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

Zhaohua Wang Zhenyu Li Minhui Xue Gareth Tyson

中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY ,CAS

THE UNIVERSITY of ADELAIDE

Queen Mary
University of London

PAM 2020

verizon media COMCAST Akamai

Table of contents

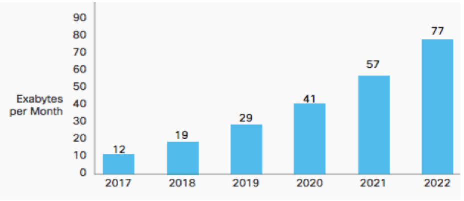
- Why study the mobile app tracking in China?
- Dataset and methodology
- How prevalent are ATSes?
- What's the community structure of ATSes?
- How are users impacted by ATSes?
- Conclusion

Table of contents

- Why study the mobile app tracking in China?
- Dataset and methodology
- How prevalent are ATSes?
- What’s the community structure of ATSes?
- How are users impacted by ATSes?
- Conclusion

Rising concerns about mobile app tracking

- CISCO projected: by 2022, there will be 1.5 mobile devices per capita and monthly global mobile data traffic will be 77 EB
- Many mobile apps are bundled with mobile Advertising and Tracking Services (ATSes) for various purposes
- Concerns:
  - Rich and sensitive user data
  - Beyond users' control



Source: Cisco VNI Mobile, 2019

GDPR

EU General Data Protection Regulation 25 May 2018

How about China?

- One of the fastest-growing countries in mobile data traffic
  - By 2022, the mobile data will reach 17.5 EB per month
- Unique local regulations and network policies
  - Many western services (e.g. Google, Facebook) are not accessible
- Chinese tracking market is poorly understood
  - Who are the major players?
  - What kind of mobile apps do trackers prefer?
  - . . . . . .

Source: marketingtochina.com, 2

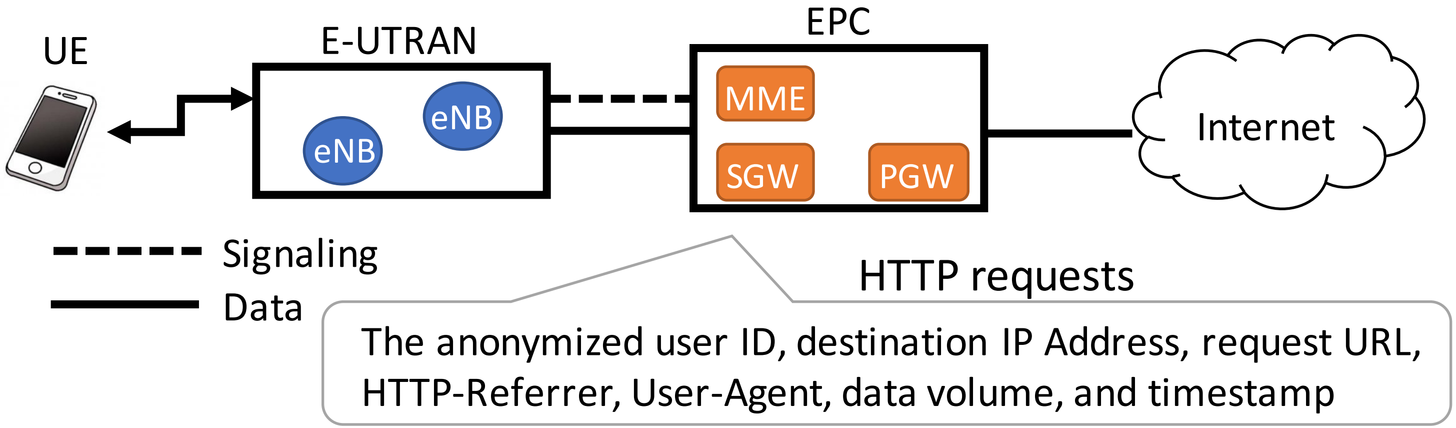
Table of contents

- Why study the mobile app tracking in China?
- Dataset and methodology
- How prevalent are ATSes?
- What's the community structure of ATSes?
- How are users impacted by ATSes?
- Conclusion

