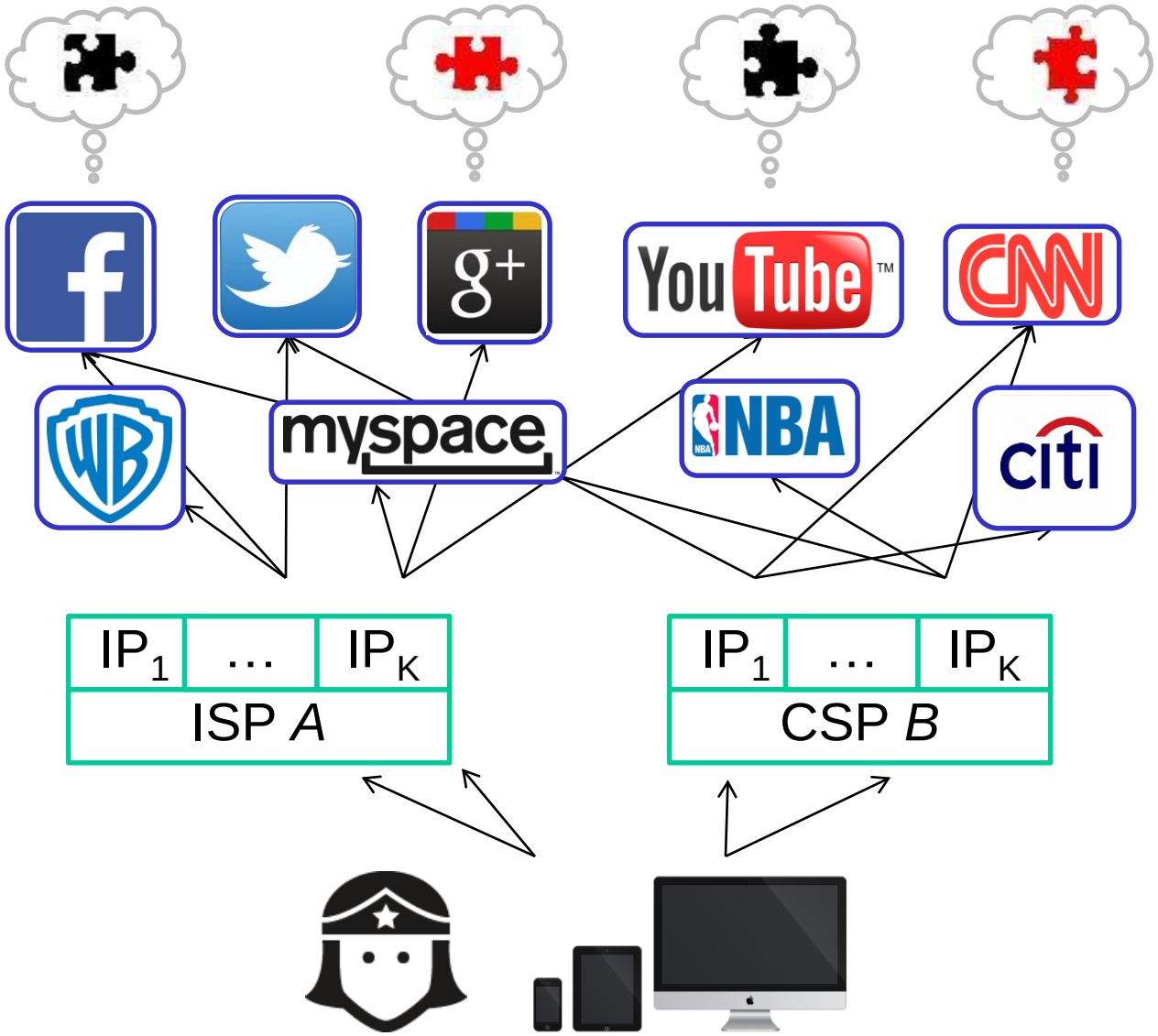
Mosaic: Quantifying Privacy Leakage in Mobile Networks

Ning Xia (Northwestern University)
Han Hee Song (Narus Inc.)
Yong Liao (Narus Inc.)
Marios Iliofotou (Narus Inc.)
Antonio Nucci (Narus Inc.)
Zhi-Li Zhang
(University of Minnesota)
Aleksandar Kuzmanovic
(Northwestern University)

