# Mobile Content Hosting Infrastructure in China: A View from a Cellular ISP

Zhenyu Li[1,2](✉), Donghui Yang[1,2], Zhenhua Li[3], Chunjing Han[1,2], and Gaogang Xie[1,2]

[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
{zyli,yangdonghui,hcj,xie}@ict.ac.cn
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Tsinghua University, Beijing, China
lizhenhua1983@gmail.com

**Abstract.** Internet users are heavily relying on mobile terminals for content access, where the content is hosted and delivered by either third-party infrastructures (*e.g.,* CDNs and clouds) or the content providers' own delivery networks, or both. China has the largest mobile Internet population in a single country, and also has unique local regulations and network policies (e.g. heavy content censorship). The content delivery ecosystem in China, as such, may show great disparity from the western one. Yet, there is little visibility into the content hosting infrastructure in Chinese cellular networks. This paper makes the first step toward filling this gap by analyzing a passive DNS trace that consists of 55 billion DNS logs collected from a national-scale cellular ISP. Our in-depth investigation of the content-related features of major ASes reveals that content objects of popular domains are replicated deep into the examined cellular ISP. On the other hand, as much as 20% of tracking traffic, which is mainly generated by trackers owned US-based companies, goes out of China. Our findings cast useful insights for cellular ISPs, CDNs and Internet policy makers.

## 1 Introduction

The ever-growing popularity of smart devices greatly promotes the content demand in cellular networks. It was projected that the mobile data traffic will grow 7-fold in the upcoming years [5]. Such an enormous demand challenges not only cellular network itself, but also the content hosting infrastructure that delivers content to massive users. Typical content hosting infrastructure includes third-party infrastructures (*e.g.,* CDNs, clouds), the content providers' own delivery networks, and a mixture of the two. Content hosting infrastructures have a significant impact on ISPs' traffic engineering, and quality of experience perceived by end users. For instance, a centralized infrastructure needs to peer its data centers with ISPs for high bandwidth [12], while a distributed one needs to deploy its servers as close to users as possible for fast content access.

Content hosting infrastructures is largely shaped by the cost, network policies as well as local regulations where they are deployed. This paper examines the


R. Beverly et al. (Eds.): PAM 2018, LNCS 10771, pp. 100–113, 2018.
https://doi.org/10.1007/978-3-319-76481-8_8

mobile content hosting infrastructures in China. China has the largest mobile Internet population in a single country, and, perhaps more interestingly, has unique local regulations and network policies. For instance, Internet Content Provider (ICP) licenses are mandatory for the sites that aim at delivering content within mainland China. This regulation prevents popular CDNs (like Akamai) from deploying their replica servers in China [17].

The above factors may lead to great disparity of the content hosting infrastructures in China than the western countries. Unfortunately, we have very limited knowledge of the infrastructures in China, despite some recent studies on that in western countries. Triukose *et al.* [16] measured Akamai, and examined the performance benefit of using distributed deployment. Pujol *et al.* [15] on the other hand studied the hosting infrastructure for advertisement trackers of DSL web users. Since DNS maps end users to specific servers [10], using DNS replies can infer what content is hosted in which locations. The web content cartography introduced in [7] was the first step to use DNS replies for this purpose. However, they focused only on a small amount of domains in wireline networks. Xue *et al.* [17] also use active DNS measurements of a few top domains to study the server selection policies used by CDNs in China.

This paper considers *all* domains requested by mobile users through cellular networks. We analyzed 55 billion DNS replies collected from all recursive resolvers of a cellular ISP. The large user coverage of the data enables us to have a comprehensive view of the content hosting infrastructure. We borrow the content-related metrics in [7] to characterize the features of the ASes that accounts for majority of the DNS queries. We further propose a clustering algorithm to identify content hosting providers, and examined the features of the major providers. Finally, this paper examines the hosting infrastructure of tracking domains, which are present in mobile webs and more prominently in mobile apps [13]. We have also discussed the implications of our major findings from different aspects. To sum up, we make the following main contributions.