PAM 2020

verizon media COMCAST Akamai

Mobile traffic Data



The anonymized user ID, destination IP Address, request URL, HTTP-Referrer, User-Agent, data volume, and timestamp

- User access logs collected from a major 4G cellular ISP
  - ~2.8 billion logs of ~3.5 million users in a major city of China
  - Identify 1,812 mobile apps, 12% of logs remain unattributed
- Ethical issues
  - Dataset is kept in the ISP's data center and sensitive user IDs are anonymized

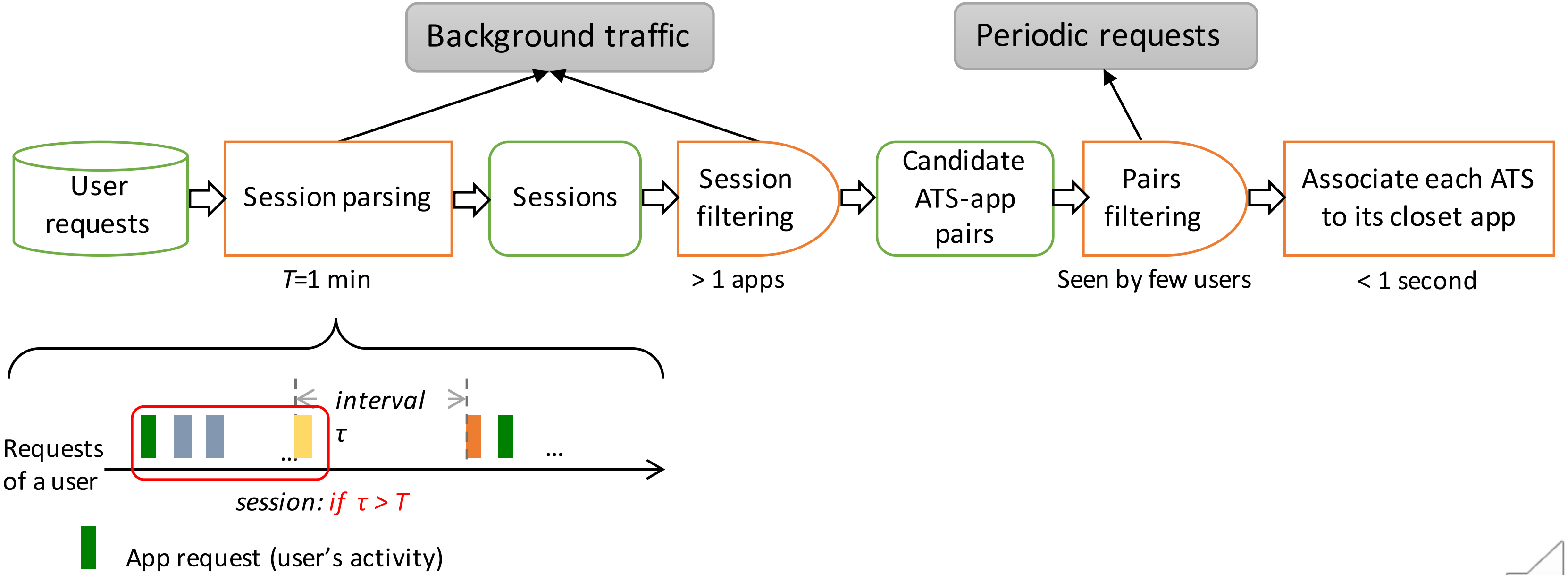
verizon media COMCAST Akamai

Data processing

- Identify ATS domains
  - 4 ATS-specific lists: AdBlock-Plus (the easylist, easyprivacy lists, and easylist China) and hpHosts (the ATS lists)
  - Apply the rules to both the URL and HTTP-Referrer
  - 260M HTTP requests (9.2%) are from ATS domains, 24,985 FQDNs and 8,773 SLDs
- Associate ATS domains to apps :
  - We focus on the top-500 apps that account for 99% of traffic
  - Heuristic approach: associate an ATS request to the closest app's request that precedes it
  - Intuition: ATS's requests should happen at a time close to the app's access (<1s)
  - Problem
    - background traffic from other mobile apps
    - periodic requests issued by some trackers

