ACM CoNEXT
Heidelberg
December 4, 2015

National Research Council of Italy

Hiding Mobile Traffic Fingerprints with GLOVE

Marco Gramaglia University Carlos III of Madrid
IMDEA Networks Institute

Marco Fiore CNR-IEIIT
Inria

Institute of Electronics Computer and Telecommunication Engineering

Context

GSMA Intelligence

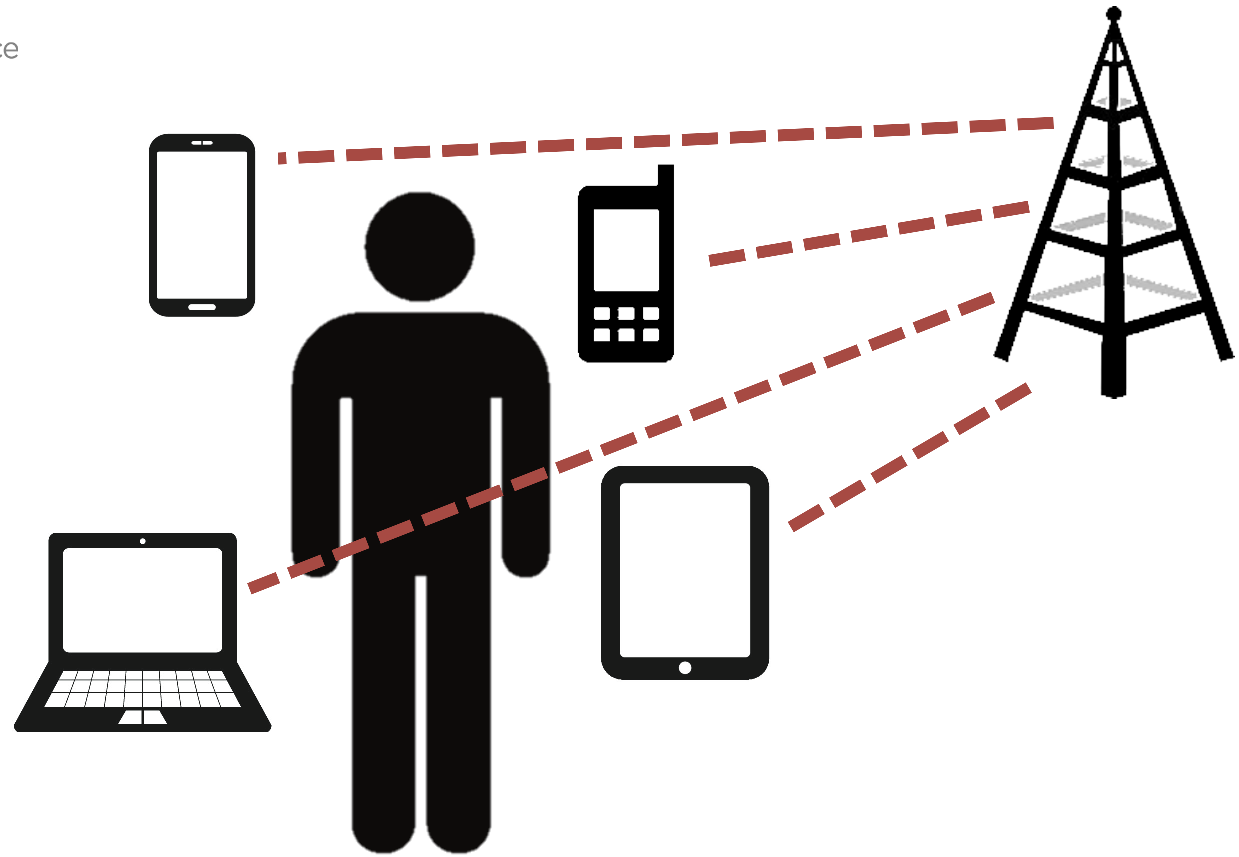