- *Hosting Infrastructure Concentration:* We find that cellular content infrastructure is concentrated in a few ISP ASes, rather than CDN ASes. This stems from the fact that content objects of popular domains have been deeply replicated into ISPs. On the contrary, there is a trend that non-popular domains outsource the hosting services to third-party clouds that currently rarely deploy caches into ISP networks.
- *Hosting Provider Identification:* We propose a clustering algorithm for identifying hosting providers from passive DNS replies of massive domains. Specifically, we apply spectral clustering on the bipartite graph formed by domains and IP /24 subnets. We show the evidence that major providers slice up their infrastructures to host different kinds of services.
- *Tracker Hosting Infrastructure:* We reveal that while the examined ISP account for the largest amount of tracking queries, over 20% of the tracking queries are still mapped to foreign ASes. Besides, we surprisingly observe that as many as 60% of the tracking servers (*i.e.,* servers used to deliver tracking content) and 52 ASes are exclusively used for tracking service.

## 2 Data and Metrics

### 2.1 Data: Passive DNS Replies

We collected our DNS data from the recursive resolvers of a cellular ISP in China. Once connected to the cellular network, mobile terminals will be automatically assigned a recursive resolver that the ISP operates. A recursive resolver receives hostname resolution requests from client hosts, and iteratively interacts with the hierarchical naming system to translate the names to IP addresses. The last step of this iterating process involves contacting the authoritative servers that maintain the mapping of the queried names to addresses. The authoritative servers often map the names to the domain hosting servers that are as proximate as possible to the recursive resolvers, in the hope that the hosting servers are also close to client hosts [10].

The examined ISP keeps a record for each DNS query at its recursive resolvers. A record consists of the recursive resolver's identifier, the timestamp, the requested domain name, the IP lists in the response, and finally the return code in the response. The records contain no specific information of client hosts for privacy concerns.

In total, we obtained 55,412,725,137 records from *all* the recursive resolvers of the examined cellular ISP for a duration of 2 days in 2015. The records are of A (IPv4) queries, *i.e.,* no AAAA (IPv6) queries were seen. By looking at the return (error) code, we observe a resolution successful ratio (*i.e.,* the ratio of records with "NOERROR") of 96.76%. Besides, over half of the hostnames map to more than one IP address. Our analysis, however, takes the first IP address as the one that the hostname is mapped to. This is reasonable because, in most cases, the first IP address is used for the following connection [9].

**Data pre-processing:** To simplify the analysis of such a huge dataset, we map the DNS FQDNs (Fully Qualified Domain Names) to their second level domains (SLDs) using the public suffix library [2]. The simplification yields 1,410,727 SLDs. The popularity of SLDs follows a power-law distribution, where less than 1% of the domains account for 80% of the queries. We further map IP addresses to their AS number (ASN) by querying Team Cymru [3]. We further aggregate the IP addresses in DNS responses into /24 subnetworks for the examination of the network footprints of domains. This aggregation granularity takes the fact that server clusters are often deployed for content hosting to achieve resilience and load balancing [7].

**Ethical issue:** The DNS dataset contains no information of individual users, and we were unable to link queries to users. It is also noteworthy that such datasets are routinely gathered by DNS servers in form of logs for security and operational purposes.

### 2.2 Content-Related Metrics

We use two metrics to characterize the content-related features of ASes. The first one is *content delivery potential* (CDP) [7], which gauges the amount of content that can be potentially served by an AS. Given a set of SLDs $R$ (*e.g.,* tracking domains), the AS $i$'s CDP is $CDP_i = \frac{|S_i|}{|R|}$, where $S_i \subseteq R$ is the set of domains that the AS can serve.

The second metric is *content monopoly index* (CMI) [7], which measures the extent to which an AS hosts content that others do not have. Let $R$ denote the set of SLDs under consideration, $S_i \subseteq R$ the set of SLDs hosted by AS $i$, and $m_j$ the number of ASes that host the SLD $j \in S_i$. The CMI of AS $i$ is $CMI_i = \frac{1}{|S_i|} \sum_{j \in S_i} \frac{1}{m_j}$. A high CMI means some content is exclusively available in the AS.

