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

Zhenyu Li Donghui Yang

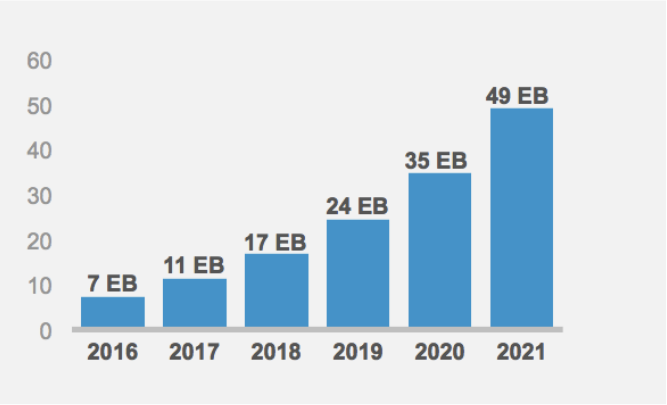
Zhenhua Li Chunjing Han Gaogang Xie

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY ,CAS

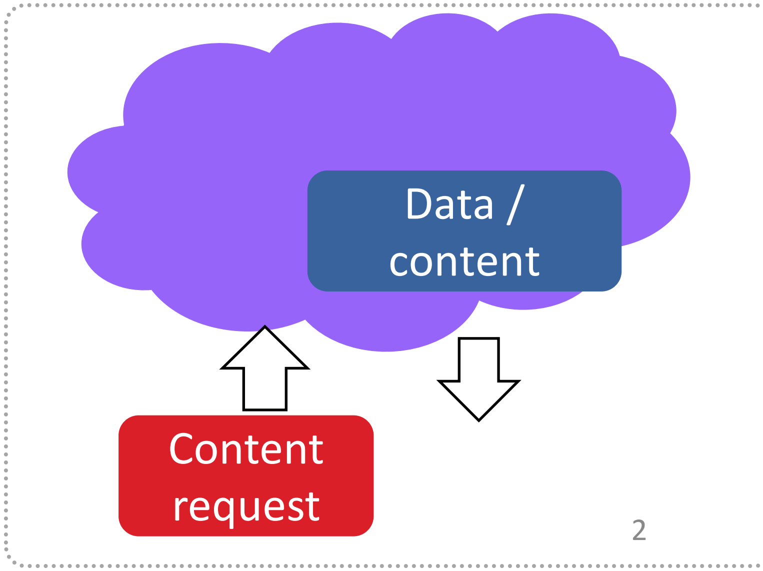
TSINGHUA UNIVERSITY

Continuous increase of mobile data

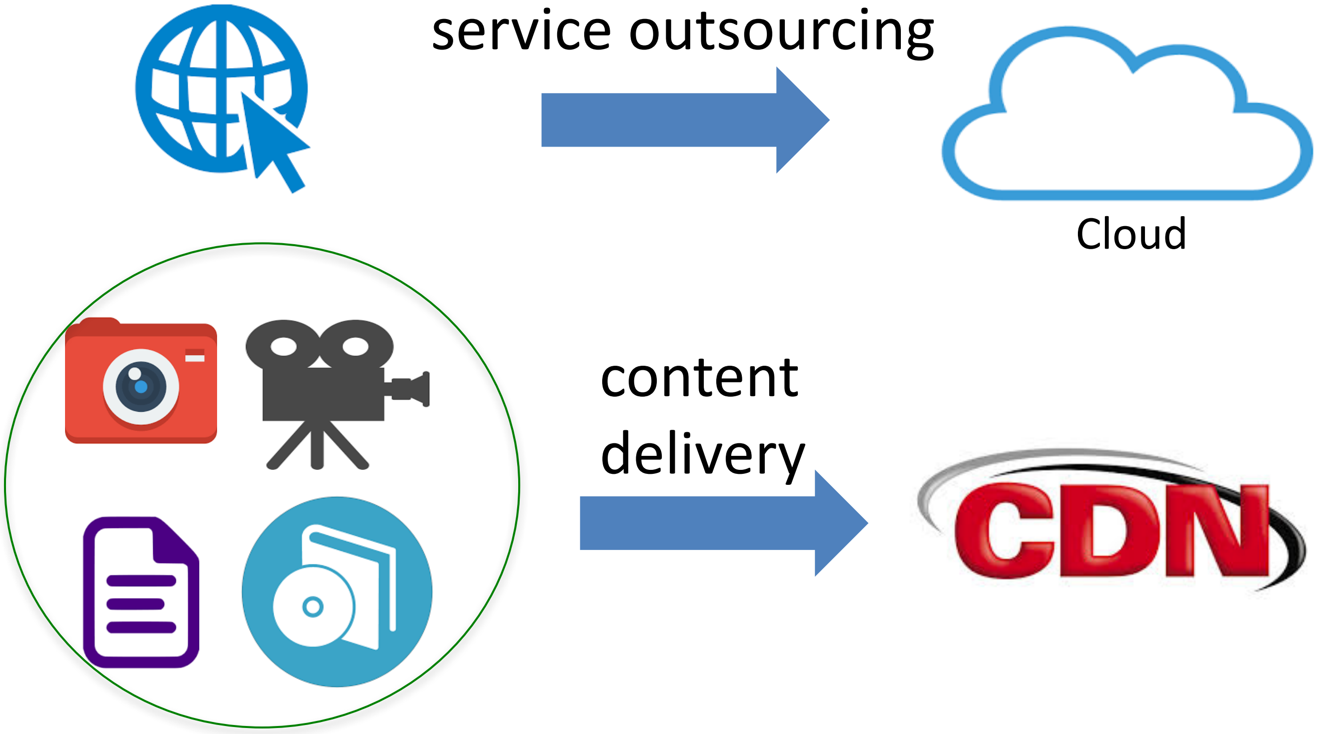
- CISCO projected: the mobile data will increase 7-fold by 2021
- The increase is largely due to rich content being available
  - Video traffic will be 78% by 2021
- The Internet is indeed a content network