50%

of the world population

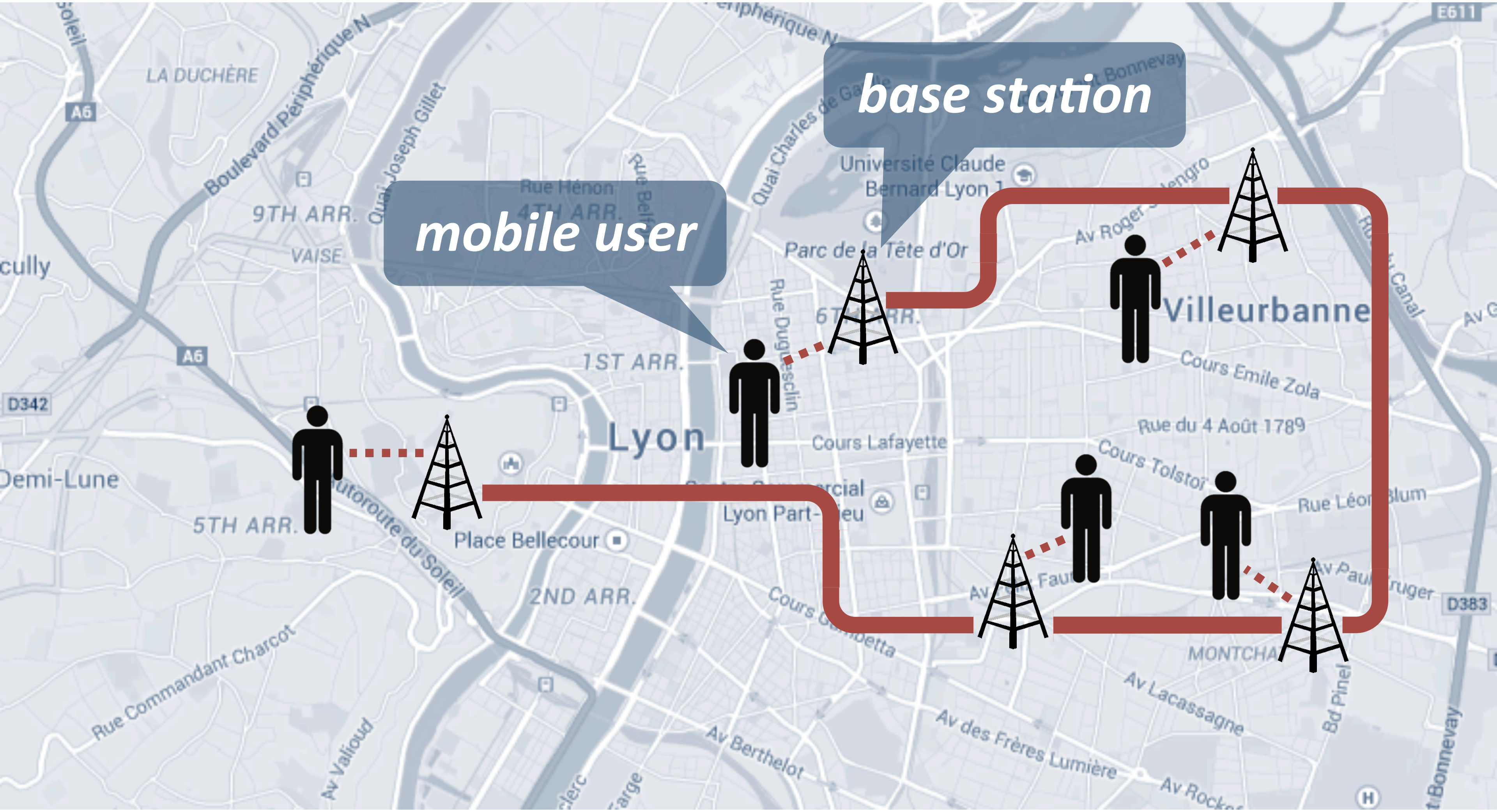
70%

by 2020

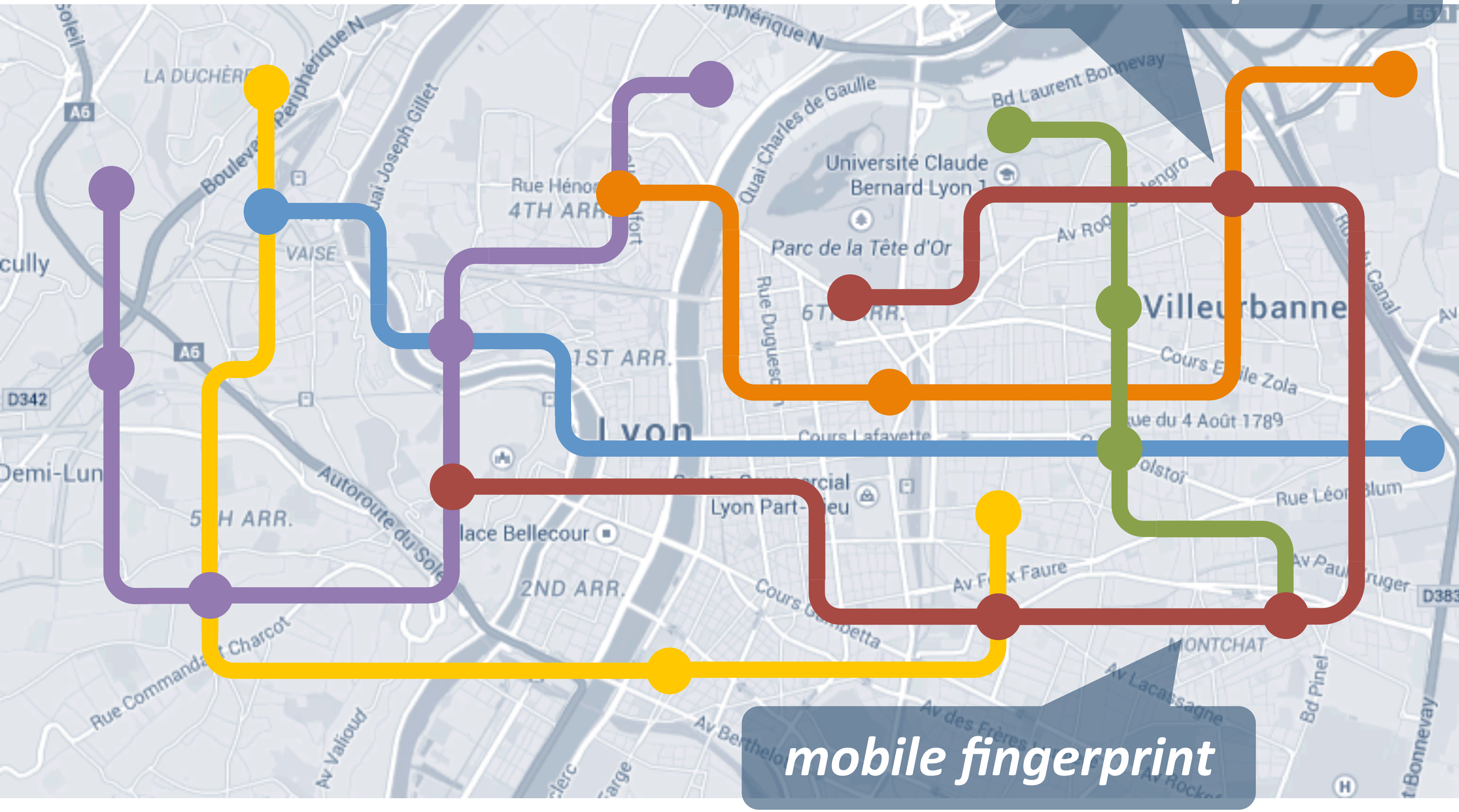
ERICSSON



Context

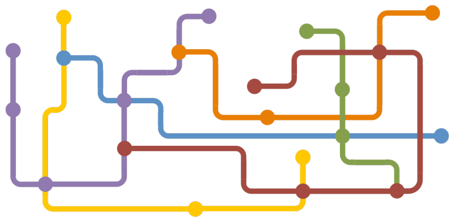


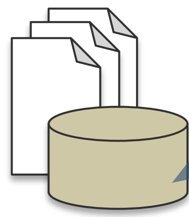
Context


spatiotemporal sample
mobile fingerprint
Lyon
Villeurbanne
Université Claude Bernard Lyon 1
Parc de la Tête d'Or
1ST ARR.
2ND ARR.
4TH ARR.
5TH ARR.
6TH ARR.
LA DUCHÈRE
VAISE
MONTCHAT
Place Bellecour
Lyon Part-Dieu
Quai Charles de Gaulle
Bd Laurent Bonnevay
Cours Émile Zola
Cours Lafayette
Rue du 4 Août 1789
Autoroute du Soleil
Quai Joseph Gillet
Av Félix Faure
Cours Gambetta
Rue Commandant Charcot
Av Lacassagne
Av Berthelot
Rue Léon Blum
A6
D342
D383