Data processing

- Associate ATS domains to apps : heuristic approach



Background traffic

Periodic requests

User requests

Session parsing

T=1 min

Sessions

Session filtering

> 1 apps

Candidate ATS-app pairs

Pairs filtering

Seen by few users

Associate each ATS to its closet app

< 1 second

interval *τ*

Requests of a user

...

...

session: if τ > T

🟩 App request (user's activity)

PAM 2020

Limitations

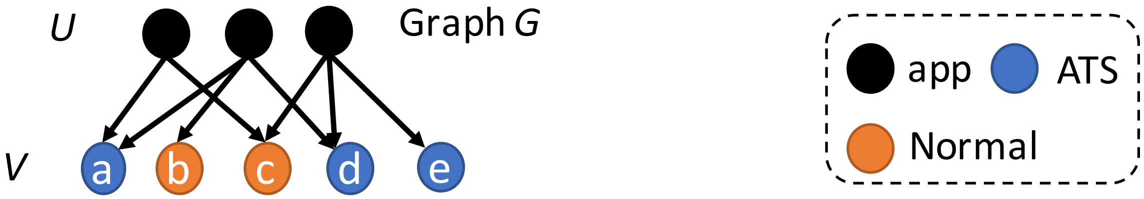
	The 4 ATS lists used for ATS identification	The heuristic method for the ATS-to-app association
limitation	They may not fully cover the current ATSes in mobile networks in China	It may not fully capture the up-to-date ATSes of individual mobile apps
Observation & Validation	Recognized ATS domains are in line with the Chinese mobile ecosystem	Manually test existing ATS domains for the top 10 most popular apps Association accuracy of F1-score 0.75 (precision: 0.7, recall: 0.82)

Table of contents

- Why study the mobile app tracking in China?
- Dataset and methodology
- How prevalent are ATSes?
- What’s the community structure of ATSes?
- How are users impacted by ATSes?
- Conclusion

Metrics

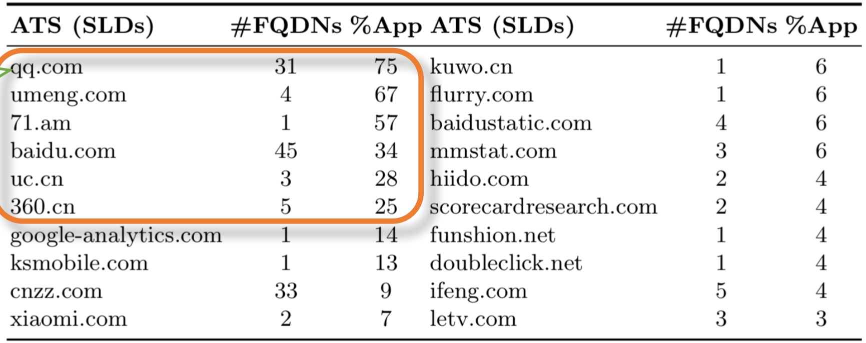
- Model a bipartite graph $G = (U, V, E)$
  - Based on the domains (FQDNs) accessed within an app
  - U: mobile apps
  - V: the ATS and normal visited domains
- G reveals the connections between ATS domains and mobile apps



