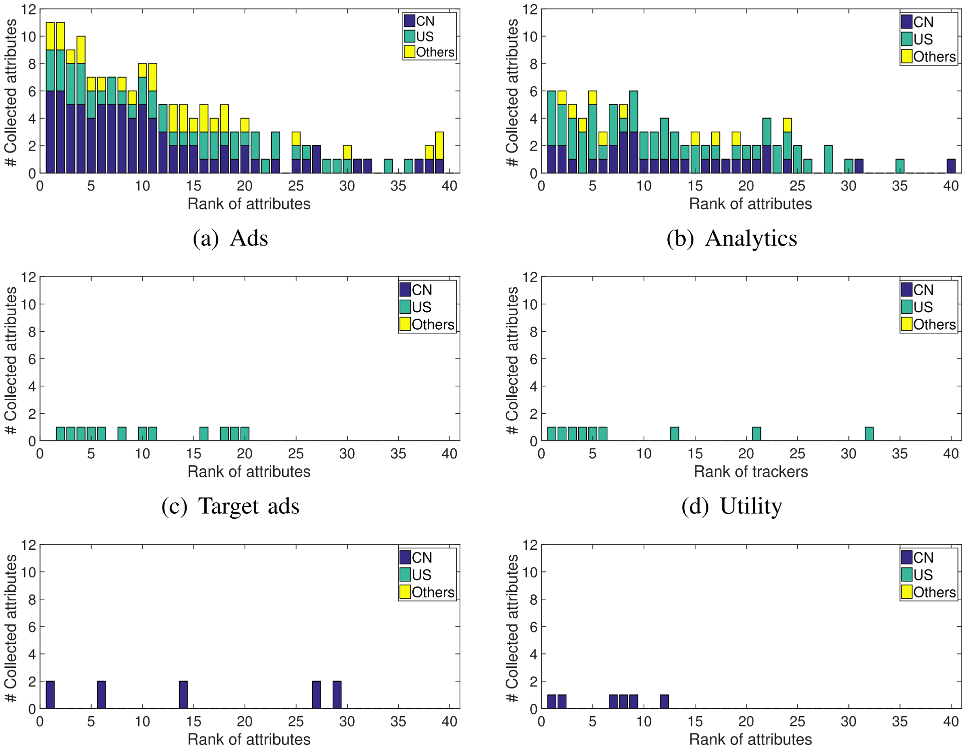
**Fig. 18.** Proportion of collected attributes in each region.



(a) Mobile

(b) Web

**Fig. 19.** Proportion of collected attributes for mobile or Web.



(a) Ads

(b) Analytics

(c) Target ads

(d) Utility

(e) Widget

(f) Search engine

**Fig. 20.** Proportion of collected attributes by category.

## 7. Discussions and conclusions

This work was made possible thanks to an exceptional DNS trace containing $10^{11}$ records relative to two-days of activity in China. However for exploiting this trace we needed to overcome several technical challenges.

The major issue was to develop new methods to overcome the issues of DNS caches, that filters DNS trace and made them incomplete. To Obtain a more realistic view of the Internet traffic we developed a rescaling method leveraging on multiple user sessions observed at different locations and times and with different local DNS cache state. In order to implement this we had to develop methods for clustering the DNS request into user sessions and to detect IP address that are NATed.

After having addressed the technical challenges, we were able to analyze the tracking and advertisement eco-system in China and beyond. Our initial results confirmed the extreme concentration of online services into a small number of corporations. We focused on the 28 sites, with at least 0.5% of the traffic, representing 67% of the whole traffic, and on the 26 trackers and advertisers representing 90% of the whole trackers/ads traffic. We observed that surprisingly, while Chinese web of services dominates the traffic , with around 3/4 of traffic, 87% of tracking and advertisement activity is directed toward US based actors.

A first question is relative to the reason of such an unbalance. Several causes might be considered. The first reason is that several popular open-source development framework, *e.g.*, in Android environment or WordPress, offer frequently by default US based advertisers/trackers. A second reason might be economic-related. The pay-per-click model used by some major US based advertisement actors is very attractive and the proposed business model of Chinese actors might no be able to compete with it. A more specific cause is relative to the importance of advertisement on other platform, like Wechat, that cannot be analyzed through DNS request. This means that a large part of the advertisement market in China is not happening through traditional means.