Scenario



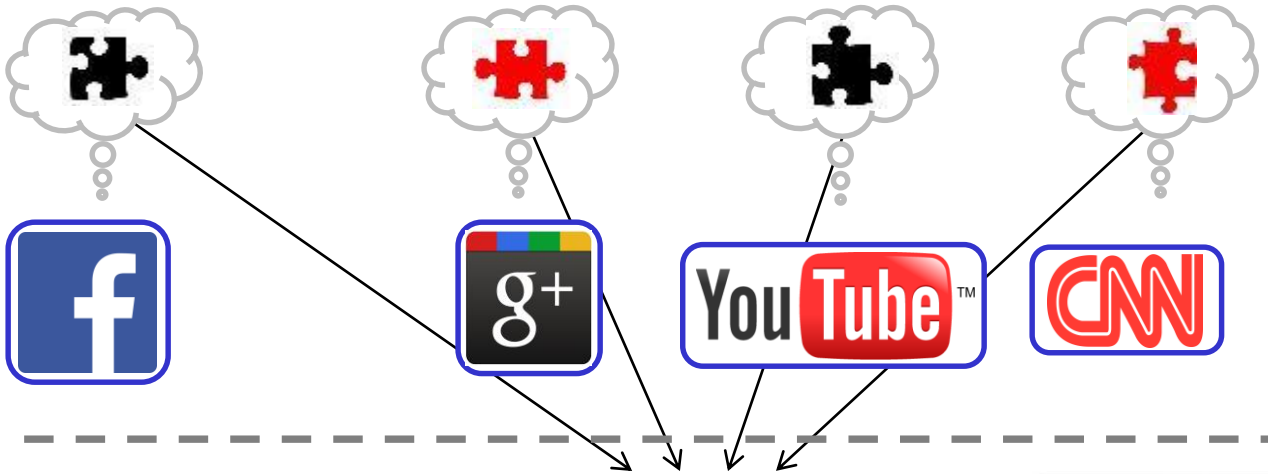
Different information

Different services

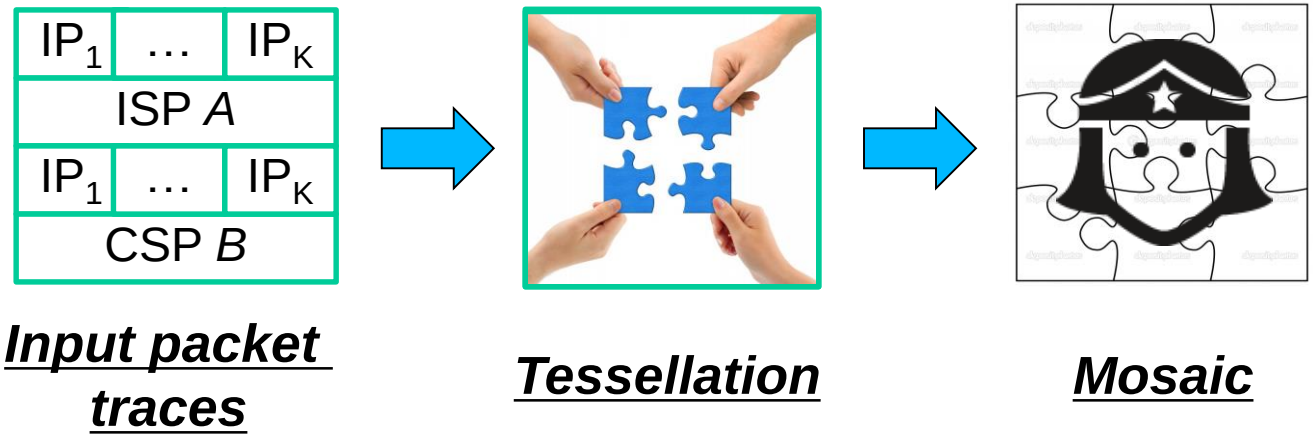
Dynamic IP, CSP/ISP

Different devices

Problem



Other research work



Input packet traces

Tessellation

Mosaic

We are here!

How much private information can be obtained and **expanded** about end users by monitoring network traffic?

Motivation

I will know everything about everyone!