Source: Cisco VNI Global Mobile Data Traf



Content hosting and delivery

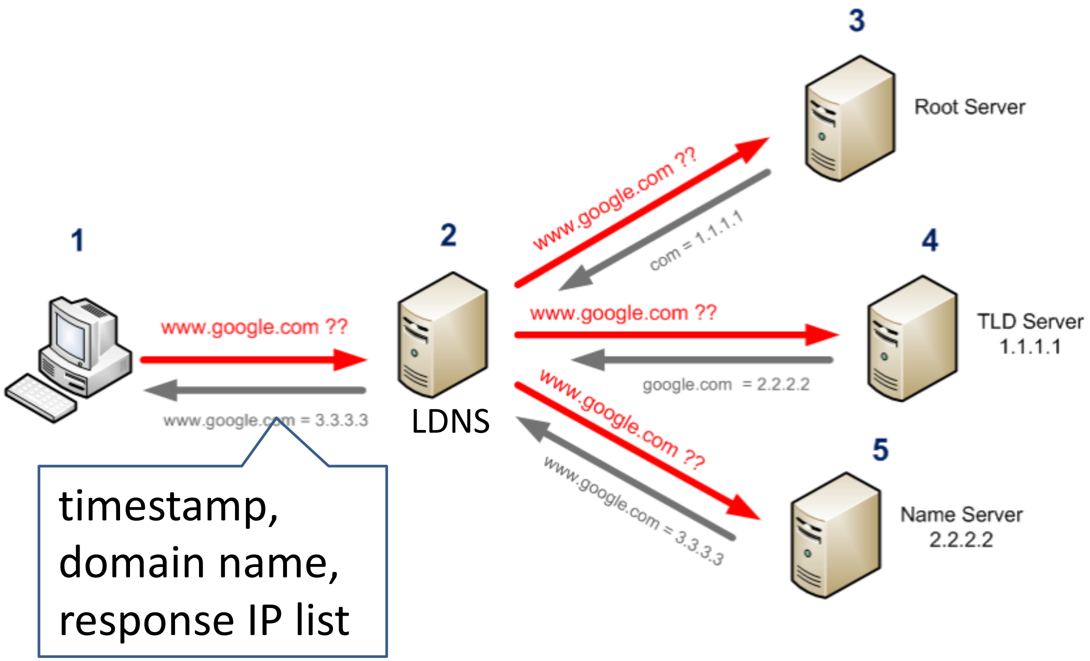


- Questions: network footprint? traffic locality?

Why China?

- The largest Internet in a single country
  - Over 800 million video users
- unique local regulations and network policies
  - Network is planned: very few ASes seen outside
  - The ICP regulation: Akamai could not deploy replica servers in mainland China
- Heavily censored visible web access. How about invisible web access (a.k.a trackers)?
  - Google is not accessible, but how about doubleclick?

Passive DNS Data



timestamp,
domain name,
response IP list

- Logs were collected from *all* recursive DNS resolvers of a major Chinese cellular ISP
  - 2 days, ~55 billion logs
- Response IP list: ~50% one single IP
  - The first one was taken as the one that the hostname was mapped to

Passive DNS Data

- Data Preprocessing
  - IP to ASN using Team Cymru
  - Aggregation IPs to /24 prefix
  - FQDN (Full Qualified Domain Names) to their second level domains (SLDs) to save analysis time
  - Invisible web access: identification of tracking domains using Easylist + EasylistChina.
- Ethical issues
  - No personal ID (client IP addresses are not available)
  - Such datasets are maintained by ISPs for maintenance purpose

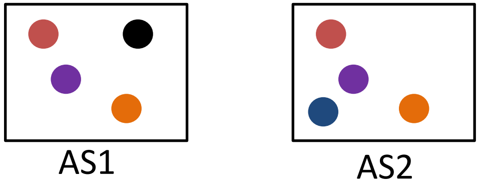
Metrics

- CDP: content delivery potential
  - Fraction of domains that an AS can serve
- CMI: content monopoly index
  - the extent to which an AS hosts content that others do not have

$$CMI_i = \frac{1}{|S_i|} \sum_{j \in S_i} \frac{1}{m_j}$$

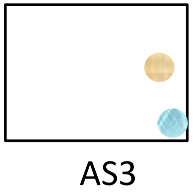
S_i: # of domains that can be served by this AS