Moreover, the observation that the tracking eco-system is dominated by US actors has two very important implications. First, as US actors are not installed inside mainland Chinese network this means that even accessing a web page or service inside China might involve to cross the GFW for accessing the tracker and having long-distance interactions. This may negatively affect the web service performance. Deploying replicas of the trackers within China might alleviate the problem, however this goes against Chinese regulations and policy. Second, our observation indicates that US corporations may have a better view of Chinese users behavior than Chinese one as US and Chinese trackers collect similar information. This raises huge concerns on advertising market, user privacy and cyber security. Enforcement of data protection regulations [13] in China could be an option to address these issues. Making mandatory the local deployment foreign trackers in China is something that Chinese cyber-security law have enforced since June 2017 by requiring mandatory in-country data storage of data collected in China. We will surely investigate the impact of this law to web tracking cartography in future work.

Another overlooked issue is relative to Internet economics that is heavily dependent on advertisement revenue to provide free access to services. The fact that a non-negligible part of this revenue is diverted abroad have an impact on the economic eco-system of Internet in China. This also mean, that the present situation offers interesting opportunities for domestic Chinese online advertising companies to improve their market share by leveraging the advantage of keeping the data inside the country.

Last but not least, the existence of pervasive trackers both Chinese and foreign, entails major privacy concerns for Chinese individuals. As far as we know, tracker blocking tools are not widely used in China, despite the fact that some tools, like Adblock, provide specific lists targeted toward Chinese trackers. Yet, we have limited knowledge of the usage of trackers blocking tools in China. That is left for future work.

## Acknowledgement

This work is supported in part by National Key R&D Program of China (Grant No. 2016YFE0133000): EU-China study on IoT and 5G(EXCITING), National Natural Science Foundation of China (Grant No. 61572475 and No. 61602448).

## References

[1] N. Schmucker, Web tracking, in: SNET2 Seminar Paper-Summer Term, 2011.
[2] Google. Google adsense. https://www.google.com/adsense.
[3] Richard Atterer, Monika Wnuk, Albrecht Schmidt, Knowing the user's every move: User activity tracking for website usability evaluation and implicit interaction, in: Proceedings of the 15th International Conference on World Wide Web, in: WWW '06, ACM, New York, NY, USA, 2006, pp. 203–212.
[4] Google. Google analytic. http://google.com/analytic.
[5] B. Krishnamurthy, C. Wills, Privacy diffusion on the web: A longitudinal perspective, in: Proceedings of the 18th International Conference on World Wide Web, in: WWW '09, 2009, pp. 541–550.
[6] Jonathan R. Mayer, John C. Mitchell, Third-party web tracking: Policy and technology, in: Security and Privacy (SP), 2012 IEEE Symposium on, IEEE, 2012, pp. 413–427.
[7] Franziska Roesner, Tadayoshi Kohno, David Wetherall, Detecting and defending against third-party tracking on the web, in: Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation, in: NSDI'12, USENIX Association, Berkeley, CA, USA, 2012, pp. 12–12.
[8] Chris Jay Hoofnagle, Nathan Good, Web privacy census. 2012.
[9] Gunes Acar, Marc Juarez, Nick Nikiforakis, Claudia Diaz, Seda Gürses, Frank Piessens, Bart Preneel, Fpdetective: dusting the web for fingerprinters, in: Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security, ACM, 2013, pp. 1129–1140.
[10] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, Claudia Diaz, The web never forgets: Persistent tracking mechanisms in the wild, in: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, ACM, 2014, pp. 674–689.
[11] Ibrahim Altaweel, Nathan Good, Chris Jay Hoofnagle, Web privacy census. 2015.
[12] Richard Gomer, Eduarda Mendes Rodrigues, Natasa Milic-Frayling, M.C. Schraefel, Network analysis of third party tracking: User exposure to tracking cookies through search, in: Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01, IEEE Computer Society, 2013, pp. 549–556.
[13] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, 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] Internet statistics 2017. https://hostingfacts.com/internet-facts-stats-2016/.
[16] Claude Castelluccia, Stéphane Grumbach, Lukasz Olejnik, Data harvesting 2.0: from the visible to the invisible web, Collect. Czechoslovak Chem. Commun. 24 (3) (2013) 760–765.
[17] David Pariag, Tim Brecht, Application bandwidth and flow rates from 3 trillion flows across 45 carrier networks, in: International Conference on Passive and Active Network Measurement, Springer, 2017, pp. 129–141.
[18] Amazon company. The top sites on the web. http://www.alexa.com/topsites.
[19] Akira Yamada, Hara Masanori, Yutaka Miyake, Web tracking site detection based on temporal link analysis, in: Advanced Information Networking and Applications Workshops (WAINA), 2010 IEEE 24th International Conference on, IEEE, 2010, pp. 626–631.
[20] Wladimir Palant, Adblock plus: Save your time and traffic. https://easylist.adblockplus.org/en/, 2017.
[21] Ricardo Bilton, Ghostery: A web tracking blocker that actually helps the ad industry. 31:2012, 2012.
[22] Ricardo Bilton, Ghostery. https://www.ghostery.com, 2017.
[23] Disconnect. malvertising list. https://disconnect.me/lists/malvertising, 2016.
[24] Adblock plus easylist china. https://easylist-downloads.adblockplus.org/easylistchina+easylist.txt, 2017.
[25] Aaron Halfaker, Oliver Keyes, Daniel Kluwer, Jacob Thebault-Spieker, Tien Nguyen, Kenneth Shores, Anuradha Uduwage, Morten Warncke-Wang, User session identification based on strong regularities in inter-activity time, in: Proceedings of the 24th International Conference on World Wide Web, in: WWW '15, 2015.
[26] Vern Paxson, Sally Floyd, Wide area traffic: The failure of Poisson modeling, IEEE/ACM Trans. Netw. 3 (3) (1995) 226–244.
[27] Carl Nuzman, Iraj Saniee, Wim Sweldens, Alan Weiss, A compound model for tcp connection arrivals for lan and wan applications, Comput. Netw. 40 (3) (2002) 319–337.
[28] A. Bianco, G. Mardente, M. Mellia, M. Munafo, L. Muscariello, Web user session characterization via clustering techniques.
[29] Jeff Bilmes, A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical report, 1998.
[30] Wikipedia. Akaike information criterion. https://en.wikipedia.org/wiki/Akaike_information_criterion, 2017.
[31] Franois Audet, Cullen Jennings, Network Address Translation (NAT) Behavioral Requirements for Unicast UDP. RFC 4787, RFC Editor, January 2007.
[32] Steven M. Bellovin, A technique for counting natted hosts, in: Proceedings of the 2Nd ACM SIGCOMM Workshop on Internet Measurment, in: IMW '02, ACM, New York, NY, USA, 2002, pp. 267–272.
[33] Gregor Maier, Fabian Schneider, Anja Feldmann, Nat usage in residential broadband networks, in: Proceedings of the 12th International Conference on Passive and Active Measurement, in: PAM'11, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 32–41.
[34] Petros S. Bithas, Nikos C. Sagias, Theodoros A. Tsiftsis, George K. Karagiannidis, Distributions involving correlated generalized gamma variables, in: Proc. Int. Conf. on Applied Stochastic Models and Data Analysis, Vol. 12, 2007.
[35] Nicaise Choungmo Fofack, Sara Alouf, Modeling modern dns caches, in: Proceedings of the 7th International Conference on Performance Evaluation Methodologies and Tools, ValueTools '13, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium, 2013, pp. 184–193.