IP_1	…	IP_K
ISP *A*		
IP_1	…	IP_K
CSP *B*		

Agencies

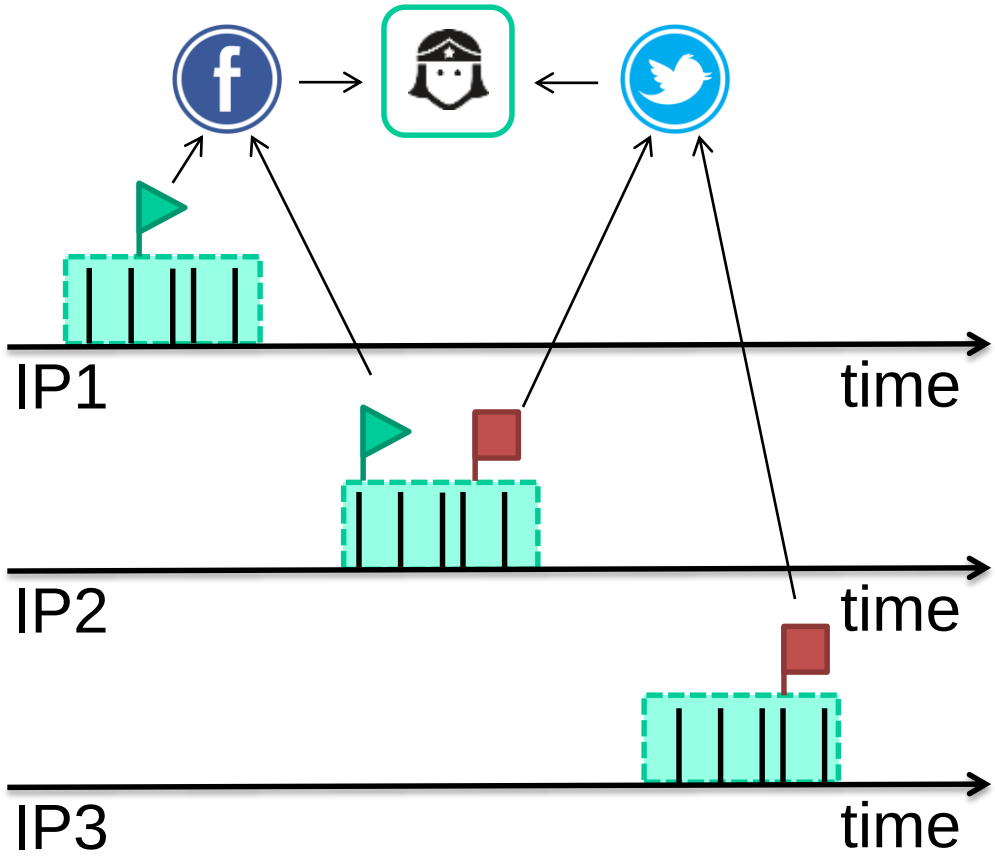
Bad guys

Mobile Traffic:

- Relevant: more personal information
- Challenging: frequent IP changes

Challenges

- How to track users when they hop over different IPs?



Sessions: |||||

Flows(5-tuple) are grouped into sessions

Traffic Markers:

Identifiers in the traffic that can be used to differentiate users

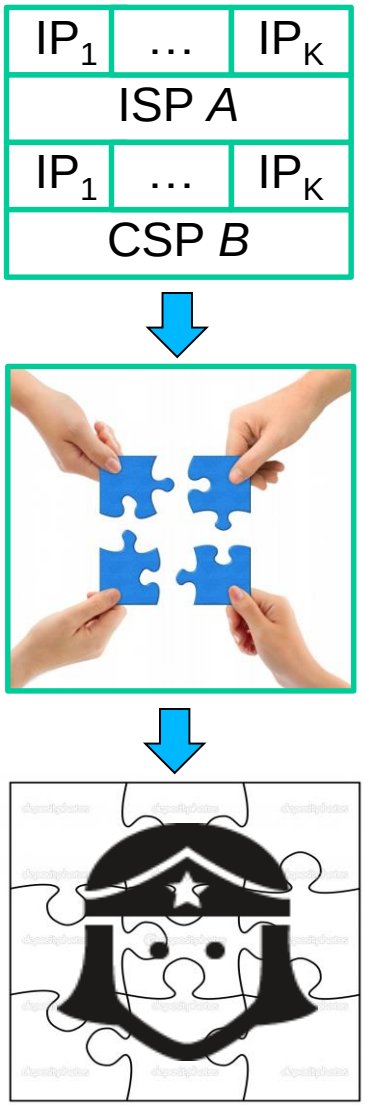
With ***Traffic Markers***, it is possible to connect the users' true identities to their sessions.

Datasets

Dataset	Source	Description
3h-Dataset	CSP-A	Complete payload
9h-Dataset	CSP-A	Only HTTP headers
Ground Truth Dataset	CSP-B	Payload & RADUIS info.

- *3h-Dataset*: main dataset for most experiments
- *9h-Dataset*: for quantifying privacy leakage
- *Ground Truth Dataset*: for evaluation of session attribution
  - RADIUS: provide session owners

Methodology Overview



Tessellation

Mapping from sessions to users

- Traffic attribution
  - Via traffic markers
  - Via activity fingerprinting
- Mosaic construction
  - Network data analysis
  - Web crawling

Combine information from both **network data** and **OSN profiles** to infer the user mosaic.

Traffic Attribution via Traffic Markers

Traffic Markers:

- Identifiers in the traffic to differentiate users
- Key/value pairs from HTTP header
- User IDs, device IDs or sessions IDs

Domain	Keywords	Category	Source
osn1.com	c_user=<OSN1_ID>	OSN User ID	Cookies
osn2.com	oauth_token=<OSN2_ID>-##	OSN User ID	HTTP header
admob.com	X-Admob-ISU	Advertising	HTTP header
pandora.com	user_id	User ID	Cookies
google.com	sid	Session ID	Cookies

How can we select and evaluate
traffic markers from network data?