# 3 On Hosting Infrastructure

## 3.1 Content Potential of ASes

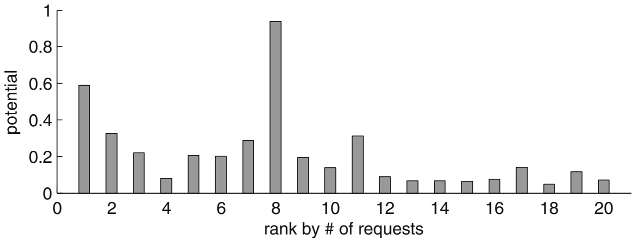
Table 1 lists the top 20 ASes in terms of the volume of DNS queries that are resolved successfully. These ASes account for over 90% of the DNS queries.

**Table 1.** Top 20 ASes ranked by the volume of queries.

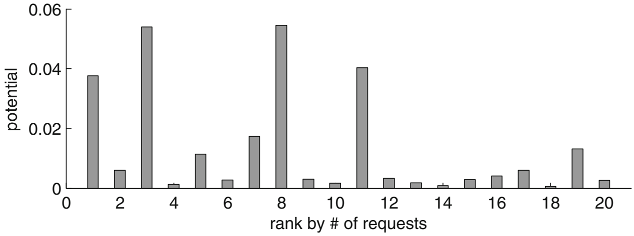
| Rank | AS name[a] | vol. (%) | $CMI_{top}$ | $CMI_{all}$ |
|---|---|---|---|---|
| 1 | ISP-AS1 | 40.99 | 0.18 | 0.63 |
| 2 | ISP-AS2 | 24.59 | 0.12 | 0.37 |
| 3 | Alibaba | 6.32 | 0.19 | 0.91 |
| 4 | Apple | 4.88 | 0.05 | 0.12 |
| 5 | Chinanet-BJ | 3.91 | 0.13 | 0.57 |
| 6 | ISP-AS3 | 2.19 | 0.09 | 0.23 |
| 7 | China169-back | 1.38 | 0.11 | 0.65 |
| 8 | ISP-AS4 | 1.33 | 0.26 | 0.52 |
| 9 | ISP-AS5 | 1.05 | 0.10 | 0.26 |
| 10 | ISP-AS6 | 0.94 | 0.07 | 0.22 |
| 11 | Chinanet-back | 0.81 | 0.13 | 0.75 |
| 12 | Akamai-ASN1 | 0.79 | 0.06 | 0.35 |
| 13 | Akamai-AS | 0.76 | 0.05 | 0.34 |
| 14 | Chinacache | 0.67 | 0.06 | 0.23 |
| 15 | CNIX | 0.56 | 0.09 | 0.73 |
| 16 | Chinanet-SN | 0.54 | 0.06 | 0.56 |
| 17 | China169-BJ | 0.54 | 0.09 | 0.65 |
| 18 | Yahoo-SG | 0.52 | 0.03 | 0.09 |
| 19 | Tencent | 0.50 | 0.11 | 0.83 |
| 20 | Google | 0.40 | 0.05 | 0.53 |

[a] Due to business considerations, we cannot reveal the name of the examined ISP. Rather, we use ISP to denote it.

Besides, most of the queries are resolved to ISPs, rather than third-party content hosting providers, like Akamai. An AS appearing in the top list is because of either hosting either very popular domains, or hosting a large quantity of domains. The content delivery potential (CDP) of ASes in Fig. 1 exactly answers this question, where in Fig. 1a only the top 10,000 popular domains are considered when computing CDP, while Fig. 1b considers all domains.




(a) considering top 10k popular domains




(b) considering all domains

**Fig. 1.** Content delivery potential of the top 20 ASes.

Figure 1a shows that the ASes of the examined ISP indeed hosts most of the popular domains. For instance, 95% of the popular domains can be served by `ISP`-AS4, and the top ranked one hosts about 60%. This observation implies that popular domains are well replicated in the examined ISP. The Apple's AS has a lower CDP, indicating that it appears in the list because of the frequent access of its domains from smartdevices, rather than hosting lots of domains.

When considering all domains in Fig. 1b, no AS hosts over 6% of the domains. This is within our expectation because most of the domains are only available in one single AS. It is also interesting to see that Alibaba cloud hosts the largest number of domains; Tencent cloud also hosts a significant fraction. The reason should be that some content owners, especially those of non-popular domains, outsource their domains to the clouds for easy maintenance and low access delay.