m_j: # of ASes that can serve this domain



CDP=4/6

CMI=1/4*(1/2+1+1/2+1/2)=5/8



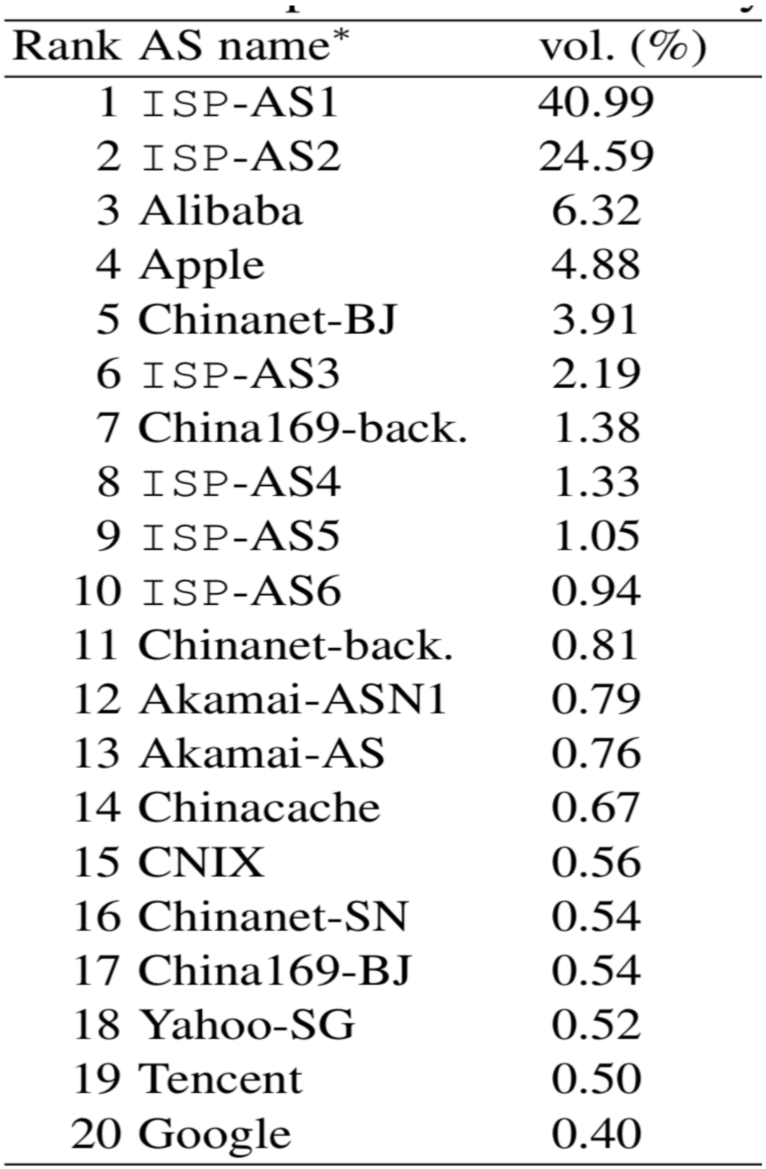
CDP=2/6

CMI=1/2*(1+1)=1

Ager, B., Muhlbauer, W., Smaragdakis, G., Uhlig, S.: Web content cartography, ACM IMC (2011)

Content Hosting Analysis

A look at the top ASes

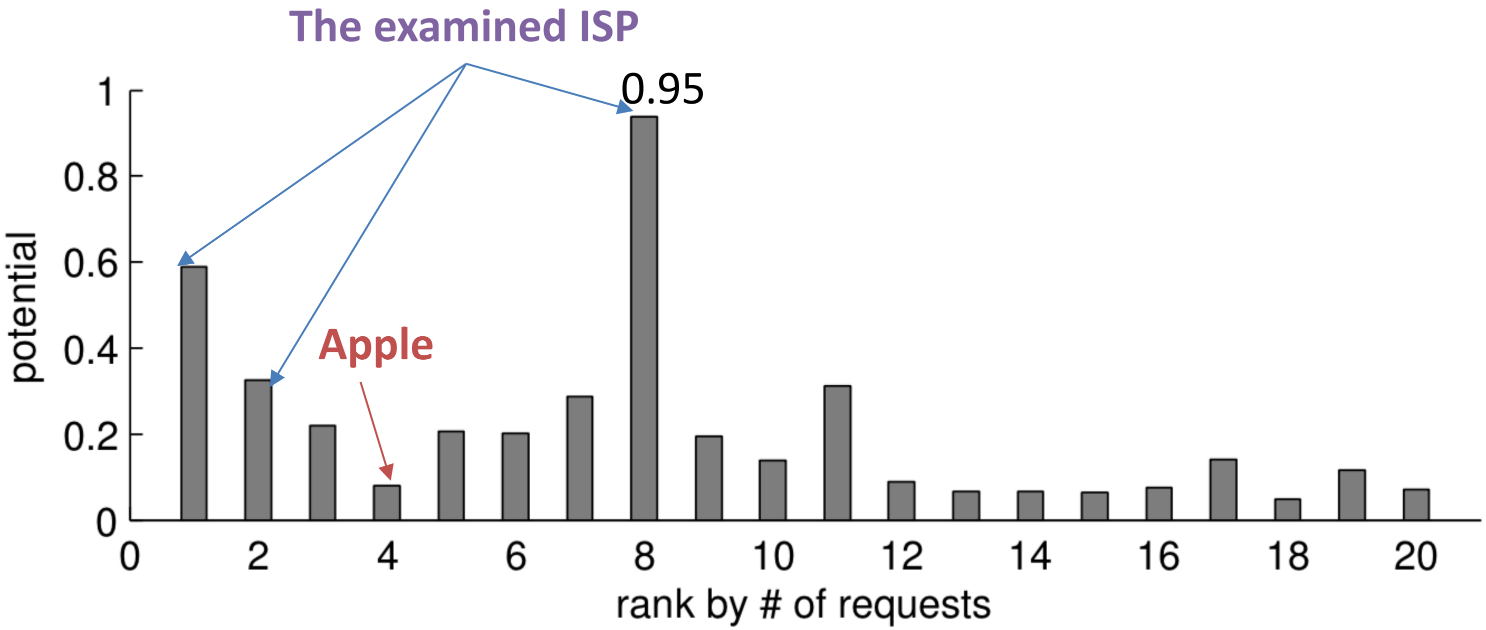


Rank	AS name*	vol. (%)
1	ISP-AS1	40.99
2	ISP-AS2	24.59
3	Alibaba	6.32
4	Apple	4.88
5	Chinanet-BJ	3.91
6	ISP-AS3	2.19
7	China169-back.	1.38
8	ISP-AS4	1.33
9	ISP-AS5	1.05
10	ISP-AS6	0.94
11	Chinanet-back.	0.81
12	Akamai-ASN1	0.79
13	Akamai-AS	0.76
14	Chinacache	0.67
15	CNIX	0.56
16	Chinanet-SN	0.54
17	China169-BJ	0.54
18	Yahoo-SG	0.52
19	Tencent	0.50
20	Google	0.40