Context

millions of fingerprints = *a dataset*





movement micro-data aka moving objects

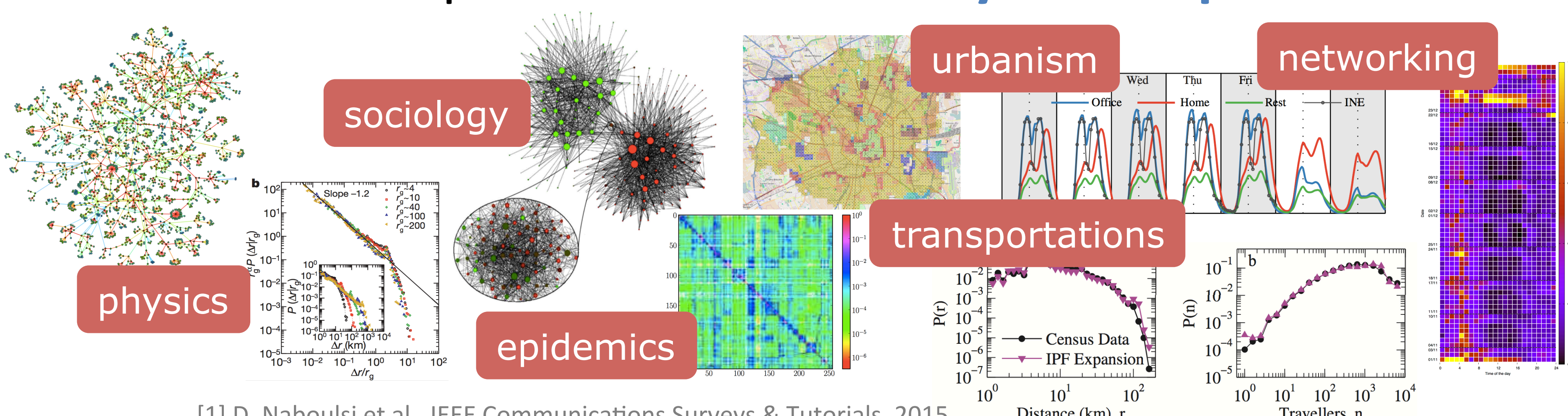
(sparse) spatiotemporal trajectories

D4D challenge data for development — orange™

- **Network operators** collect datasets...
  - using passive monitoring mainly

BIGDATACHALLENGE2015

- ...which are put to use in a **variety of disciplines** [1]



[1] D. Naboulsi et al., IEEE Communications Surveys & Tutorials, 2015

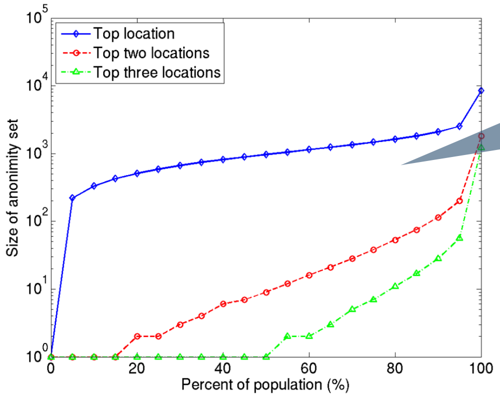
1

The privacy issue

with mobile user fingerprint datasets

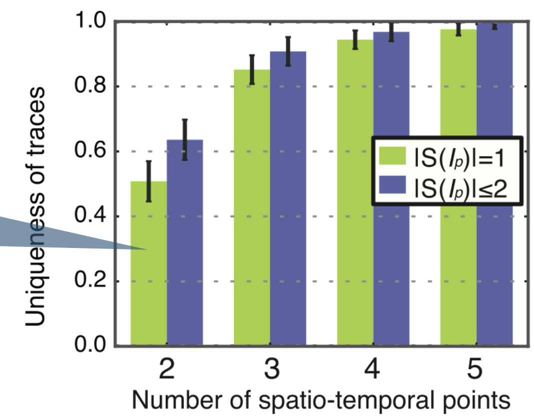
Fingerprint uniqueness

- Mobile fingerprints are **highly unique** [2,3]

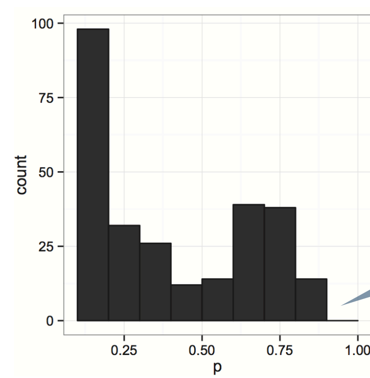


three top locations pinpoint 50% subscribers

five random points pinpoint 95% subscribers



- **Uniqueness** does not imply **re-identification**…
  - …but it is a first step towards it! [4]
- An issue for **data publishing**
  - current solution: NDAs
  - limit research/reproducibility