We further examine whether the listed ASes serve different or similar content in Fig. 2. An AS is associated with a content serving vector, and the $i$-th element is $< h_i, c_i >$, where $h_i$ is a SLD and $c_i$ is the number of queries on $h_i$ that are mapped to the AS. We compute the similarity between two ASes using

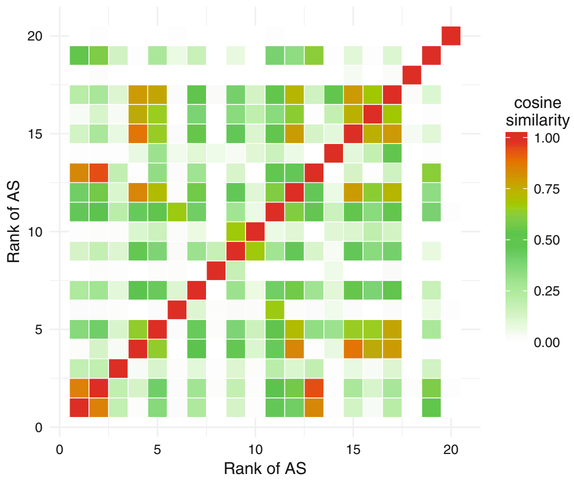



**Fig. 2.** Cosine similarity between each pair of ASes

the cosine similarity between their content serving vectors. Several observations are notable. First, the similarity values between the examined ISP's ASes are relative low, despite of the high content availability in the ASes. It seems that the ISP hosts different content in different ASes. Second, the relatively high similarity between Akamai's ASes and ISPs' ASes is an evidence that Akamai has already replicated content into ISPs, including the examined one and those that own China169 and Chinanet. Third, Alibaba and Tencent clouds host content that other ASes do not have, evidenced by the low similarity with other ASes. This is also confirmed by the high CMI values of these two ASes (see the last column of Table 1). Fourth, the high similarity between Apple and the ASes of Chinanet implies that Apple's content is available in Chinanet's ISP, but not the examined one. This is a potential performance bottleneck for the ISP's users to access Apple's content. Last but not least, Yahoo and Google's ASes host totally different content from others. This is because domains like google.com, flickr.com are blocked for access in China, so their content is not replicated to the ASes under consideration [8].

The above analysis, however, does not reveal the reasons of the presence of ISPs' ASes in Table 1. In fact, there are two possibilities. First, ISPs host content of popular domains that other ASes do not have, so that the queries of these domains can only be mapped to the ISPs' ASes. Second, content hosting providers deploy their content servers into the ISPs to boost the content delivery performance. Both cases may lead to a high CDP of an AS.

The content monopoly index (CMI, see Sect. 2) is used to investigate the first possibility. We observe low CMI values when considering only the top 10,000 domains (see the 4th column of Table 1), indicating that these ASes do not exclusively host content of *popular* domains that others do not have. When considering all domains, we observe high CMI values for some ASes, because they host lots of *non-popular* domains that are only available in the ASes. Moreover, the extremely high CMI values of the Alibaba cloud and Tencent cloud are further evidence of the trend of outsourcing content to clouds for non-popular domains that are less replicated.

For the second possibility, we aim at identifying the major content hosting providers. We do so in the next subsection by applying spectrum clustering on the bipartite graph formed by IP /24 subnets and domains.

### 3.2 Content Hosting Provider Analysis

We identify major hosting providers by clustering servers (identified by IP addresses) that are run by the same hosting provider. For this purpose, we form a bipartite graph, where one type of nodes is SLD, and the other is IP /24 subnet. An edge is present between a SLD node and a subnet node if the domain is mapped to the subnet in our dataset. Each edge is associated with a weight, which is defined later in this section. The key idea of clustering is that the /24 subnets used by a hosting provider serve the similar domains, and thus are densely connected through domains. Graph partitioning algorithms can thus be used for the clustering purpose.