Traffic Attribution via Traffic Markers

OSN IDs as Anchors:

- The most popular user identifiers among all services
- Linked to user public profiles

OSN	Source	Session Coverage
OSN1 ID	HTTP URL and cookies	1.3%
OSN2 ID	HTTP header	1.0%

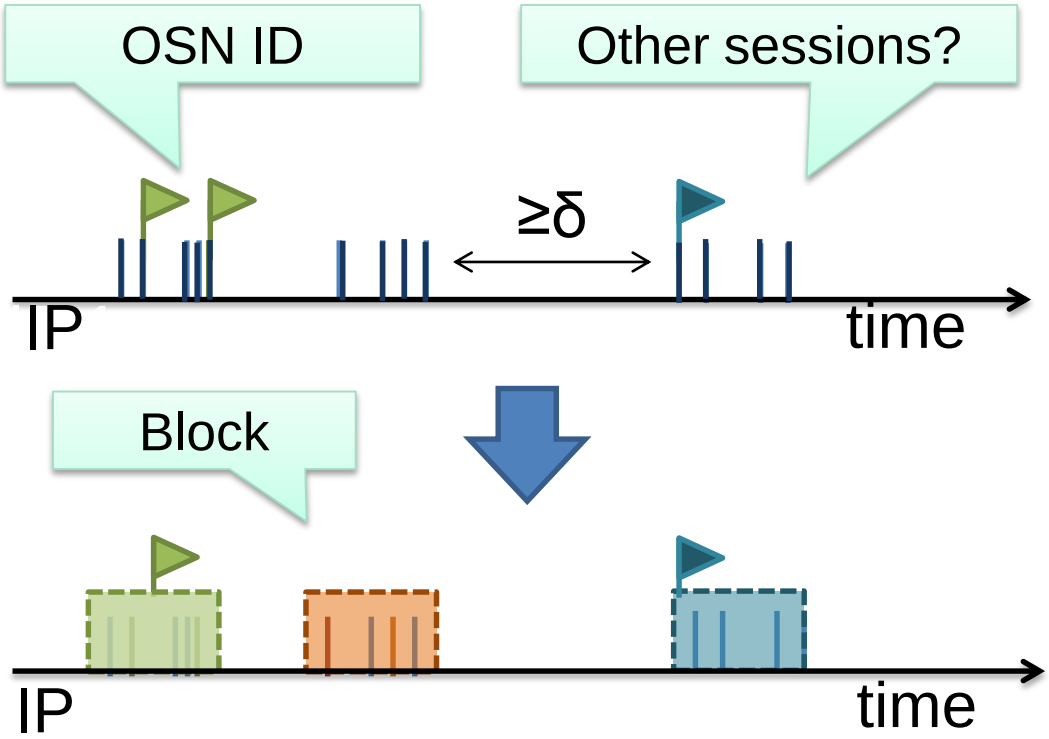
Top 2 OSN providers from North America

Only 2.3% sessions contain OSN IDs

OSN IDs can be used as anchors, but their coverage on sessions is too small

Traffic Attribution via Traffic Markers

Block Generation: Group Sessions into Blocks



Session interval δ

- Depends on the CSP
- δ=60 seconds in our study

Block

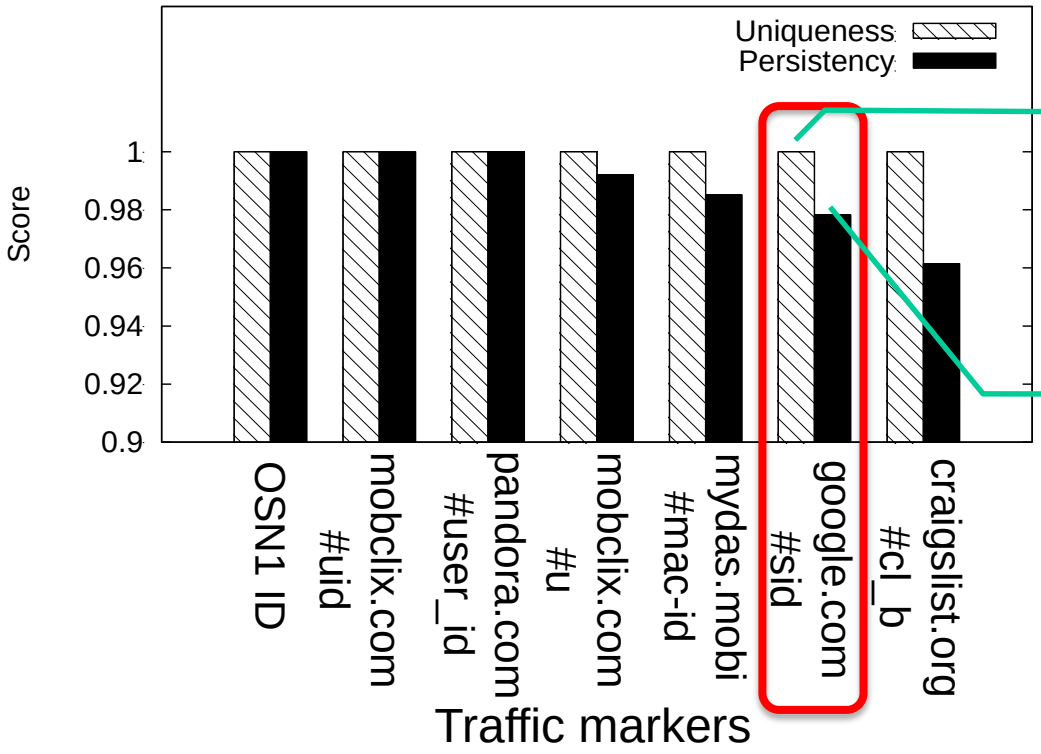
- Session group on the same IP within a short period of time
- Traffic markers shared by the same block

99K session blocks generated from the 12M sessions

Traffic Attribution via Traffic Markers

Culling the Traffic Markers: OSN IDs are not enough

- Uniqueness: Can the traffic marker differentiate between users?
- Persistency: How long does a traffic marker remain the same?



