Exploring the Eastern Frontier: A First Look at Mobile App Tracking in China

Zhaohua Wang[1,2], Zhenyu Li[1,2,3], Minhui Xue[4], and Gareth Tyson[5]

[1] Institute of Computing Technology, Chinese Academy of Sciences, China
[2] University of Chinese Academy of Sciences, China
[3] Purple Mountain Laboratories, China
[4] The University of Adelaide, Australia
[5] Queen Mary University of London, United Kingdom

Abstract. Many mobile apps are integrated with mobile advertising and tracking services running in the background to collect information for tracking users. Considering China currently tops mobile traffic growth globally, this paper aims to take a first look at China's mobile tracking patterns from a large 4G network. We observe the dominance of the top popular *domestic* trackers and the pervasive tracking on mobile apps. We also discover a very well-connected tracking community, where the non-popular trackers form many local communities with each community tracking a particular category of mobile apps. We further conclude that some trackers have a monopoly on specific groups of mobile users and 10% of users upload Personally Identifiable Information (PII) to trackers (with 90% of PII tracking flows local to China). Our results consistently show a distinctive mobile tracking market in China. We hope the results can inform users and stakeholders on the interplay between mobile tracking and potential security and privacy issues.

1 Introduction

Many mobile apps are bundled with mobile Advertising and Tracking Services (**ATSes**). These are used for various purposes, including monetization, app maintenance, and audience understanding [15, 27, 34]. This, however, can result in such apps exposing a wide variety of information to (third-party) services, often without a clear understanding of how it may be used. Due to the sensitive nature of data accumulated on mobile devices, their prevalence has therefore been a cause for concern [4,6,17,22,29,30]. This is particularly the case as tracking behavior often cannot be controlled by users, particularly after granting apps permissions [11, 40].

Due to the importance of this topic, there has been a large body of recent research in this area, including studies that have used static app analysis [1, 2, 11], dynamic device monitoring [12, 25, 26, 28], and the inspection of network traffic [13,32]. They have revealed a number of insights, including the prominence of a small number of ATS platforms, the presence of privacy invasive leaks (*e.g.* phone numbers), and various attempts at cross-device tracking. Despite this

range of insights, these studies have one common bias: they near exclusively focus on western countries, primarily in North America and Europe. Although these countries are both important and relevant, we posit that this bias introduces a deficiency into the mobile ATS research landscape. Specifically, we have little evidence related to how the above trends may generalize to the Chinese market. As one of the fastest growing countries in terms of mobile traffic [7], we argue that this deficiency must be addressed.

This paper performs the first characterization of mobile ATS traffic patterns in China. Using a dataset containing 28 billion anonymized access logs from mobile users, we explore the distinctive properties of the tracking market in China. Our analysis reveals a highly active ecosystem dominated by a set of (poorly understood) major players. Due to the presence of the Great Firewall of China (which blocks certain western services), a number of trackers are quite distinct from those observed in past works.

Our main findings are summarized as follows:

- We reveal a distinctive mobile tracking market in China that is dominated by several popular *domestic* trackers. A handful of trackers (35%) are present in 2 or more mobile apps, implying the prevalence of cross-tracking of users. Notably, the prominence of tracking in some types of apps (*e.g. InputMethod*) raises particular concerns for user privacy.
- Popular trackers regularly co-occur with non-popular ones. Non-popular trackers, however, tend to cluster into local communities; each community tends to track a particular relevant type of app.
- China's tracking services reach a majority of users, with some trackers showing a tendency to exclusively track specific groups of users. As many as 10% of users send PII data to trackers, implying the possibility of privacy leakage. Nevertheless, 90% of PII data is confined to China.

2 Dataset and Methodology

2.1 Data Description

Our dataset contains user access logs in a major 4G cellular ISP. The access logs are generated by combining the traces of Deep Packet Inspection (DPI) deployed at Serving Gateway (SGW) and the information provided by the Mobility Management Entity (MME). Each log corresponds to an HTTP request, and contains the following major fields: the anonymized unique ID of the user that initiates the request, destination IP Address, request URL, HTTP-Referrer, User-Agent, the data volume, and the timestamp of the request initiation. In addition, to identify the mobile apps which generate each HTTP request, the DPI appliances uses a rule-based approach introduced in SAMPLES [39]. To train the rule-set in SAMPLES, a crawl-download-execution pipeline is run across the major Chinese app markets The rule-set is then deployed on the DPI appliances for app identification, and is updated routinely to include new apps. In total, we identify 1,812 unique mobile apps.

Note that we naturally cannot extract URLs from HTTPS, accounting for around 20% of the mobile traffic observed. However, we note that many apps that use HTTPS *also* use HTTP. For instance, WeChat, the most popular mobile app among Chinese diaspora, relies on HTTPS for third-party APIs, but also issues requests to `imgcache.gtimg.cn` for cached images via HTTP. This means that, even though our vantage is constrained, we can still observe activities. Indeed, the Kendall correlation between the top-100 most popular apps in our dataset and that obtained from [8] is 0.85, suggesting that our app traffic is reflective of general usage. In total, the dataset contains 2,811,233,521 access logs of 3,516,828 users in a major city of China.