Let $M \in \mathbb{R}^{m \times n}$ be the matrix representation of the bipartite graph. $M$ would be a sparse matrix, where rows are domains (*i.e.,* SLDs), and columns represent /24 subnets. $M_{ij}$ is the weight of the edge between the $i$-th domain and the $j$-th subnet. We set $M_{ij} = 1.0 + log(q_{ij})$, where $q_{ij}$ is the number of queries of the $i$-th domain that are mapped to the $j$-th subnet. The intuition of weight setting is that the higher $q_{ij}$ is, the more likely that the $j$-th subnet belongs to the hosting providers that deliver the $i$-th domain's content. We discard domains that are mapped to only one subnet to reduce the dimensionality. We finally apply a graph partitioning algorithm based on spectral clustering [14], as summarized in Algorithm 1, on $M$.

**Algorithm 1.** Spectral clustering of /24 subnets for the identification of content hosting providers

**Input**: $M \in \mathbb{R}^{m \times n}$
**Output**: Clusters of /24 subnets

1. $S \leftarrow M^T \cdot M$;
2. compute the first $k$ eigenvectors from $S$;
3. $k$ eigenvectors form $Q \in \mathbb{R}^{n \times k}$;
4. $v_i \leftarrow$ the $i$-th column of $Q^T$ $(1 \leq i \leq n)$, *i.e.,* the dimension-reduced representation of the $i$-th /24 subnet;
5. Cluster the vectors (*i.e.,* /24 subnets) $\{v_i\}_{i=1,\ldots,n}$ using the X-means clustering alg.

Each cluster yielded from the above algorithm represents a content hosting provider. To label the owners of clusters, we resort to the IP usage information from the examined ISP as well as third parties. The examined ISP maintains a table recording who uses which IP addresses (often represented in IP ranges). If there is one /24 subnet in a cluster belonging to the examined ISP, we looked up the table using the /24 subnet as key and get the entity name of the /24 subnet, which is further used as the owner of the clusters. Otherwise, we looked up third-party databases (*e.g.,* `whois` utility, MaxMind[1]) to infer cluster owners.

We find two exceptions during the labeling process. First, some clusters are labeled multiple owners. This happens because some domains, may leverage several CDNs for content distribution. For instance, Both Netflix and Hulu use three CDNs: Akamai, LimeLight and Level-3 [6]. The /24 subnets of these CDNs may be clustered into one cluster as they are connected by the same domains. We label them as `mixed`. Second, some owners have multiple clusters. This happens because an owner may provide multiple types of services, and it slices up its hosting infrastructure to host different services. For instance, Tencent uses one cluster of subnets for multimedia objects delivery and one for social network service hosting. In this case, we further infer the major services that a cluster provides by examining the domains in the cluster.

In total, we get 922 clusters. Table 2 lists the top 15 clusters, along with their network footprints and owners[2] These clusters account for over 50% of the queries. We can see the owners indeed are the major providers that provide a large amount of mobile content in China. As expected, the `mixed` ones contain more /24 subnets and have footprints in much more ASes than other clusters, because they contain several CDNs. The major CDN players in China, like ChinaCache and ChinaNetCenter, are included in the `mixed` clusters, because they are used by several popular Internet video providers (*e.g.,* PPTV, iQiyi).

The four clusters owned by Tencent distinguish from each other in the services that they provide. For instance, the first-ranked cluster hosts Tencent multimedia objects, while the second hosts Tencent social networks. We make similar observations for the Baidu's clusters. Xiaomi (a smartphone maker) appears in the list, because of the huge number of users using its smartphones, which frequently contact its cloud center for storage/retrieval of personal data, software download etc. Alibaba, on the other hand, hosts its own services (like alipay), as well as the outsourced content to it.

Akamai's clusters were identified by the prevalence of `akacdn.com` and `akamaiedge.net` in the clusters. Nevertheless, the /24 subnets do not necessarily belong to Akamai's AS, but the partners that Akamai collaborate with in China. Finally, we see Apple and Google in the list because of their prevalence in

---

[1] MaxMind: www.maxmind.com.

[2] We manually cross-checked the CNAMEs of the popular domains (FQDNs) in non-mixed clusters to validate the clustering approach. For example, the popular domains in both Baidu clusters use the CNAMEs with the same suffix `shifen.com`, which is run by Baidu.