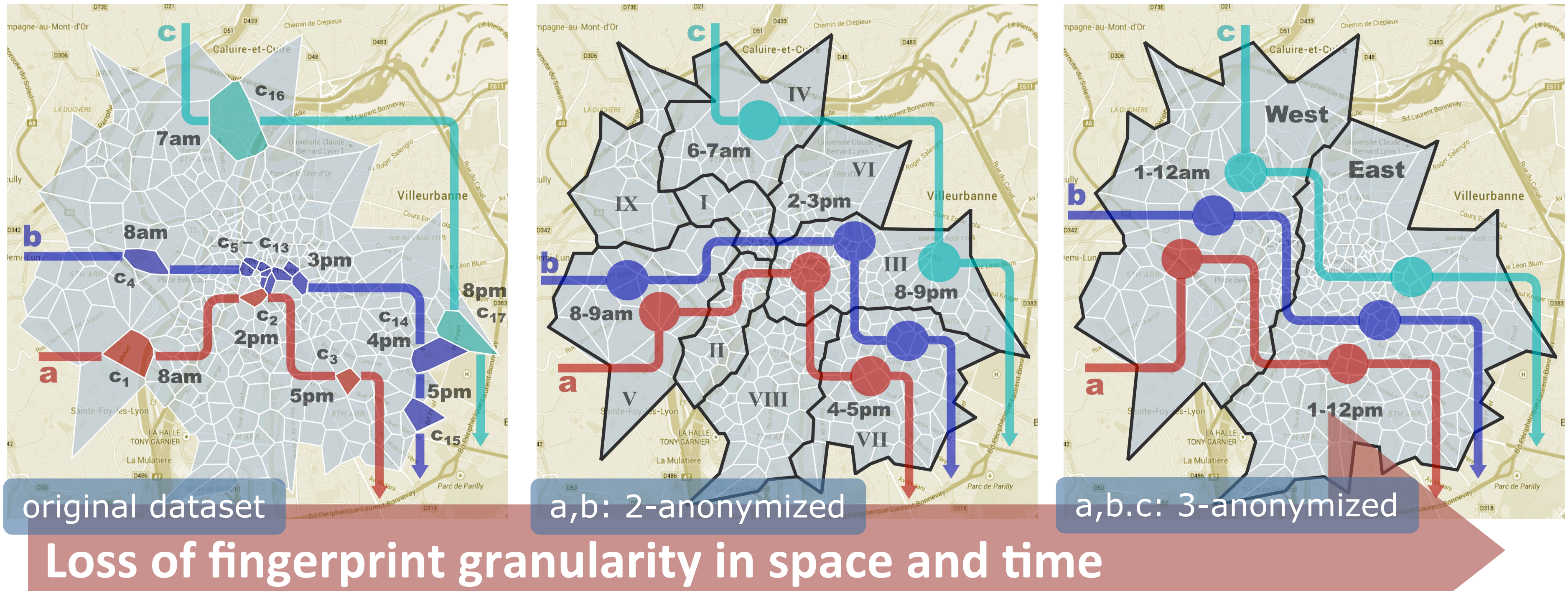


data mixing with Flickr/Twitter can re-identify users with 90%+ confidence

[2] H. Zang, J. Bolot, ACM MobiCom, 2011
[3] Y. De Montjoye et al., Nature Scientific Reports, 2013
[4] A. Cecaj et al., IEEE PerCom Workshops, 2014

How to cope with uniqueness?