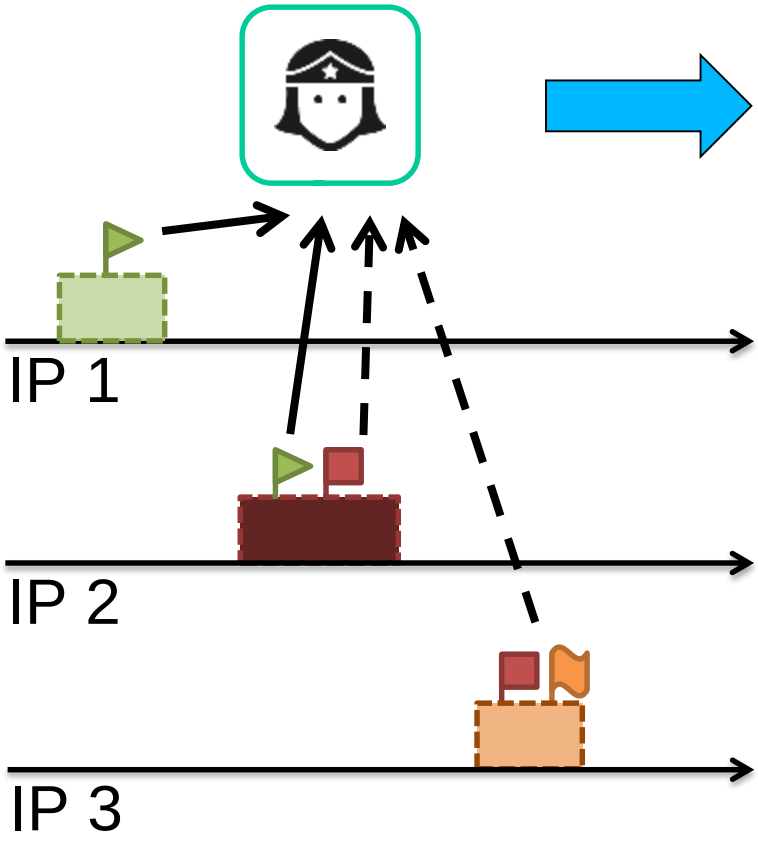
Uniqueness = 1
No two users will share the same google.com#sid value

Persistency ~= 1
The value of Google.com#sid remains the same for the same user nearly all the observation duration

We pick 625 traffic markers with uniqueness = 1, persistency > 0.9

Traffic Attribution via Traffic Markers

Traffic Attribution: Connecting the Dots



Tessellation User T_i

( )

Same OSN ID

Same traffic marker

Traffic markers are the key in attributing sessions to the same user over different IP addresses

Traffic Attribution via Activity Fingerprinting

- What if a session block has no traffic markers?

Assumption (Activity Fingerprinting):

- Users can be identified from the DNS names of their favorite services

DNS names:

- Extracted 54,000 distinct DNS names
- Classified into 21 classes

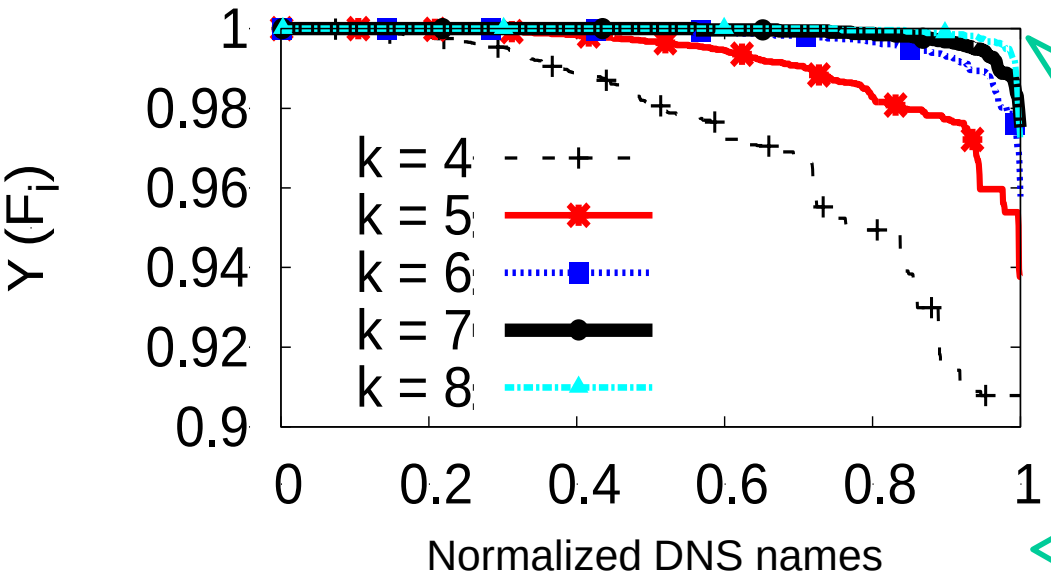
Activity Fingerprinting:

- Favorite (top-*k*) DNS names as the user's "fingerprint"

Service classes	Service providers
Search	bing, google, yahoo
Chat	skype, mtalk.googl.com
Dating	plentyoffish, date
E-commerce	amazon, ebay
Email	google, hotmail, yahoo
News	msnbc, ew, cnn
Picture	Flickr, picasa
…	…

Traffic Attribution via Activity Fingerprinting

- F_i : Top *k* DNS names from user as “activity fingerprint”
- $\Psi(F_i)$: Uniqueness of the fingerprint



Y-axis:
closer to 1, more distinct the fingerprint is

X-axis:
normalized by the total number of DNS names

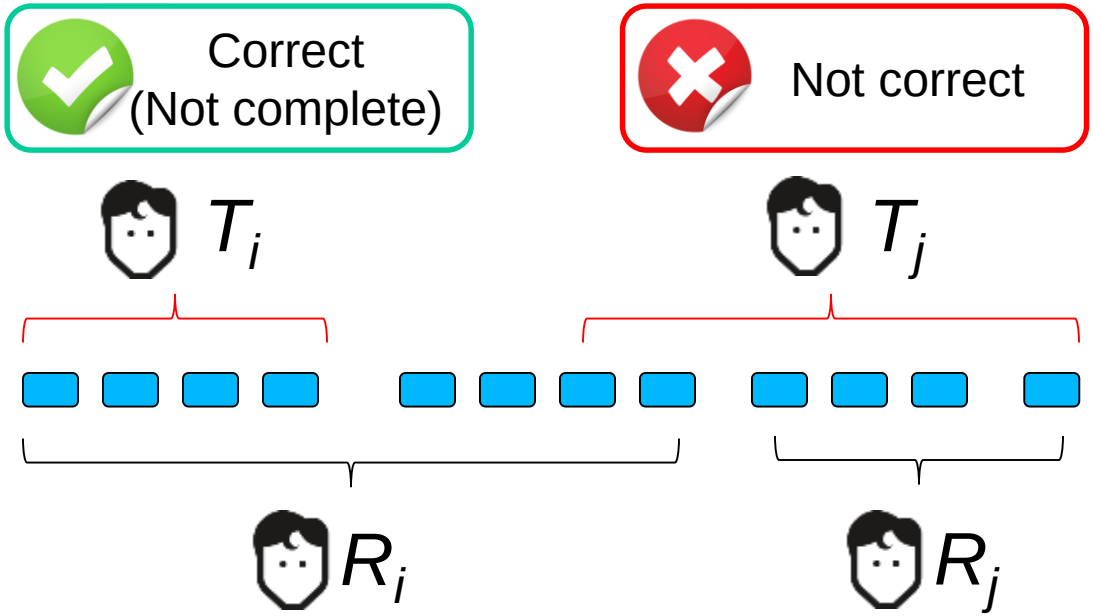
Mobile users can be identified
by the DNS names from their preferred services

Traffic Attribution Evaluation

Session

R_i RADIUS user (Ground Truth)

T_i Tessellation user (Correct?)