**Table 2.** Top 15 clusters in terms of query volume

| Rank | volume (%) | # /24 subs | # ASes | Owner |
|---|---|---|---|---|
| 1 | 8.5 | 11 | 2 | Tencent |
| 2 | 7.0 | 4 | 1 | Tencent |
| 3 | 6.7 | 37 | 16 | `mixed` |
| 4 | 4.2 | 5 | 3 | Xiaomi |
| 5 | 3.9 | 3 | 1 | Akamai[a] |
| 6 | 3.6 | 3 | 1 | Tencent |
| 7 | 3.2 | 2 | 2 | Baidu |
| 8 | 2.9 | 6 | 1 | Alibaba |
| 9 | 2.6 | 4 | 2 | Baidu |
| 10 | 2.4 | 2 | 2 | Akamai[a] |
| 11 | 2.4 | 3 | 1 | Tencent |
| 12 | 2.3 | 81 | 30 | `mixed` |
| 13 | 2.3 | 47 | 24 | `mixed` |
| 14 | 2.1 | 8 | 3 | Google |
| 15 | 1.8 | 5 | 1 | Apple |

[a] The /24 subnets belong to a Chinese CDN provider, with which Akamai collaborates for content delivery.

mobile phone market. The Apple cluster mainly provides service for apple.com, and thus the volume share is less than the Apple AS showed in Table 1.

### 3.3 Summary and Discussion

Our analysis in this section has revealed that the cellular content infrastructure is mostly concentrated in the examined ISP's ASes. This implies a significant locality of cellular traffic. Besides, it means cellular users can get their content mostly within only one AS hop, since the ASes of the examined ISP are often peered with each other.

Our analysis also shows the trend of outsourcing non-popular domains to clouds. This implies cloud providers have already taken the niche market of content hosting. As this trend continues, cloud providers will become *de-facto* content providers that deliver a large amount of content that other ASes do not have (see Table 1). In fact, Tencent has already offered CDN service based on its cloud platform [4]. This may change the ecosystem of content hosting.

The proposed clustering algorithm provides a tool for content hosting provider identification from large-scale passive DNS datasets. The above analysis provides evidence of slicing-up infrastructure by hosting providers to deliver different kinds of content.

## 4 Tracker Hosting Infrastructure

This section examines the hosting infrastructure of tracking domains (*a.k.a.* trackers), because tracking is prevalent in mobile web service and mobile apps. Mobile users are getting concerned about the possible privacy leakage. Besides, it would be interesting to see the impact of content censorship on tracking behavior.

To identify the tracking domains, we used lists of trackers proposed by Ad blockers. More precisely, we merged two lists: *EasyList* (combined with the EasyList China supplementary list)[3] and *Simple Malvertising*[4]. Each queried hostname is labeled as tracker or normal depending on the suffix match with a hostname in the lists. In total, we find 124,235 tracking domains, which are further aggregated to 1,456 second-level domains.

### 4.1 Top Trackers

We first examine the top 10 tracking domains in terms of DNS query volume and their network features in Table 3. These domains account for 90% of total tracking queries, showing a very biased distribution of tracking traffic. It is surprising to see only 2 tracking domains are based in China, and most in US. We conjecture the prevalence of Android phones and the availability of mobile third-party analytics libraries are the main reasons for this observation [11].

**Table 3.** Top 10 tracking (second-level) domains

| Domain | Vol.% | Type* | #ASes | Owner |
|---|---|---|---|---|
| flurry.com | 35.07 | `an` | 11 | Yahoo |
| crashlytics.com | 25.25 | `an` | 18 | Google |
| scorecardresearch.com | 18.53 | `an` | 21 | comScore |
| doubleclick.net | 3.38 | `ad` | 24 | Google |
| adsmogo.com | 1.77 | `ad` | 9 | Alibaba |
| tapjoy.com | 1.71 | `ad` | 11 | Tapjoy |
| inmobi.com | 1.61 | `ad` | 14 | InMobi |
| tapjoyads.com | 1.56 | `ad` | 4 | Tapjoy |
| 51yes.com | 1.31 | `an` | 20 | 51yes |
| vungle.com | 0.84 | `ad` | 9 | Vungle |

* `an`: analytics, `ad`: advertiser

[3] https://easylist.to.
[4] https://disconnect.me/lists/malvertising.