- Observations
  - **Biased distribution**: top 2 accounting for 2/3
  - **ISPs dominate**: not CDNs /cloud
  - **Good locality**: ~70% queries resolved to IPs of the examined ISP
- Possible reasons
  - ISPs provide IDC or even servers to CDNs for content replication
  - Only ISPs and some giant enterprises have their own ASes in China

ISP is the one where we obtained data

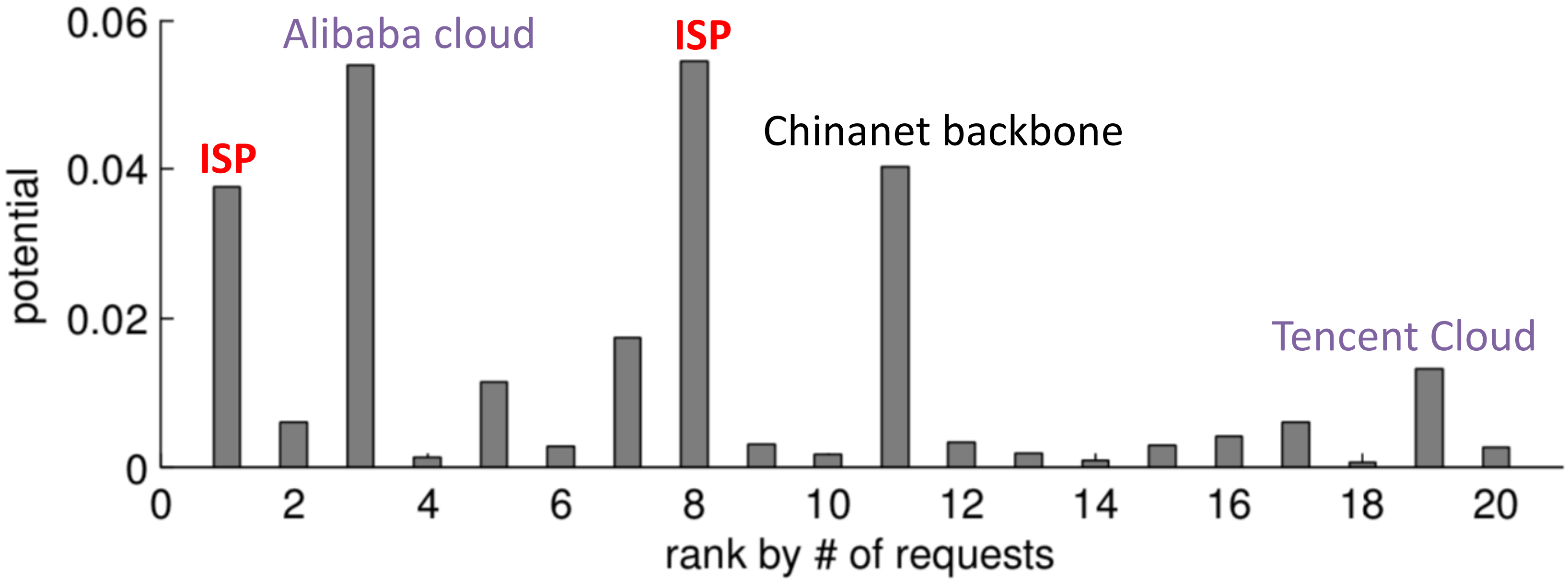
CDP of Top ASes: *popular* domains



(a) considering top 10k popular domains

- Popular content is well replicated into the examined cellular ISPs
  - Good for performance
- Apple AS: low CDP, but higher rank in terms of requests
  - Host of its own services that are frequently requested (by smart devices)