2.2 Identifying ATS Domains

Inspired by [18, 27], we utilize four ATS-specific lists provided by: AdBlock-Plus [10] (the *easylist* and *easyprivacy* lists) and hpHosts [23] (the *ATS* list). We further incorporate the EasyList China supplementary list given that we target China's Internet. These contain a set of string matching rules, and are commonly used by ad blockers. We apply the rules to both the URL *and* HTTP-Referrer of each flow, such that we can also identify cases where a URL that is not classified as an ATS was requested by an ATS [16].

In total, we attribute 260M HTTP requests (9.2%) to ATS domains, in which 16.4% are unattributable flows labeled as *others* as mentioned above. These cover 24,985 unique fully-qualified domain names (FQDNs) and 8,773 unique second-level domains (SLDs). Note that our focus is not only on third-party tracking services like [3, 15, 33] where the first-party domains are considered to be trusted by users (even though they can still track users). Instead, we also inspect first-party trackers that collect personal data (contained within EasyPrivacy [24]).

2.3 Associating ATS Domains to Apps

Next, we identify the trackers that are used by individual mobile apps. Casual analysis [20, 38] immediately reveals a highly skewed popularity distribution of mobile apps. The most popular app (WeChat) is accessed by 92% of users in a single day, whereas the majority of services (outside the top 500) are accessed by less than 0.1% of users per day. Hence, to simplify analysis, we focus on the the top-500 mobile apps, which account for 86.7% of HTTP flows in our dataset.[6]

The easiest way to associate trackers with apps is to use the `HTTP-Referrer` and `User-Agent` in the ATS requests [13]. However, for the majority of ATS HTTP requests from unattributed apps, the HTTP-Referrers are empty and the User-Agents do not meet the specification required to identify apps. As such, we turn to an alternative heuristic approach inspired by [31]. The intuition is that if an ATS is associated with a mobile app, its requests should happen at a time close to the app's access. Hence, we can associate an ATS request to the

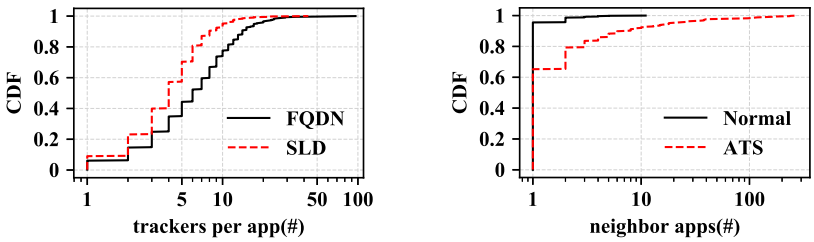
[6] Among the top 500 apps, 29 mobile browsers are excluded in further analysis to avoid potential inflation or bias caused by web trackers bundled in web pages.

closest app's request that precedes it. A problem here is that some apps may send background traffic, which may appear between the app's requests and the requests of the associated trackers. To mitigate this effect, we divide a user's requests into sessions [31], where a session corresponds to a set of user activities before an obvious pause. The session interval is set to 1 minute, which is learned empirically as in [14].

Using the above approach, we obtain 193,527,553 sessions in total, and filter out the sessions that contain requests from more than one app. For the remaining sessions (4,238,015) containing only one identified app request, we can safely associate an ATS domain with the app. For each app, s, this results in a vector R_s, in which an element $< d_i, n_i >$ is an ATS domain and the number of users seeing their association. We further mitigate another possible effect that is relevant to the periodic requests issued by some trackers (*e.g.* statistic tracking services): One potential flaw in the above approach is that certain trackers may very rarely issue requests. Thus, these requests may appear in the sessions that contain only a single app's request (*i.e.* even when the ATS is not associated with the app). Given that this happens only occasionally, for an app s, we filter out those ATS domains T from R_s if $n_i < q$ $(i \in T)$, where q takes the mean of all $n_{j \in R_s}$. Finally, based on the inferred ATS domains of each app above, we process all access logs for each user to associate the ATS request with its host app (assuming the app's request precedes the ATS request less than 1 second). Importantly the filtered sessions include all of the top 500 apps, and are only used for ATS-to-app association. For other analysis (*cf.* Section 3.3), we use *all* access logs.

2.4 Limitations

It is important to highlight potential limitations in our data. The four ATS lists that we utilize for identifying ATS domains may not fully cover the current ATSes in mobile networks in China. But we have identified a number of prominent and recognized mobile tracker domains, which are in line with the Chinese mobile ecosystem. Additionally, the heuristic method for the ATS-to-app association may not fully capture the up-to-date ATSes of individual mobile apps. We utilize both the app Lumen [27] and the Lightbeam tool [21] to manually test existing ATS domains (SLDs) for the top 10 most popular apps. Our inspection revealed an association accuracy of F1-score 0.75 (precision: 0.7, recall: 0.82). Taking the popular video app Youku, for example, among 9 trackers inferred by our approach, 6 dominant ones can also be detected by Lumen or Lightbeam. One domain is not detected by our method but only monitored by Lumen; however, this domain has never been accessed in our dataset and is perhaps an additional tracker after our dataset was collected. Finally, although it has been shown in [39] that the rule-based approach for app identification can achieve a high accuracy, we are not aware of the exact accuracy because the DPI provider keeps its implementation details confidential. Thus, we cannot evaluate its accuracy, nor can we tell how false positives/negatives bias our results. Nevertheless,



(a) CDF of #ATSes per app.