### 4.2 Tracker Hosting Infrastructure

We then focus on the tracking servers (identified by IPs) that host the trackers. We say a server is a tracking server if more than 10 tracking queries are resolved to the server's IP address[5]. In total, 7,404 tracking servers are identified.

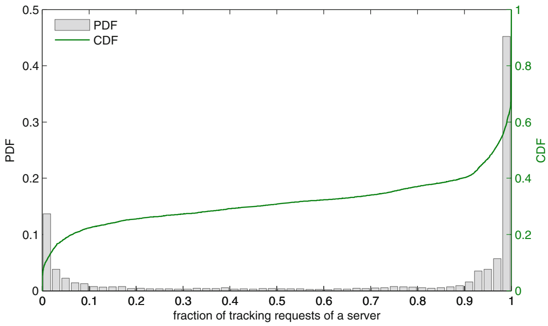



**Fig. 3.** Distr. of the ratio of tracking queries of servers. The left $y$-axis is for probability distr. function, while the right one is for cumulative distr. function.

A tracking server may host both tracking domains and non-tracking ones. For each tracking server, we compute the ratio of tracking queries (*i.e.,* queries to tracking domains) to all queries resolved to it, and plot the distribution in Fig. 3. We observe a bimodal distribution, where most of tracking servers either deliver a very small ratio of tracking queries (*i.e.,* $< 0.1$), or dedicate most of its capacity for tracker hosting. As in [15], we consider a server dedicated exclusively for hosting tracking service if the ratio of tracking queries exceeds 0.9. This is reasonable given that the two lists used for tracker identification may not cover all tracking domains in our trace. We surprisingly find as many as 4,427 (59.8%) tracking servers are exclusively used for hosting trackers, and these servers account for half of the tracking queries in our dataset.

Next, we study the ASes that host most of the trackers in Table 4. The tracking traffic is also mostly concentrated in the examined ISP's ASes. It means the hosting service of trackers has also been deployed into cellular networks. Besides ISPs, we also see cloud providers (Amazon and Internap), CDNs (Akamai) and search engines (Google), implying diverse infrastructures being used for tracking services. Despite of the traffic concentration in the examined ISP, a considerable fraction ($> 20\%$) still goes to the ASes that have rare footprints in China.

We then report the ratio of tracking queries to all queries of ASes (see the 3rd column in Table 4). It is surprising to see some ASes (*e.g.,* Internap) being exclusively used for tracker content delivery. We then compute the ratio of tracking queries for the ASes that have at least 1,000 tracking queries resolved to them.

[5] Due to DNS caching, we may underestimate the queries mapped to individual IP addresses.

**Table 4.** Top 10 ASes by tracking requests

| AS name | % tracking in trace | % tracking in AS | CDP | CMI |
|---|---|---|---|---|
| ISP-AS1 | 35.27 | 1.89 | 0.03 | 0.12 |
| ISP-AS2 | 24.10 | 0.77 | 0.12 | 0.42 |
| Amazon-AES | 7.96 | 54.79 | 0.09 | 0.30 |
| Internap-B.4 | 7.01 | 100.00 | $< 0.01$ | 0.11 |
| ISP-AS3 | 5.64 | 25.29 | $< 0.01$ | 0.05 |
| ISP-AS4 | 3.89 | 3.84 | 0.34 | 0.35 |
| Amazon-02 | 2.96 | 14.77 | 0.11 | 0.36 |
| GoogleCN | 2.32 | 27.28 | $< 0.01$ | 0.17 |
| NTT | 1.43 | 34.33 | 0.06 | 0.16 |
| Akamai-ASN | 1.04 | 1.74 | 0.09 | 0.20 |

Again, we use the threshold 0.9 to determine whether an AS exclusively hosts trackers or not. As many as 52 ASes are identified as exclusive tracking ones. They are either cloud providers (*e.g.,* Internap, Carpathia), or owners of trackers that run their own ASes (*e.g.,* Crashlytics).

We finally report the content delivery potential (CDP) and content monopoly index (CMI) of the ASes when considering only tracking domains in the last two columns of Table 4. We see low CDP ($< 0.1$) for most of ASes because they host only several popular tracking domains. The CMI is also relatively low, meaning that the tracking domains hosted by these ASes are also available in other ASes.