CDP of Top ASes: *all* domains

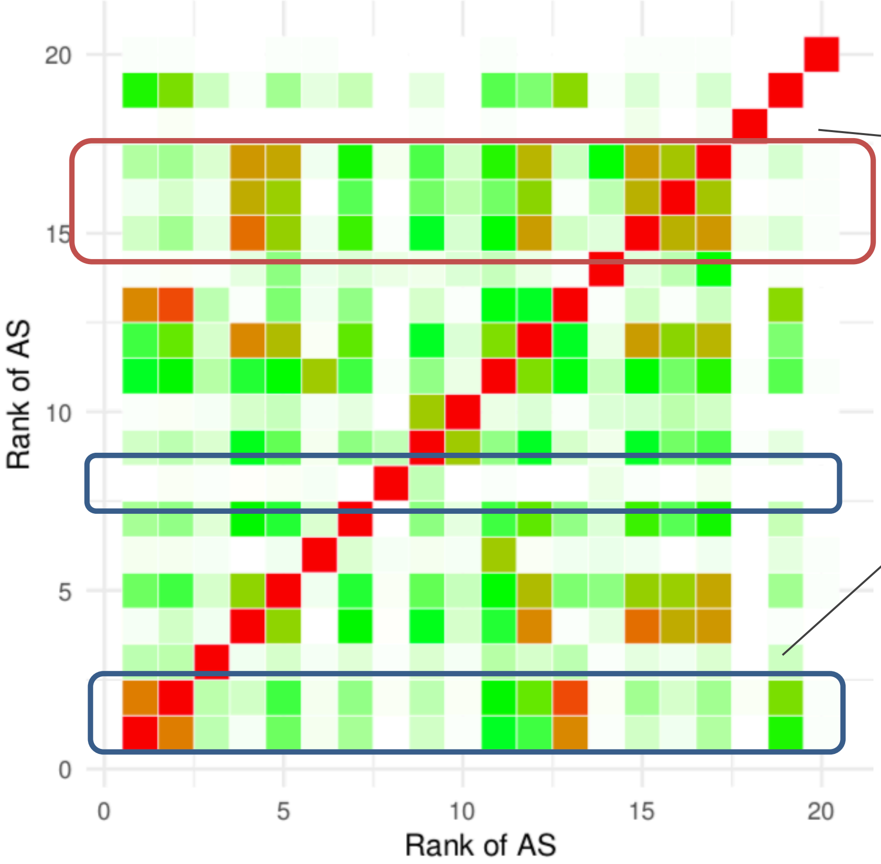


(b) considering all domains

- CDP values for all ASes are relatively low (<0.06)
  - Because of huge volume of non-popular domains
- **The rise of cloud**
  - Cloud platforms provide easy-to-use hosting services for individuals

Content similarity between ASes

- Cosine Similarity
  - One vector for each AS: an element is <*domain name, # of queries*>

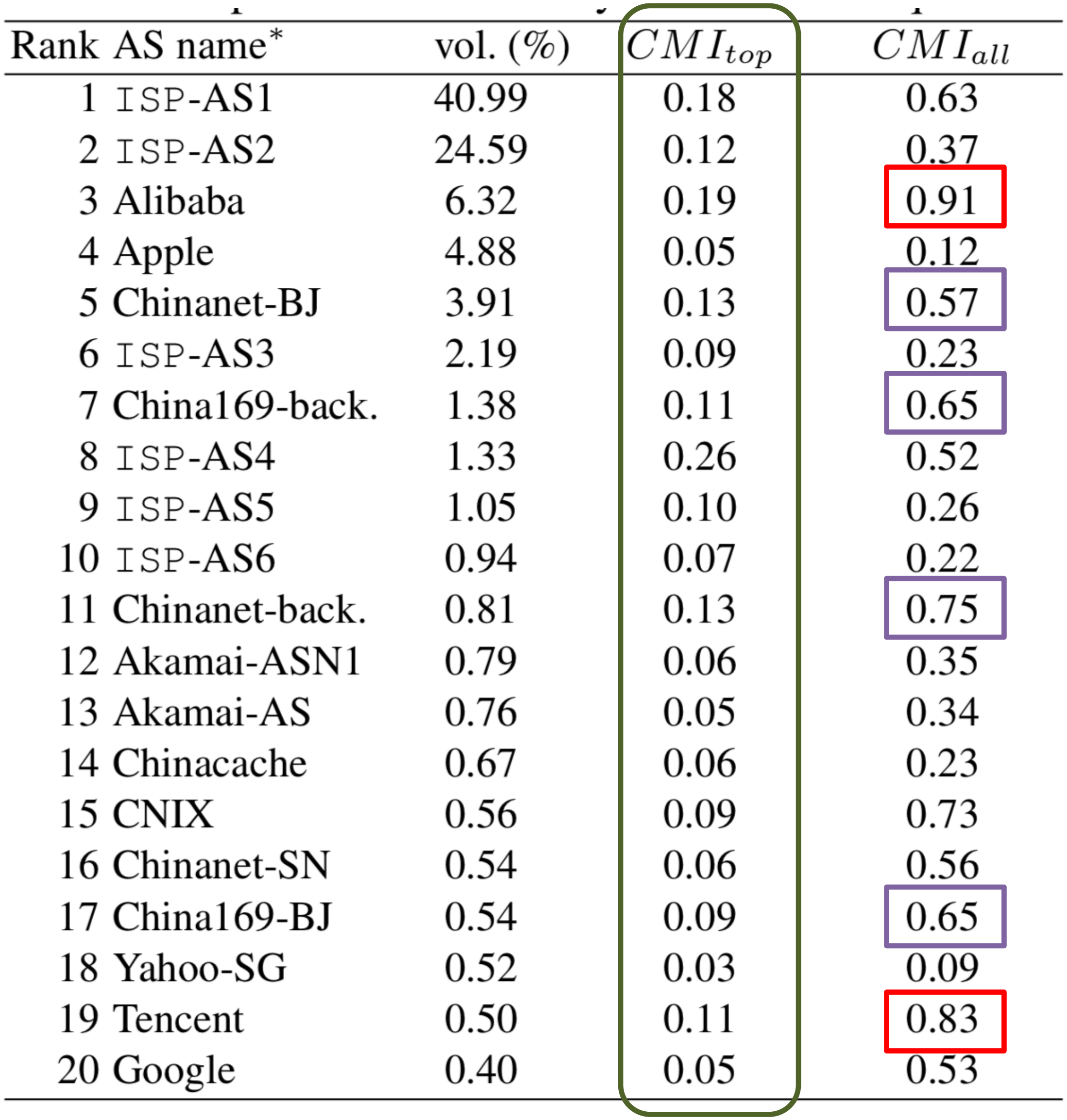


Chinanet:
- giant network

Cloud
- very low similarity (hosting non-popular sites)

The examined ISP
- Low similarity: high content availability
- Exception: Akamai ASes (#12 and #13)
  - ✓ caused by the domain aggregation?