(b) CDF of #neighbors per ATS & other FQDN.

Fig. 1. The presence of ATSes among mobile apps

we find that 12% of the HTTP requests cannot be attributed to particular apps in our dataset and are labeled as *others*.

2.5 Ethical Considerations

The ISP routinely collects user access logs for the purpose of improving their service quality and security. When users subscribe to the ISP network, they are notified that the ISP may collect and analyze their personal and access information for the above purpose (including but not limited to tracking behavior), and may share the information with the research community for research purposes after anonymization. The dataset is kept in the ISP's data center with access being granted only to the authors' affiliation. Several precautions for protecting users' privacy have been taken by the ISP before access is granted. For instance, the unique user IDs are substituted with random numbers to delink the activities with specific users; all sensitive user data (*e.g.* IMEI) has been encrypted by hashing. We have obtained the approval from the ISP for accessing the request URL, HTTP-Referrer and User-Agent fields.

3 Results and Analysis

3.1 How prevalent are ATSes?

Presence of ATSes. Based on user request sessions produced in Section 2.3, we model the domains (FQDNs) accessed within an app as a bipartite graph $G = (U, V, E)$, where U denotes mobile apps, V represents the ATS domains and normal visited domains, and E is the set of edges connecting vertices in U to vertices in V. This 2-mode graph reveals connections between ATS domains and mobile apps. We first analyze the number of ATSes present in each app in graph G and present its CDF distribution in Fig. 1(a). Unsurprisingly, we confirm that ATSes are widely used by mobile apps. The median number of trackers observed per app is 6 for FQDNs, and 4 when classified by SLDs.

We also inspect the number of apps neighbored with each ATS domain in graph G in order to understand how well mobile trackers are connected with

Table 1. Presence of the top 20 ATS domains (SLDs) on mobile apps.

ATS (SLDs)	#FQDNs	%App	ATS (SLDs)	#FQDNs	%App
qq.com	31	75	kuwo.cn	1	6
umeng.com	4	67	flurry.com	1	6
71.am	1	57	baidustatic.com	4	6
baidu.com	45	34	mmstat.com	3	6
uc.cn	3	28	hiido.com	2	4
360.cn	5	25	scorecardresearch.com	2	4
google-analytics.com	1	14	funshion.net	1	4
ksmobile.com	1	13	doubleclick.net	1	4
cnzz.com	33	9	ifeng.com	5	4
xiaomi.com	2	7	letv.com	3	3

different apps. Fig. 1(b) shows that ATS domains tend to appear on much more apps than normal ones: over 30% of trackers appear in at least two apps. To further get a handle on the "popularity" of ATSes among app developers, Table 1 presents the top 20 ATS domains (SLDs), as measured by the number of apps they are used by. The number of FQDNs associated with each SLD is also shown in Table 1. We see a skewed distribution, whereby the top 3 ATS domains are accessed by over half of all apps, while the bottom 12 ATS domains are used by under 10% of apps.

The ATS domains of qq.com are the most popular and accessed by over 70% of all mobile apps observed, showing its pervasive tracking. 31 FQDNs of qq.com are identified as mobile trackers and the top 5 are pingma.qq.com, zxcv.3g.qq.com, omgmta.qq.com, sngmta.qq.com and mi.gdt.qq.com, accounting for 70% of flows of SLDs. They provide services for link share, advertising aggregation and mobile analytics. Notably, unlike Europe which relies heavily on US trackers, China's tracking ecosystem is dominated by key *domestic* trackers: the top 6 most popular SLDs are all domestic (Chinese) ATS domains. Foreign trackers (*e.g.* google-analytics.com, flurry.com, scorecardresearch.com) make up the minority of ATS traffic: they are used by under 20% of apps. Many factors, including Internet censorship, language and unique local regulations, contribute to this unique ecosystem that differs greatly from the western countries.

App's ATS Usage. An obvious question is which apps are responsible for utilizing this wide range of ATSes within their code. To this end, we group the mobile apps into 23 categories collected from several Android app markets using [35]. The categorization is mostly based on the functionality of apps. Table 2 lists the number of apps, user popularity (measured by the share of users) as well as the percentage of ATS domains in each category. [7]

There is a strong propensity towards certain app categories, with *communication* apps (*e.g.* messaging services) being used by 98% of users. The percentage of trackers indeed is dependent on the number of apps of each category and also

[7] As mentioned in Section 2.3, we do not show the number of trackers of the browser apps.

Table 2. App Categories, sorted by the user penetration percentage.