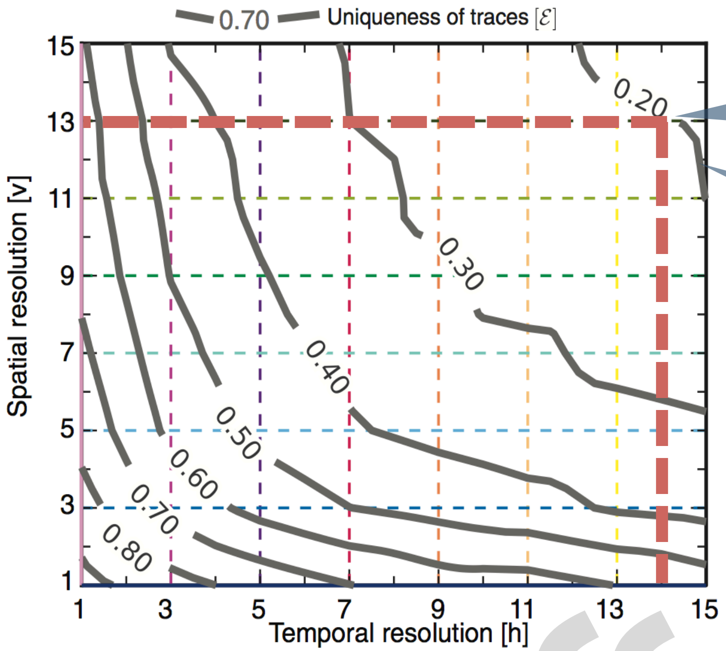
- ***k*-anonymity** is a well-suited privacy model
  - hide each individual in a crowd of k-1 other individuals
  - implemented via **spatiotemporal generalization**



- *k*-anonymity has a **cost** in terms of data accuracy

How much does *k*-anonymity cost?

- Mobile user fingerprint datasets are **extremely expensive** to *2*-anonymize via generalization [3]



granularity reduced to to 14 hours and 13 base station cells

data utility is disrupted!

20% of fingerprints is still unique

no 2-anonymity!

assuming knowledge of just five random points...

- Our contribution: ***"achieving k-anonymity of all full-length fingerprints while preserving substantial data accuracy"***

[3] Y. De Montjoye et al., Nature Scientific Reports, 2013

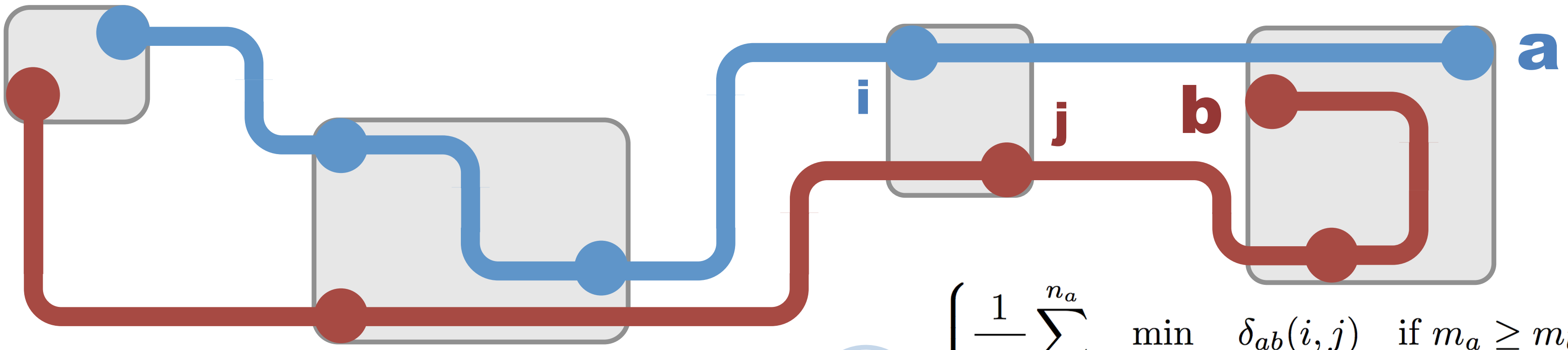
2

Measuring *k*-anonymizability

of mobile user fingerprints

Measure definitions

- **Sample stretch effort**
$$\delta_{ab}(i,j) = w_\sigma \phi_\sigma \left(\sigma_i^a, \sigma_j^b\right) + w_\tau \phi_\tau \left(\tau_i^a, \tau_j^b\right)$$
  - cost (granularity loss) to **make two samples indistinguishable**



- **Fingerprint stretch effort**
$$\Delta_{ab} = \begin{cases} \frac{1}{m_a} \sum_{i=1}^{n_a} \min_{j=1,\dots,m_b} \delta_{ab}(i,j) & \text{if } m_a \geq m_b \\ \frac{1}{m_b} \sum_{j=1}^{n_b} \min_{i=1,\dots,m_a} \delta_{ab}(i,j) & \text{otherwise.} \end{cases}$$
  - cost to **make two complete fingerprints indistinguishable**