Presence of ATSes

- ATSes are widely used by mobile apps
  - 6 ATSes for FQDNs (4 ATSes for SLDs) per app in median
- Cross-app tracking of users
  - Over 30% of ATSes appear in at least 2 apps
- China's tracking ecosystem is dominated by key domestic trackers

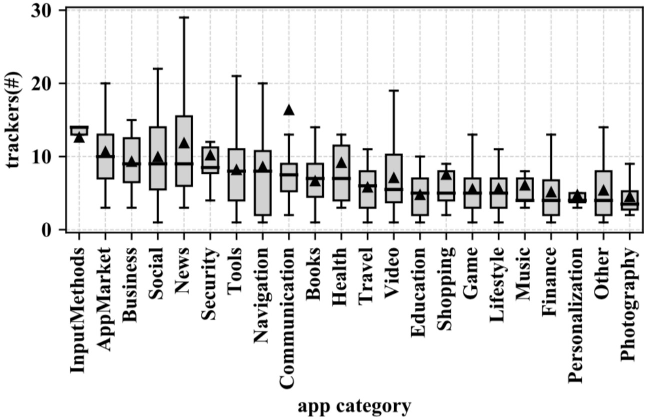
pingma.qq.com,
zxcv.3g.qq.com,
omgmta.qq.com,
sngmta.qq.com,
mi.gdt.qq.com …



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

The top 20 ATS domains (SLDs) measured by the number of apps they are used by

App’s ATS usage



The distribution of tracker domains (FQDNs) by different app categories, each box is ranked in descending order by the median

- Apps are grouped into 23 categories based on their functionalities
- Trackers tend to be active in some app categories, for example
  - *InputMethods* has the most trackers (13 ATSes) per app
  - *Communication* has the highest mean value of 16 ATSes per app
  - Top 5% of *News* apps use over 26 ATSes

PAM 2020

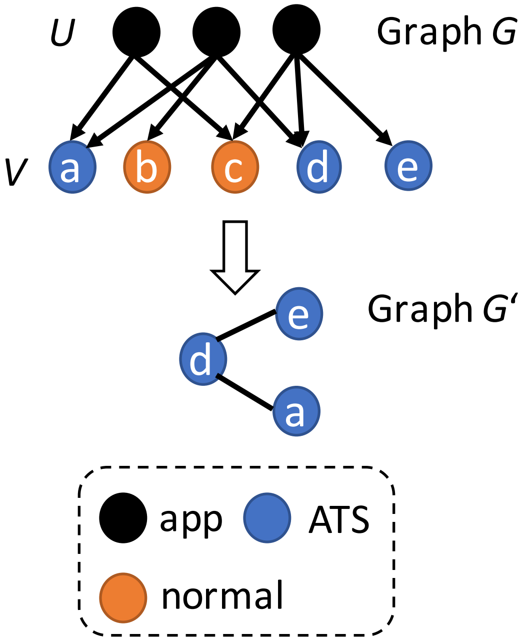
verizon media COMCAST Akamai

Table of contents

- Why study the mobile app tracking in China?
- Dataset and methodology
- How prevalent are ATSes?
- What’s the community structure of ATSes?
- How are users impacted by ATSes?
- Conclusion

Metrics

- Model a bipartite graph $G = (U, V, E)$
  - Based on the domains (FQDNs) accessed within an app
  - U: mobile apps
  - V: the ATS and normal visited domains
- 1-mode ATS-projection graph $G' = (V', E')$
  - Create from the largest connected component in G
  - V' : the ATS domains in V
  - E' : if two vertices share a common neighbor (app) in G
- G' captures the co-location of multiple ATSes used within individual apps