CMI of Top ASes

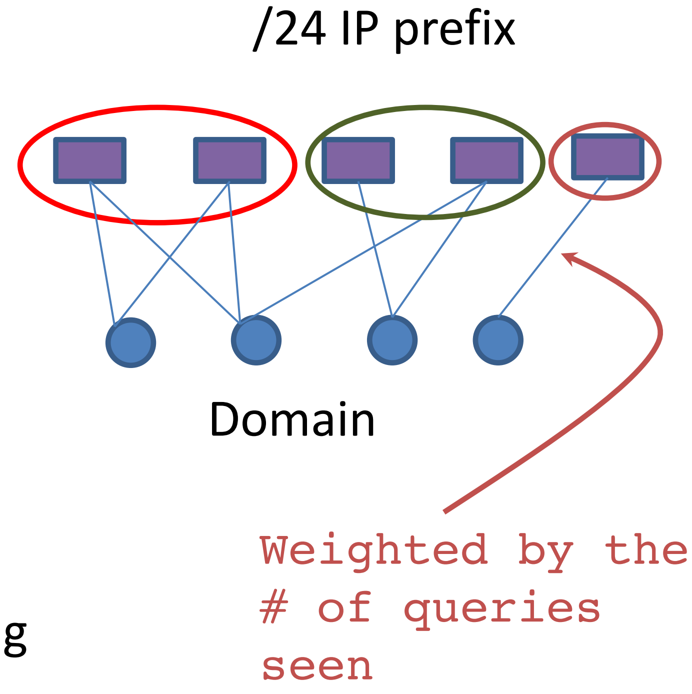


Rank	AS name*	vol. (%)	CMI_{top}	CMI_{all}
1	ISP-AS1	40.99	0.18	0.63
2	ISP-AS2	24.59	0.12	0.37
3	Alibaba	6.32	0.19	0.91
4	Apple	4.88	0.05	0.12
5	Chinanet-BJ	3.91	0.13	0.57
6	ISP-AS3	2.19	0.09	0.23
7	China169-back.	1.38	0.11	0.65
8	ISP-AS4	1.33	0.26	0.52
9	ISP-AS5	1.05	0.10	0.26
10	ISP-AS6	0.94	0.07	0.22
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

- Top 10k domains
  - low CMI values for all ASes
- All domains
  - Very high for the two cloud platforms
  - Moderately high for Chinanet's ASes

On Content Providers

- ***Questions:*** who deployed the replicas into the cellular ISP? How about their network footprints?
- Identification of major providers
  - WhoIs utility: not accurate
  - Last CNAME: not available
- spectrum clustering on the bipartite graph
  - Intuition: a provider uses a set of IP prefixes to serve same sites ➔ clustering IP prefixes