Category	App	User(%)	ATS(%)	Category	App	User(%)	ATS(%)
Commu.	15	98	23	**Input.**	5	37	5
Browsers	29	85	-	**Security**	12	36	8
Navigation	16	75	7	**Photo.**	4	31	2
Tools	45	64	12	**Lifestyle**	38	19	15
Shopping	27	63	13	**Books**	21	18	11
News	27	60	28	**Business**	8	15	8
AppMarket	25	59	11	**Education**	24	11	9
Video	42	57	23	**Person.**	5	6	2
Finance	46	57	12	**Health**	10	4	9
Social	16	53	14	**Travel**	13	4	6
Music	21	41	10	**Other**	14	5	7
Game	37	41	9				

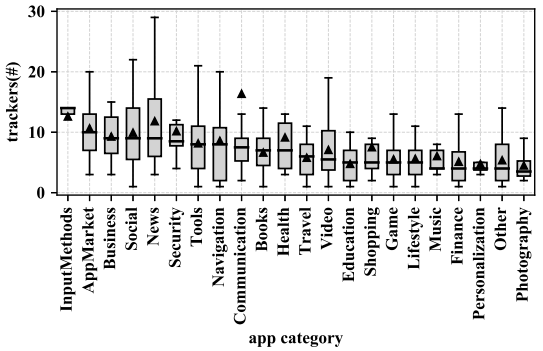


Fig. 2. The distribution of tracker domains (FQDNs) by different app categories.

the apps' functionality. For instance, the communication category, which contains moderate number of apps, has over 23% trackers. This is probably because apps like WeChat are not only communication tools, but platforms for many third-party services (*e.g.* online payments). Trackers serving different purposes will thus likely be embedded in these apps. A closer look at the trackers of video apps shows the dominance of statistic services that collect QoE related metrics.

To mitigate the effect of the number of apps in each category, we count the number of unique trackers of each app and present the box-plot distribution of ATSes (FQDNs) across app categories in Fig. 2. We rank each box in descending order by the median, which ranges from 4 to 13. It is notable that the number of trackers per app varies based on category (*i.e.* its functionality). *InputMethods* apps, which include five third-party keyboards, have the most trackers per app. This is particularly worrying, as they have incentives to log and collect user input to improve their services [5]. *Communication* apps hold the highest mean value of 16 ATSes per app; this is largely driven by certain extremely popular apps (*e.g.* WeChat and QQ). The category with the greatest diversity is *News*: although the median number is 9, the top 5% of news apps use over 26 ATSes. We note that this differs greatly from past western-oriented studies, where games

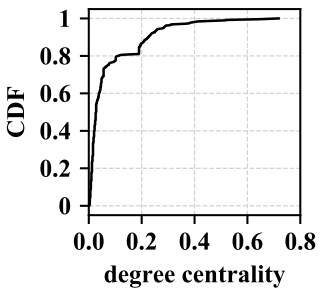


Fig. 3. The normalized degree centrality of ATS domains in projection graph G'.

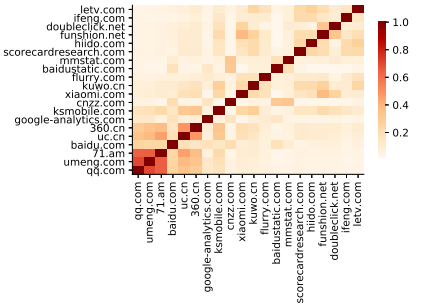


Fig. 4. The co-occurrence prob. distr. of the top 20 ATSes (SLDs).

and education apps are tracked by the highest number of third-party ATSes, and news and entertainment apps are exposed to a wide range of ATSes [27].

Takeaway. China's tracking market differs greatly from the western one. It is dominated by several popular domestic trackers. Over 30% of mobile trackers tend to be present in at least 2 apps, implying the prevalence of cross-app tracking of users. Tracking behavior varies across app categories mainly due to their functionality. The prominence of some types of apps (*e.g. InputMethods*) in tracking raises particular concerns for user privacy.

3.2 What is the community structure of ATSes?

Co-location of ATSes. The mobile trackers usually appear on as many apps as possible to enable cross-tracking of users, which leads to implicit connections between trackers through mobile apps. Inspired by [19], we further focus on the co-location of ATS domains within mobile apps by inspecting the trackers' community structure. To this end, we create a 1-mode ATS-projection graph G' from the largest connected component in G. In G', the vertices only contain the ATS domains in V and the edges are created if any two vertices share a common neighbor (app) in G. We find that trackers are very well-connected: nearly 99% of trackers appear in the largest connected component.

The ATS-projection graph G' captures the co-location of *multiple* tracking services used within individual apps. We first use the degree centrality (normalized by $N-1$, where N is the number of vertices in G') to measure how likely a tracker tends to co-locate with others (see Fig. 3). We can clearly identify two types of trackers: the *popular ones* with the normalized degree centrality over 0.2, the rest are *non-popular* ones that sparsely connect with others in the graph. Indeed, the popular ones are present more pervasively among apps than the non-popular ones. We further utilize the global clustering coefficient to measure the degree to which nodes in the graph G' tend to cluster together [36]. The coefficient is as high as 0.52. We also calculate the clustering coefficients for individual nodes — the results reveal low coefficients for the popular trackers, but high coefficients for the non-popular ones. These results imply that G' a

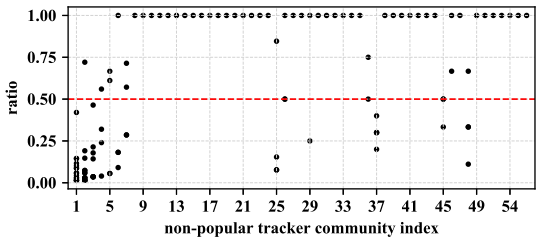


Fig. 5. Tracker Specialization Index distr. of non-popular tracker communities.

well connected graph, where the non-popular trackers form local communities,[8] while the popular trackers densely co-occur with the non-popular ones.