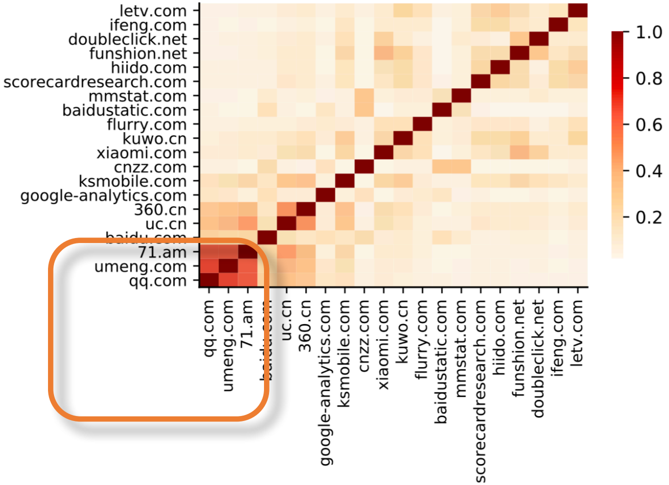


The structure of graph G'

- Identify two types of trackers with the degree centrality of ATSes in G'
  - Popular ATS (>0.2) and non-popular ATS
  - Popular ATSes are present more pervasively among apps
- Popular trackers are densely connected with the non-popular ones
  - High global clustering coefficient of G', but low coefficients for popular trackers
- Non-popular trackers form 56 local communities
  - Clauset-Newman-Moore greedy method for inferring community structure
  - 10 communities and 46 isolated components

Co-location of ATSes

- The popular trackers tend to co-locate in the same apps with each other
  - qq.com (Tencent), umeng.com (Alibaba), 71.am (Baidu)

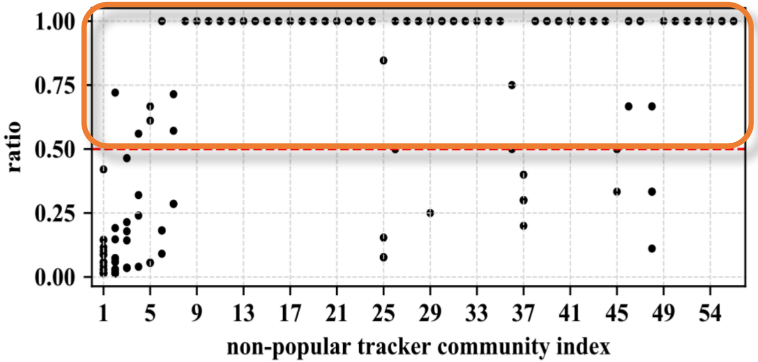


The co-occurrence probability distribution of the top 20 ATSes (SLDs), Quantified by the Jaccard Similarity Coefficient and ranked by the popularity

PAM 2020

Specialization of ATSes

- The local community of non-popular trackers is dedicated to specific app categories
  - Tracker Specialization Index (TSI): $\frac{|U(a) \cap U(b)|}{|U(a)|}$, $U(a)$ and $U(b)$ are sets of trackers in the local community a and app category b



TSI distribution of non-popular tracker communities

Non-popular ATS local communities tend to be specialized in only one or two app categories with TSI ≥ 0.5

We observe that they provide specialized tracking services relevant to particular apps, e.g. *education apps*

Table of contents

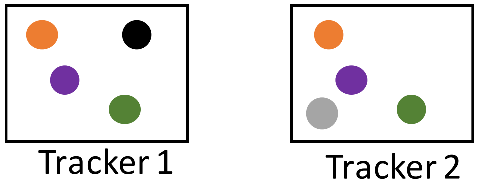
- Why study the mobile app tracking in China?
- Dataset and methodology
- How prevalent are ATSes?
- What's the community structure of ATSes?
- How are users impacted by ATSes?
- Conclusion

ATS Monopolies

- To test whether ATSes have a monopoly on certain users' data
- UTP : user tracking potential
  - Fraction of users that a tracker can track
- TMI: tracking monopoly index
  - The extent to which a tracker reaches users that others do not have
  - $TMI_i = \frac{1}{|S_i|}\sum_{j\in S_i}\frac{1}{|m_j|}$,