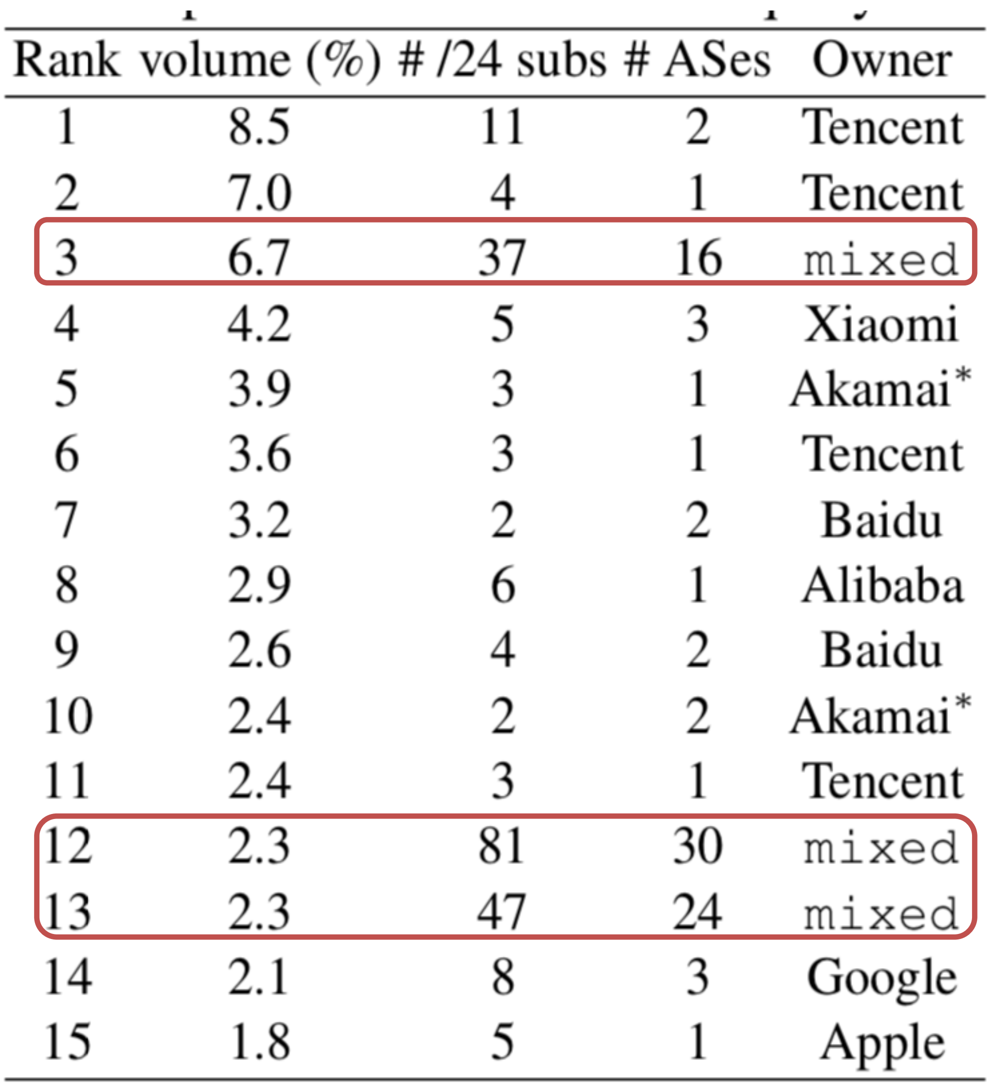


/24 IP prefix

Domain

Weighted by the # of queries seen

On Content Providers

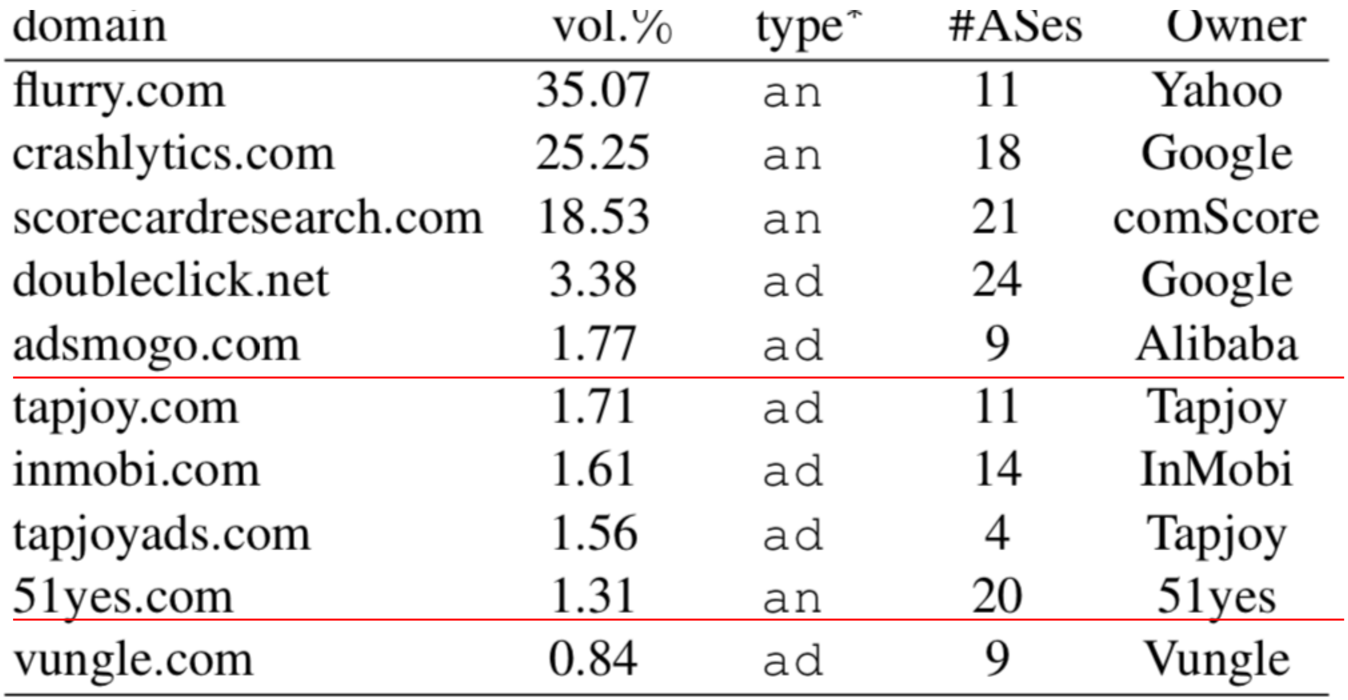


Rank	volume (%)	# /24 subs	# ASes	Owner
1	8.5	11	2	Tencent
2	7.0	4	1	Tencent
3	6.7	37	16	mixed
4	4.2	5	3	Xiaomi
5	3.9	3	1	Akamai*
6	3.6	3	1	Tencent
7	3.2	2	2	Baidu
8	2.9	6	1	Alibaba
9	2.6	4	2	Baidu
10	2.4	2	2	Akamai*
11	2.4	3	1	Tencent
12	2.3	81	30	mixed
13	2.3	47	24	mixed
14	2.1	8	3	Google
15	1.8	5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

- 15 out of 900+ clusters account for ~50% query volume
- Giant players in mobile Internet dominate, e.g. Baidu, Alibaba, and Tencent
- *Mixed*: may contain one or more CDNs
- 4 Tencent clusters provide 4 different services