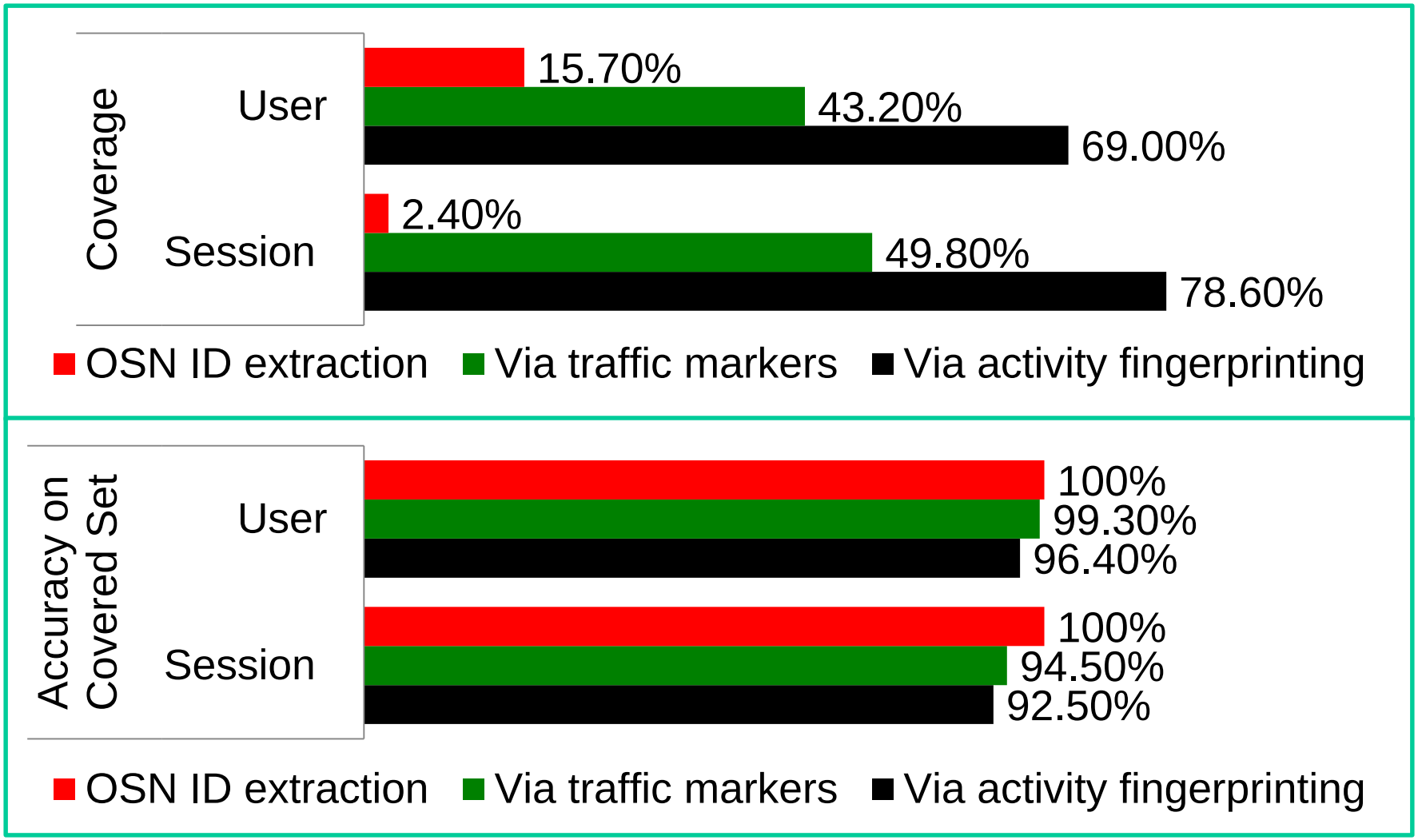


$$\text{Coverage} = \frac{\text{identified sessions/users}}{\text{total sessions/users}}$$

$$\text{Accuracy on Covered Set} = \frac{\text{correctly identified sessions/users}}{\text{total identified sessions/users}}$$