To verify the above conjecture on the structure of G', we remove the popular trackers from G' and obtain a graph G'' consisting of non-popular trackers. Approximately 62% of non-popular trackers appear in the largest connected component of G'' and the others consist of 46 isolated components in G''. We leverage the Clauset-Newman-Moore greedy method [9] for inferring community structure. We discover a total of 56 local communities, where 10 communities constitute the largest connected component. The global clustering coefficient of G'' is as high as 0.78. These results confirm the structure of G'. As we will show later, the trackers of each community tend to track one particular app category.

We next examine the popular trackers to see whether they are co-located in the same apps with each other. To this end, we compute the *Jaccard Similarity Coefficient* to quantify how likely two popular trackers, a and b, are to co-occur within the same target app. We calculate $\frac{|U(a)\bigcap U(b)|}{|U(a)\bigcup U(b)|}$, where $U(a)$ and $U(b)$ are the sets of apps tracked by a and b. Fig. 4 presents the coefficients between each of the top 20 popular ATS SLDs. The lower left portion of the heatmap exhibits high levels of co-location, primarily among tracking domains operated by qq.com, umeng.com, and 71.am, indicating that these popular trackers tend to co-occur with each other. Since their holding companies are Tencent, Alibaba, and Baidu, respectively, these three (Chinese) tech giants generally offer complementary, albeit competitive, services. In contrast, there are a number of trackers which show negligible correlation. Most prominently, international rival services, such as baidu.com and google-analytics.com, tend not to co-occur.

Specialization of ATSes. The above analysis leads us to explore the specialization of non-popular trackers, *i.e.* whether a local community of ATSes intends to occur in some specific app categories. To this end, we compute the *tracker specialization index* (TSI) to measure the extent to which an ATS local community is dedicated to a certain app category. The TSI is calculated as $\frac{|U(a)\bigcap U(b)|}{|U(a)|}$, where $U(a)$ and $U(b)$ are the sets of trackers in the ATS local community a and app category b.

[8] Communities are groups of vertices that are well-connected internally while sparsely connected with others.

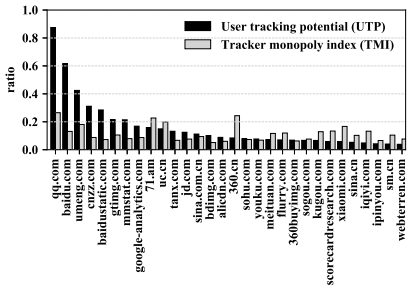


Fig. 6. UTP and TMI distr. of the top 30 tracker domains (SLDs).

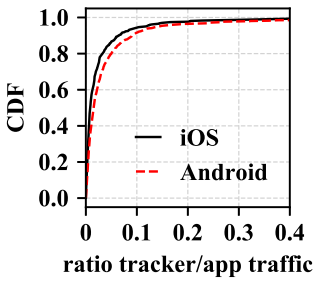


Fig. 7. The distribution of the ratio of tracker/app traffic volume for each user.

We plot the distribution of the *tracker specialization index* for 56 non-popular tracker communities in Fig. 5. We observe that ATS local communities tend to be specialized in only one or two app categories with $TSI \geq 0.5$, *i.e.* they provide specialized tracking services relevant to particular apps. For instance, the *Education* apps are mostly tracked by some ATS local communities run by the companies providing educational related services. Specifically, the parenting app Yaolan is mostly tracked by the following ATS local communities: <yaolan.com, yaolanimage.cn> run by Yaolan itself and <pcbaby.com.cn, pconline.com.cn> run by the app PCbaby that also provides parenting or educational services.

Takeaway. Mobile trackers are interconnected because popular trackers are regularly co-occur (in the same apps) with non-popular ones. The non-popular trackers, however, form many local communities, and the trackers in each local community tend to track a special category of mobile apps. The very top ATSes are often co-located in the same app, implying pervasive tracking.

3.3 How are users impacted by ATSes?

ATS Monopolies. The heavy-tailed distribution of ATS popularity leads us to conjecture that some may have a monopoly on certain user's data, *i.e.* a user may exclusively be tracked by a single ATS. To test this, we compute two metrics. First, *user tracking potential* (UTP) measures the number of users that can be potentially tracked by a mobile tracker. Given the set of all mobile users R, the tracker i's UTP is $UTP_i = \frac{|S_i|}{|R|}$, where $S_i \subset R$ is the set of users that the tracker i can reach. Second, *tracking monopoly index* (TMI) measures the extent to which a tracker reaches users that others do not have. Let m_j denote the number of mobile trackers that can reach the user $j \in S_i$. The TMI of the tracker i is $TMI_i = \frac{1}{|S_i|} \sum_{j \in S_i} \frac{1}{|m_j|}$. A high TMI indicates that some users are exclusively reached by the tracker and maybe due to trackers' high prevalence or specific coverage on mobile users.