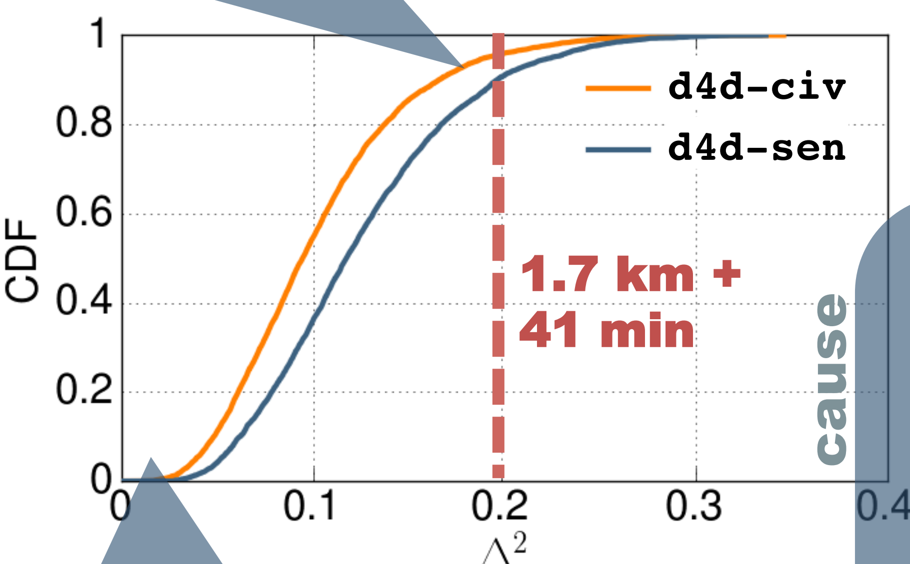
- ***k*-gap**
$$\Delta_a^k = \frac{1}{k-1} \sum_{b \in \mathbb{N}_a^{k-1}} \Delta_{ab}$$
  - **minimum mean** cost to hide one fingerprint with ***k*-1** other

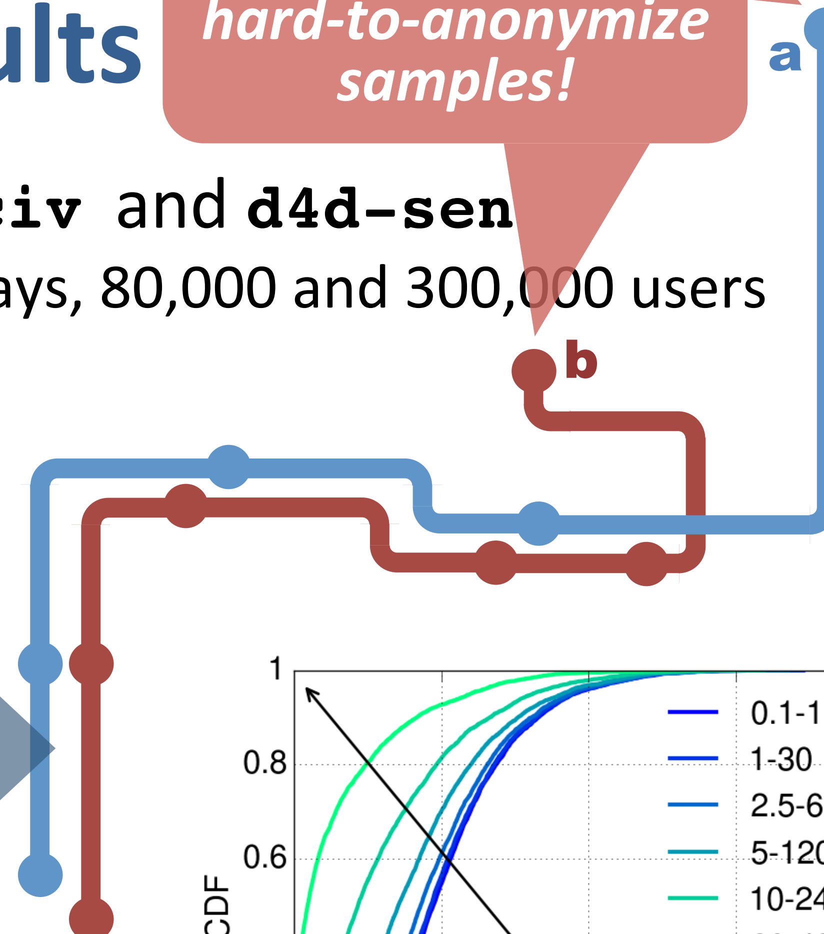
Results

“heavy tail” of hard-to-anonymize samples!