### 4.3 Summary and Discussion

We observe that the tracking queries are concentrated in a small number of trackers, of which most are US based. Moreover, over 20% of the tracking traffic goes out of China. These observations raise privacy and cybersecurity concerns. The analysis also reveals that multiple types of infrastructures are used for tracker service hosting.

The bimodal distribution of the tracking query ratio shows that 60% of the tracking servers exclusively provide tracking services. Monitoring the traffic going to these servers may help us find new trackers that also rely on these servers for content delivery. ISPs and mobile apps can also use this observation to block tracking activities for privacy and security concerns.

## 5 Conclusion

This paper uses passive DNS traces from a Chinese cellular ISP to investigate the mobile content hosting infrastructure in China. To this end, we proposed a clustering algorithm to identify hosting providers and used content-related metrics to characterize hosting infrastructure. Our key observation is that ISPs and hosting providers have collaborated to extensively replicate popular content into cellular networks. On the contrary, content of many non-popular domains and tracking domains tends to be available only in particular networks, resulting content monopoly by these networks.

Our findings provide evidences that the ISPs and CDNs in China follow the global trends of close collaboration [1,12]. However, care should be given when generalizing our findings to other countries. In addition, our dataset was collected from only one cellular ISP and the observation period is only two days. We are collecting DNS data from multiple ISPs with longer observation period, in the hope of providing an up-to-date picture of the content hosting infrastructure in China.

**Acknowledgments.** The authors would like to thank Rocky Chang for shepherding our paper and PAM reviewers for their feedback. 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 61502460).

## References

1. Akamai and AT&T renew global alliance (2017). https://goo.gl/b2uHMT
2. Public suffix list (2017). https://publicsuffix.org
3. Team Cymru (2017). http://www.team-cymru.org/
4. Tencent Cloud CDN (2017). https://www.qcloud.com/en/product/cdn.html
5. VNI mobile forecast highlights (2017). http://www.cisco.com/assets/sol/sp/vni/forecast_highlights_mobile
6. Adhikari, V.K., Guo, Y., Hao, F., Hilt, V., Zhang, Z.L., Varvello, M., Steiner, M.: Measurement study of Netflix, Hulu, and a tale of three CDNs. IEEE/ACM Trans. Networking **23**(6), 1984–1997 (2015)
7. Ager, B., Mühlbauer, W., Smaragdakis, G., Uhlig, S.: Web content cartography. In: Proceedings of the ACM IMC (2011)
8. Calder, M., Fan, X., Hu, Z., Katz-Bassett, E., Heidemann, J., Govindan, R.: Mapping the expansion of Google's serving infrastructure. In: Proceedings of the ACM IMC (2013)
9. Callahan, T., Allman, M., Rabinovich, M.: On modern DNS behavior and properties. SIGCOMM Comput. Commun. Rev. **43**(3), 7–15 (2013)
10. Chen, F., Sitaraman, R.K., Torres, M.: End-user mapping: Next generation request routing for content delivery. In: Proceedings of the ACM SIGCOMM (2015)
11. Chen, T., Ullah, I., Kaafar, M.A., Boreli, R.: Information leakage through mobile analytics services. In: Proceedings of the ACM HotMobile (2014)
12. Frank, B., Poese, I., Lin, Y., Smaragdakis, G., Feldmann, A., Maggs, B., Rake, J., Uhlig, S., Weber, R.: Pushing CDN-ISP collaboration to the limit. SIGCOMM Comput. Commun. Rev. **43**(3), 34–44 (2013)
13. Han, S., Jung, J., Wetherall, D.: A study of third-party tracking by mobile apps in the wild. Technical report, UW-CSE-12-03-01, March 2012
14. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: analysis and an algorithm. In: Proceedings of the NIPS (2001)

15. Pujol, E., Hohlfeld, O., Feldmann, A.: Annoyed users: ads and ad-block usage in the wild. In: Proceedings of the ACM IMC (2015)
16. Triukose, S., Wen, Z., Rabinovich, M.: Measuring a commercial content delivery network. In: Proceedings of the WWW (2011)
17. Xue, J., Choffnes, D., Wang, J.: CDNs meet CN: An empirical study of CDN deployments in China. In: IEEE Access (2017)