Fig. 6 shows the distribution of *user tracking potential* and *tracking monopoly index* of the top 30 ATS domains (SLDs). We rank the tracker domains in descending order by the UTP values. The result reveals a high penetration of the

Table 3. Common UIDs host on mobile devices.

UID	Description	UID	Description
IMSI	SIM ID	**MAC**	Unique hardware ID
IMEI	Device ID	**ADID/IDFA**	Advertising ID
ICCID	SIM number		

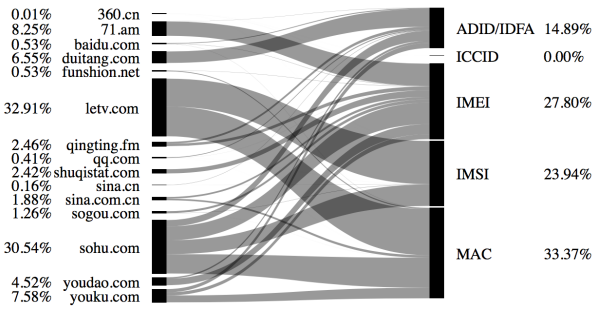


Fig. 8. Tracking domains (SLDs) that collect PII.

tech giants in China. For example, *qq.com* (owned by Tencent) holds a high UTP (over 0.8) and TMI (about 0.3) metrics, which reveals its high popularity and tracking monopoly. In addition, although under 20% of mobile users are tracked by *71.am* (owned by Baidu), *uc.cn* (owned by Alibaba) and *360.cn* (owned by 360 security), these trackers have relatively high TMIs. This indicates that there is a significant pooling of tracker data within this small elite, similar to that achieved by companies such as Google and Facebook in the western context.

ATS* vs. *App Traffic Volumes. Regardless of privacy implications, the data sent to trackers creates increased resource usage (on devices and within the network). We are next curious to see what volume of each user's traffic is generated by ATSes. Thus, we compute the ratio of tracking traffic to app traffic for individual users, and plot the distribution in Fig. 7. The median ratio is around 1%. Nevertheless, 5% of users send over 10% of their traffic to trackers. That said, the tracking traffic ratio per user is actually lower than that observed in an equivalent European 3G ISP [32], possibly due to the pervasive availability of online videos (used by 57% of users) in the 4G network. Interestingly, the device OS also has an impact on this ratio: iOS users (median 0.9%) tend to send less data to trackers than Android users (median 1.5%). This observation is in accord with the 3G network [32].

PII Leakage in ATS and Regional Destination. We next proceed to explore if any personally identifiable information (PII) is uploaded to ATS domains. We process each URL from all user access logs in our dataset to test for the presence of any PII. We use regular expressions to detect the common UIDs on mobile devices, *e.g.* `*\?imei=*` or `*&imei=*`. Table 3 summarizes the things we check for, as inspired by [27, 29]. In our analysis all the UIDs collected are anonymized to protect user privacy. To check whether the identified UIDs indeed contain PII, we leverage a small dataset of about 10K access logs collected at our

lab's wireless access point for one day.[9] Each log in this dataset contains similar information to the ones used in this paper. We applied the UID detection to this dataset and found that 80% of identified IMEIs, 95% of IMSIs, 83% of MACs and 92% of ADIDs/IDFAs indeed contain PII. This lends evidence to the claim that inferred UID exposure detected from the DPI dataset is often correct.

Our analysis reveals a worrying volume of PII leakage: as many as 10% of users send their PII to trackers via their mobile apps. Fig. 8 shows the distribution of how several popular tracker SLDs receive PII from apps. For each ATS domain, say sohu.com, the percentage on the left represents the number of flows that contain UIDs. For each type of PII, the percentage on the right represents the number of flows that belong to each of the SLDs. IMEI, IMSI, and MAC are equally likely to be collected by these trackers. The ATSes that upload the largest volume of PII are letv.com (ad online video service) and sohu.com (a mixture of services including ads and video): a remarkable 60% of PII relevant flows belong to them. Each ATS shows clear preferences towards certain PII (shown in Fig. 8). For instance, letv.com mainly collects IMSI and MAC information, while sohu.com shows balanced interests across four types of PII. In contrast, ICCID is only accessed by 360.cn (security service).

A particular concern is whether PII is sent across borders to other countries or regions [18]. We find that more than 90% of PII tracking flows are inside mainland China by mapping IP geo-locations in China [37]. This may be largely driven by the predominance of Chinese ATSes and the blocking of several key US trackers (*e.g.* Google, Facebook), as well as the extensive support for HTTPs in the majority of western countries (which is excluded from our analysis).