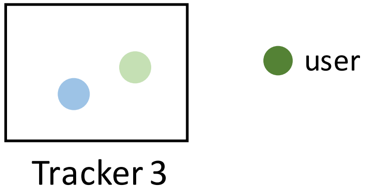
S_i : the set of users that can be reached by tracker i

m_j : the number of trackers that can reach user j



UTP=4/7

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

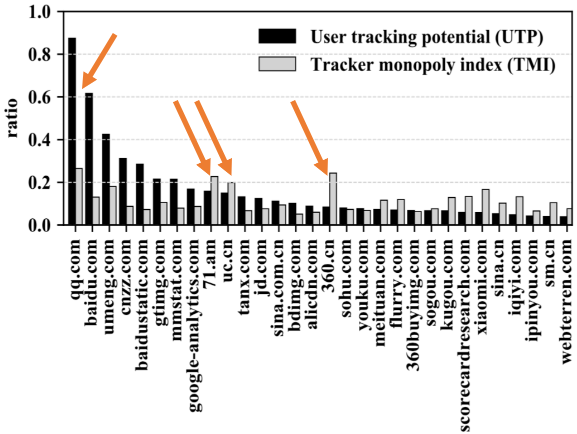


UTP=2/7

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

ATS Monopolies

- High penetration of the tech giants, for example
  - *qq.com* (Tencent) holds a high UTP (over 0.8) and TMI (about 0.3) metrics
  - *71.am* (Baidu), *uc.cn* (Alibaba), *360.cn* (360 Security) track under 20% of users, but have relatively high TMIs (about 0.3)



UTP and TMI distribution of the top 30 tracker domains (SLDs), ranked in descending order by the UTP values

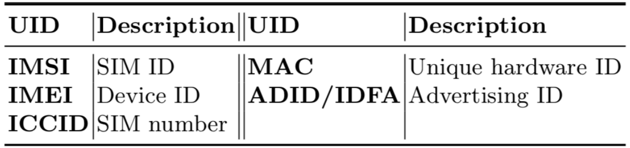
PAM 2020

verizon media COMCAST Akamai

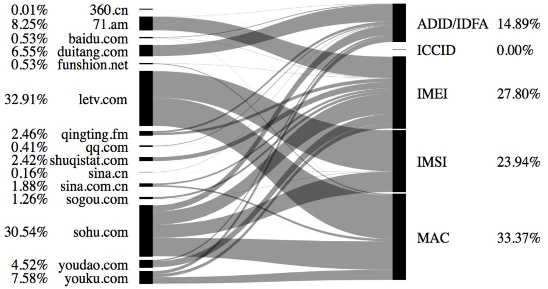
ATS traffic consumption & PII leakage

- ATS *v.s.* app traffic volumes
  - 5% of users send over 10% of app traffic to trackers
  - iOS users tend to send less data to trackers than Android users
- PII leakage and regional destination
  - Detect the common UIDs in URLs
  - 10% of users send their PII to trackers
  - IMEI, IMSI, and MAC are equally likely to be collected by trackers
  - 90% of PII tracking flows are inside mainland China

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		



Tracking domains (SLDs) that collect PII

Table of contents

- Why study the mobile app tracking in China?
- Dataset and methodology
- How prevalent are ATSes?
- What's the community structure of ATSes?
- How are users impacted by ATSes?
- Conclusion

Summary of contributions

- The first study on China's mobile app tracking from a 4G cellular network
- Key finding 1: a distinctive mobile tracking market dominated by several top popular domestic trackers
  - Prevalent cross-tracking of users and Prominent tracking in some types of apps
- Key finding 2: a well-connected tracking community
  - Popular trackers regularly co-occur with non-popular ones
  - Non-popular trackers cluster into local communities, each community tends to track a particular relevant type of apps
- Key finding 3: most of the PII data are confined to China
  - 10% of users send PII data to trackers

Thank you

Any question?

wangzhaohua@ict.ac.cn