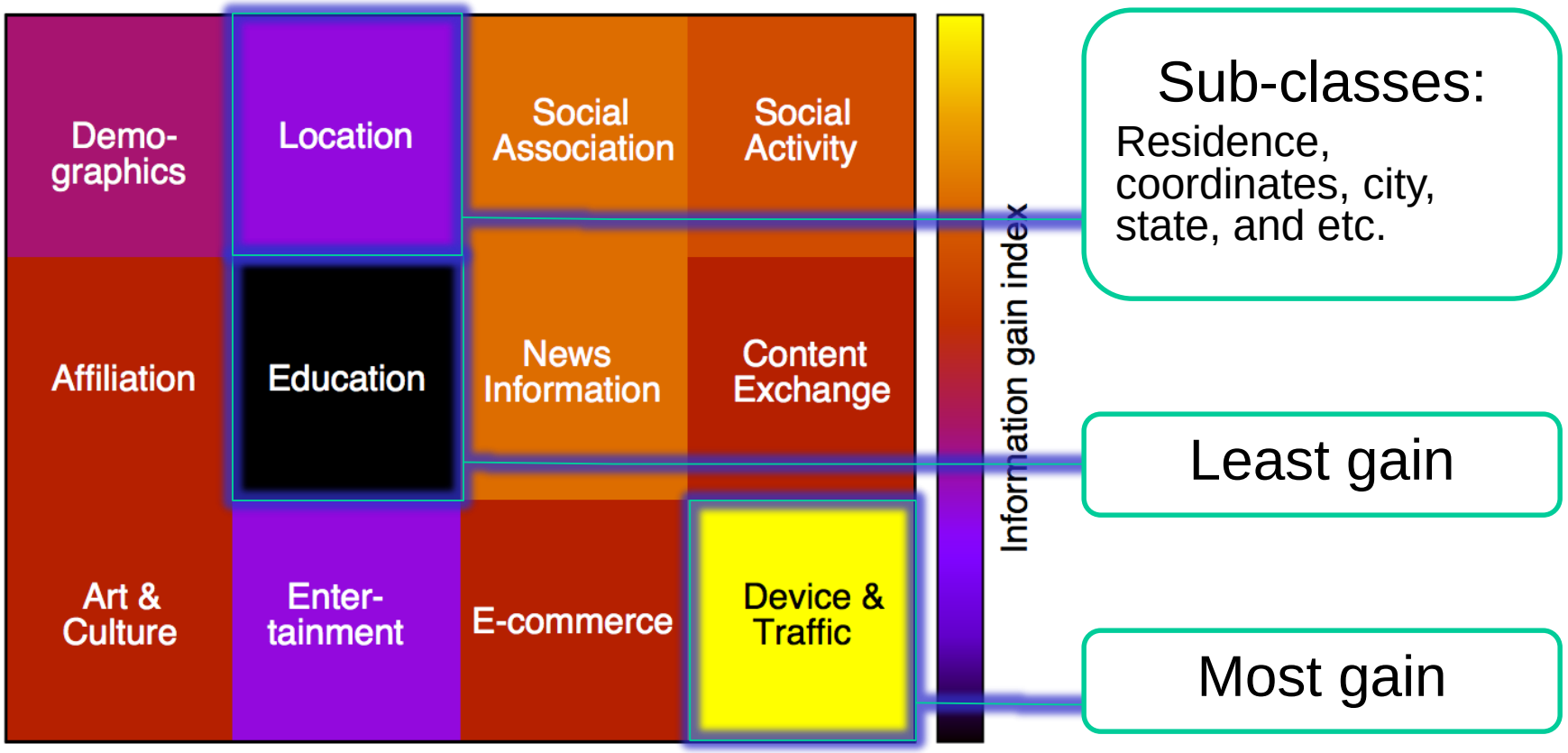
Traffic Attribution Evaluation

- Evaluation Results



Construction of User Mosaic

- Mosaic of Real User

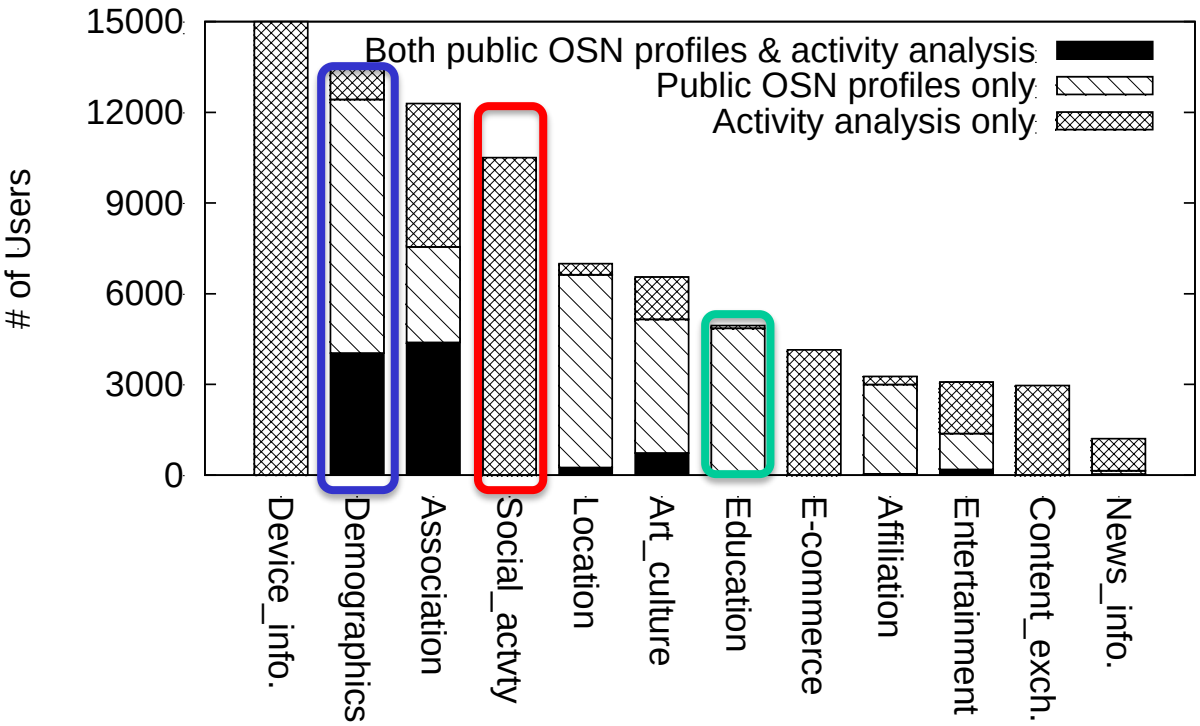


MOSAIC with 12 information classes(***tesserae***):

- Information (Education, affiliation and etc.) from OSN profiles
- Information (Locations, devices and etc.) from users' network data

Quantifying Privacy Leakage

- Leakage from OSN profiles vs. from Network Data



OSN profiles provide ***static*** user information (education, interests)

Analysis on network data provides ***real-time*** activities and locations

Information from both sides can corroborate to each other

Information from OSN profiles and network data can complete and corroborate each other

Preventing User Privacy Leakage

Protect traffic markers

- Traffic markers (OSN IDs and etc.) should be limited and encrypted

Restrict 3rd parties

- Third party applications/developers should be strongly regulated

Protect user profiles

- OSN public profiles should be carefully obfuscated

Conclusions

- Prevalence in the use of OSNs leaves users' true identities available in the network
- Tracking techniques used by mobile apps and services make traffic attribution easier
- Sessions can be labeled with network users' true identities, even without any identity leaks
- Various types of information can be gleaned to paint rich digital Mosaic about users

Mosaic: Quantifying Privacy Leakage in Mobile Network

Thanks!