Takeaway. Several tech giants in China track the majority of users. Some specialized trackers, while having relatively small user coverage, track specific groups of users that others do not track. For 5% of users, 10% of their traffic is attributable to ATS flows. 10% users are exposed to PII leakage. Nevertheless, 90% of the PII data is local to China.

4 Conclusion and Discussion

This paper provides insights into the distinctive mobile tracking behavior in China. We make several interesting observations with respect to ATS popularity and community structure, user monopoly patterns, and PII collection. This study not only validates many previous findings, but also facilitates fresh analysis of tracking behavior in China. We believe that our first look at China's mobile tracking patterns has significant implications for many stakeholders in the mobile tracking community (*e.g.* app vendor, tracker provider, adblocker). For instance, adblockers can leverage the community structure for new tracker detection and the prevalence of cross-app tracking raises serious privacy concerns. Many of the findings are indeed worth further exploration, such as the tracker detection, the PII collection, and the business relationships between mobile trackers.

[9] Every member in the lab was notified about this experiment and consented.

Acknowledgments

We would like to thank David Choffnes for shepherding our paper and PAM reviewers for their useful feedback. This work was supported, in part, by National Key R&D Program of China under Grant No. 2018YFB1800201 and the Youth Innovation Promotion Association CAS.

References

1. Arzt, S., Rasthofer, S., Fritz, C., Bodden, E., Bartel, A., Klein, J., Le Traon, Y., Octeau, D., McDaniel, P.: Flowdroid: Precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for android apps. Acm Sigplan Notices **49**(6), 259–269 (2014)
2. Backes, M., Bugiel, S., Derr, E.: Reliable third-party library detection in android and its security applications. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. pp. 356–367. ACM (2016)
3. Binns, R., Zhao, J., Kleek, M.V., Shadbolt, N.: Measuring third-party tracker power across web and mobile. ACM Transactions on Internet Technology (TOIT) **18**(4), 52 (2018)
4. Book, T., Wallach, D.S.: An empirical study of mobile ad targeting. arXiv preprint arXiv:1502.06577 (2015)
5. Chen, J., Chen, H., Bauman, E., Lin, Z., Zang, B., Guan, H.: You shouldn't collect my secrets: Thwarting sensitive keystroke leakage in mobile {IME} apps. In: 24th {USENIX} Security Symposium ({USENIX} Security 15). pp. 657–690 (2015)
6. Chen, T., Ullah, I., Kaafar, M.A., Boreli, R.: Information leakage through mobile analytics services. In: Proceedings of the 15th Workshop on Mobile Computing Systems and Applications. p. 15. ACM (2014)
7. Cisco: Visual networking index: Global mobile data traffic forecast update, 2017–2022 white paper. Tech. rep., Cisco (2019)
8. CIW: ebook: Top 200 mobile apps in china (2018), https://www.chinainternetwatch.com/ebook/top-mobile-apps/
9. Clauset, A., Newman, M.E., Moore, C.: Finding community structure in very large networks. Physical review E **70**(6), 066111 (2004)
10. Easylist: The primary filter list that removes most adverts from international webpages (2016), https://easylist.to/
11. Egele, M., Kruegel, C., Kirda, E., Vigna, G.: Pios: Detecting privacy leaks in ios applications. In: NDSS. pp. 177–183 (2011)
12. Enck, W., Gilbert, P., Han, S., Tendulkar, V., Chun, B.G., Cox, L.P., Jung, J., McDaniel, P., Sheth, A.N.: Taintdroid: an information-flow tracking system for realtime privacy monitoring on smartphones. ACM Transactions on Computer Systems (TOCS) **32**(2), 5 (2014)
13. Gill, P., Erramilli, V., Chaintreau, A., Krishnamurthy, B., Papagiannaki, K., Rodriguez, P.: Follow the money: understanding economics of online aggregation and advertising. In: Proceedings of the 2013 conference on Internet measurement conference. pp. 141–148. ACM (2013)
14. Halfaker, A., Keyes, O., Kluver, D., Thebault-Spieker, J., Nguyen, T., Shores, K., Uduwage, A., Warncke-Wang, M.: User session identification based on strong regularities in inter-activity time. In: Proceedings of the 24th International Conference on World Wide Web. pp. 410–418. International World Wide Web Conferences Steering Committee (2015)