[36] Aaron Clauset, Cosma Rohilla Shalizi, M.E.J. Newman, Power-law distributions in empirical data, SIAM Rev. 51 (4) (2009) 661–703.

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

[38] Jun Xu, Jinliang Fan, M.H. Ammar, S.B. Moon, Prefix-preserving ip address anonymization: measurement-based security evaluation and a new cryptography-based scheme, in: 10th IEEE International Conference on Network Protocols, 2002. Proceedings, 2002, pp. 280–289.

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

[40] Vito Latora, Massimo Marchiori, Efficient behavior of small-world networks, Phys. Rev. Lett. 87 (19) (2001) 198701.

[41] Muhammad Ikram, Mohamed Ali Kaafar, A first look at ad-blocking apps, in: IEEE NCA, 2017.

[42] Muhammad Ikram, Narseo Vallina-Rodriguez, Suranga Seneviratne, Mohamed Ali Kaafar, Vern Paxson, An analysis of the privacy and security risks of android vpn permission-enabled apps, in: Proceedings of the 2016 Internet Measurement Conference, in: IMC '16, ACM, New York, NY, USA, 2016, pp. 349–364.

[43] Muhammad Ikram, Hassan Jameel Asghar, Mohamed Ali Kaafar, Balachander Krishnamurthy, Anirban Mahanti, Towards seamless tracking-free web: Improved detection of trackers via one-class learning. PETS, 2017.

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

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

[46] Wikipedia. Cosine similarity. https://en.wikipedia.org/wiki/Cosine_similarity, 2017.

[47] Balachander Krishnamurthy, Craig E. Wills, Generating a privacy footprint on the internet, in: Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement, ACM, 2006, pp. 65–70.

[48] Balachander Krishnamurthy, Konstantin Naryshkin, Craig Wills, Privacy leakage vs. protection measures: the growing disconnect, in: Proceedings of the Web, Vol. 2, 2011, pp. 1–10.