(Invisible web) tracker hosting infrastructure

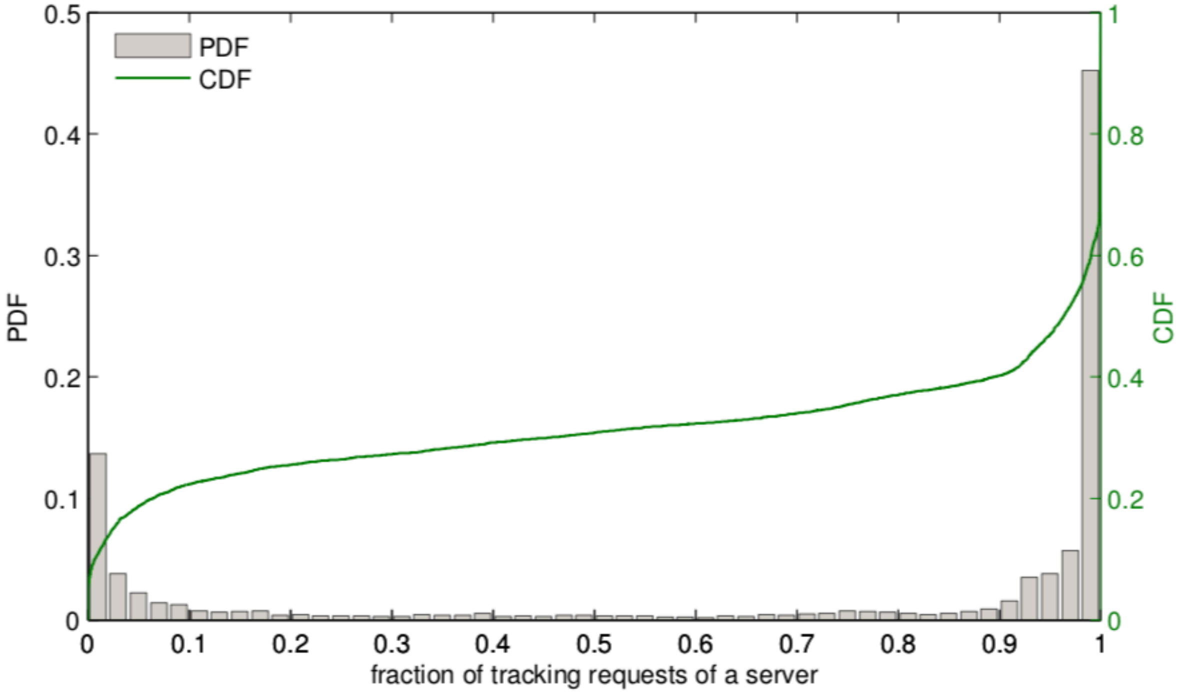
A look at trackers



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

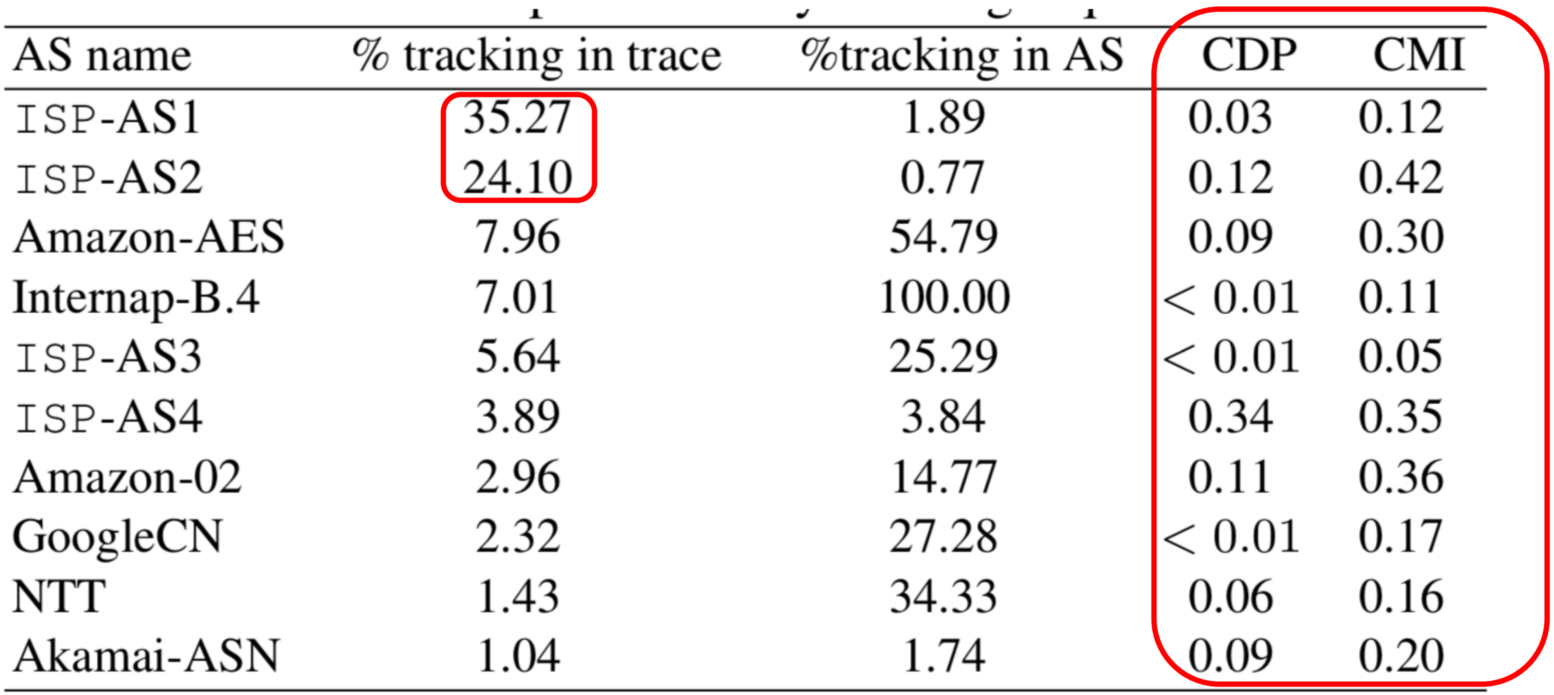
- Only 2 trackers are based in China
  - a potential cyber-security vulnerability
- Trackers are well-replicated into several networks

Tracking server



- Bimodal distribution: either seldom used by tracking service, or exclusively for trackers
  - Monitoring traffic goes to the servers that are exclusively for trackers could provide insights into trackers usage

Tracking from the net perspective



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

- Trackers have also been replicated into the examined cellular network, but still 20% goes abroad
- Low CDP, low CMI
  - trackers are replicated into several ASes, and each AS hosts very few

Summary

- One of the first studies on content hosting infrastructure in cellular network from the Chinese perspective
  - Finding 1: great traffic locality in the examined ISP network
  - Finding 2: raise of cloud platforms
  - Finding 3: most of the popular trackers are non-China based
  - Methodology: clustering over bipartite graph to infer providers

- On-going work
  - Data: One ISP ➔ all major ISPs, with CNAME being available
  - Vision: an up-to-date picture of the content hosting infrastructure in China

Thanks

http://fi.ict.ac.cn