15. Han, S., Jung, J., Wetherall, D.: A study of third-party tracking by mobile apps in the wild. Univ. Washington, Tech. Rep. UW-CSE-12-03-01 (2012)
16. Ikram, M., Masood, R., Tyson, G., Kaafar, M.A., Loizon, N., Ensafi, R.: The chain of implicit trust: An analysis of the web third-party resources loading. Web Conference (2019)
17. Ikram, M., Vallina-Rodriguez, N., Seneviratne, S., Kaafar, M.A., Paxson, V.: An analysis of the privacy and security risks of android vpn permission-enabled apps. In: Proceedings of the 2016 Internet Measurement Conference. pp. 349–364. ACM (2016)
18. Iordanou, C., Smaragdakis, G., Poese, I., Laoutaris, N.: Tracing cross border web tracking. In: Proceedings of the Internet Measurement Conference 2018. pp. 329–342. ACM (2018)
19. Kalavri, V., Blackburn, J., Varvello, M., Papagiannaki, K.: Like a pack of wolves: Community structure of web trackers. In: Karagiannis, T., Dimitropoulos, X. (eds.) Passive and Active Measurement (2016)
20. Li, H., Lu, X., Liu, X., Xie, T., Bian, K., Lin, F.X., Mei, Q., Feng, F.: Characterizing smartphone usage patterns from millions of android users. In: Proceedings of the 2015 Internet Measurement Conference. pp. 459–472. ACM (2015)
21. Lightbeam: Shine a light on who is watching you (2019), https://addons.mozilla.org/fr/firefox/addon/lightbeam-3-0/
22. Liu, M., Wang, H., Guo, Y., Hong, J.: Identifying and analyzing the privacy of apps for kids. In: Proceedings of the 17th International Workshop on Mobile Computing Systems and Applications. pp. 105–110. ACM (2016)
23. MalwareBytes: hphosts (2019), http://hosts-file.net/
24. policy, E.: Filter evaluation (2011), https://easylist.to/2011/08/31/what-is-acceptable-first-party-tracking.html
25. Qiu, L., Zhang, Z., Shen, Z., Sun, G.: Apptrace: Dynamic trace on android devices. In: 2015 IEEE International Conference on Communications (ICC). pp. 7145–7150. IEEE (2015)
26. Rao, A., Sherry, J., Legout, A., Krishnamurthy, A., Dabbous, W., Choffnes, D.: Meddle: middleboxes for increased transparency and control of mobile traffic. In: CoNEXT Student Workshop (2012)
27. Razaghpanah, A., Nithyanand, R., Vallina-Rodriguez, N., Sundaresan, S., Allman, M., Gill, C.K.P.: Apps, trackers, privacy, and regulators. In: 25th Annual Network and Distributed System Security Symposium, NDSS. vol. 2018 (2018)
28. Razaghpanah, A., Vallina-Rodriguez, N., Sundaresan, S., Kreibich, C., Gill, P., Allman, M., Paxson, V.: Haystack: In situ mobile traffic analysis in user space. arXiv preprint arXiv:1510.01419 pp. 1–13 (2015)
29. Ren, J., Rao, A., Lindorfer, M., Legout, A., Choffnes, D.: Recon: Revealing and controlling pii leaks in mobile network traffic. In: Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services. pp. 361–374. ACM (2016)
30. Seneviratne, S., Seneviratne, A., Mohapatra, P., Mahanti, A.: Your installed apps reveal your gender and more! ACM SIGMOBILE Mobile Computing and Communications Review **18**(3), 55–61 (2015)
31. Su, J., Li, Z., Grumbach, S., Ikram, M., Salamatian, K., Xie, G.: A cartography of web tracking using dns records. Computer Communications **134**, 83–95 (2019)
32. Vallina-Rodriguez, N., Shah, J., Finamore, A., Grunenberger, Y., Papagiannaki, K., Haddadi, H., Crowcroft, J.: Breaking for commercials: characterizing mobile advertising. In: Proceedings of the 2012 Internet Measurement Conference. pp. 343–356. ACM (2012)

33. Vallina-Rodriguez, N., Sundaresan, S., Razaghpanah, A., Nithyanand, R., Allman, M., Kreibich, C., Gill, P.: Tracking the trackers: Towards understanding the mobile advertising and tracking ecosystem. arXiv preprint arXiv:1609.07190 (2016)
34. Wang, H., Guo, Y.: Understanding third-party libraries in mobile app analysis. In: 2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C). pp. 515–516. IEEE (2017)
35. Wang, H., Liu, Z., Liang, J., Vallina-Rodriguez, N., Guo, Y., Li, L., Tapiador, J., Cao, J., Xu, G.: Beyond google play: A large-scale comparative study of chinese android app markets. In: Proceedings of the Internet Measurement Conference 2018. pp. 293–307. ACM (2018)
36. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world'networks. nature **393**(6684), 440 (1998)
37. Xiang, C., Wang, X., Chen, Q., Xue, M., Gao, Z., Zhu, H., Chen, C., Fan, Q.: No-jump-into-latency in china's internet!: Toward last-mile hop count based ip geo-localization. In: Proceedings of the International Symposium on Quality of Service. pp. 42:1–42:10. IWQoS '19, ACM (2019)
38. Xu, Q., Erman, J., Gerber, A., Mao, Z., Pang, J., Venkataraman, S.: Identifying diverse usage behaviors of smartphone apps. In: Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. pp. 329–344. ACM (2011)
39. Yao, H., Ranjan, G., Tongaonkar, A., Liao, Y., Mao, Z.M.: Samples: Self adaptive mining of persistent lexical snippets for classifying mobile application traffic. In: Proceedings of the 21st Annual International Conference on Mobile Computing and Networking. pp. 439–451. ACM (2015)
40. Zang, J., Dummit, K., Graves, J., Lisker, P., Sweeney, L.: Who knows what about me? a survey of behind the scenes personal data sharing to third parties by mobile apps. Technology Science **30** (2015)