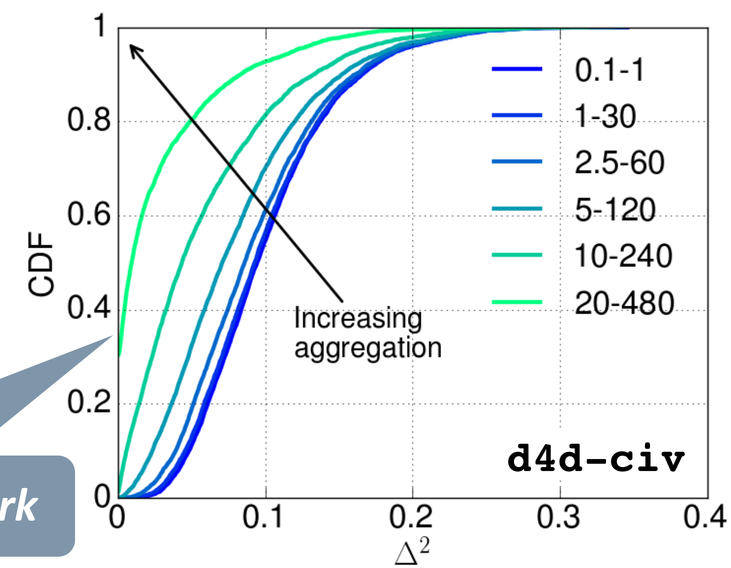
- Reference datasets: **`d4d-civ`** and **`d4d-sen`**
  - calls/sms, nationwide, 14 days, 80,000 and 300,000 users

2-anonymity comes at a low average cost for a vast majority of the users





cause



no user is 2-anonymous

generalization does not work

3

Achieving *k*-anonymity

in mobile user fingerprint datasets

The GLOVE algorithm

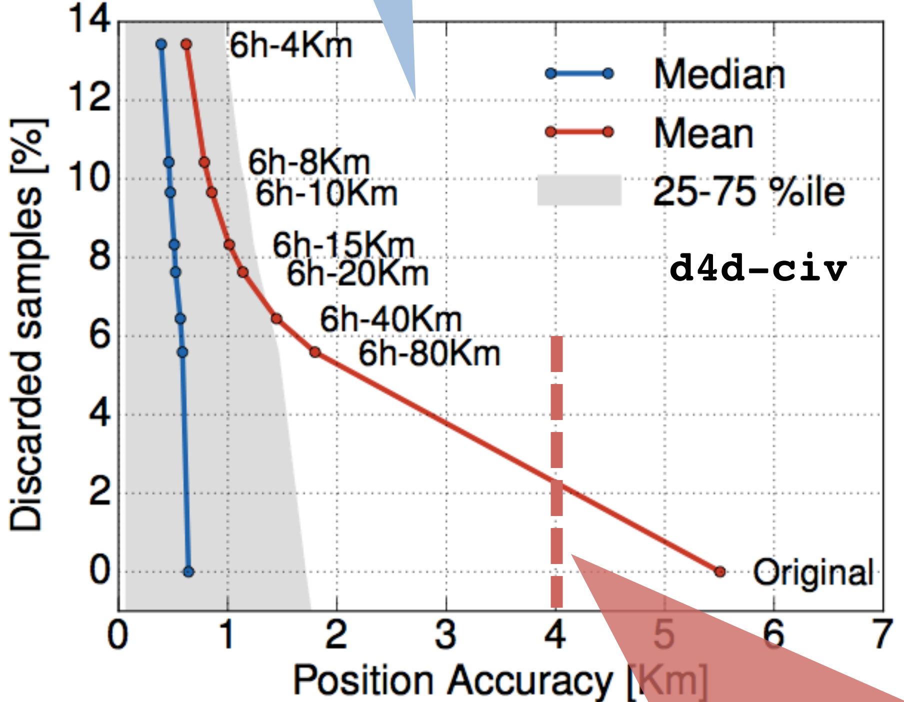
- **Fingerprint stretch effort**
  - captures the diversity between two fingerprints
  - can operate on pairs of fingerprint groups as well
- GLOVE: **greedy hierarchical clustering** algorithm based on **fingerprint stretch effort distances**
  - privacy level *k* is the sole algorithm parameter
  - stopping rule: the desired anonymization level *k* is attained
  - can be combined with **suppression**

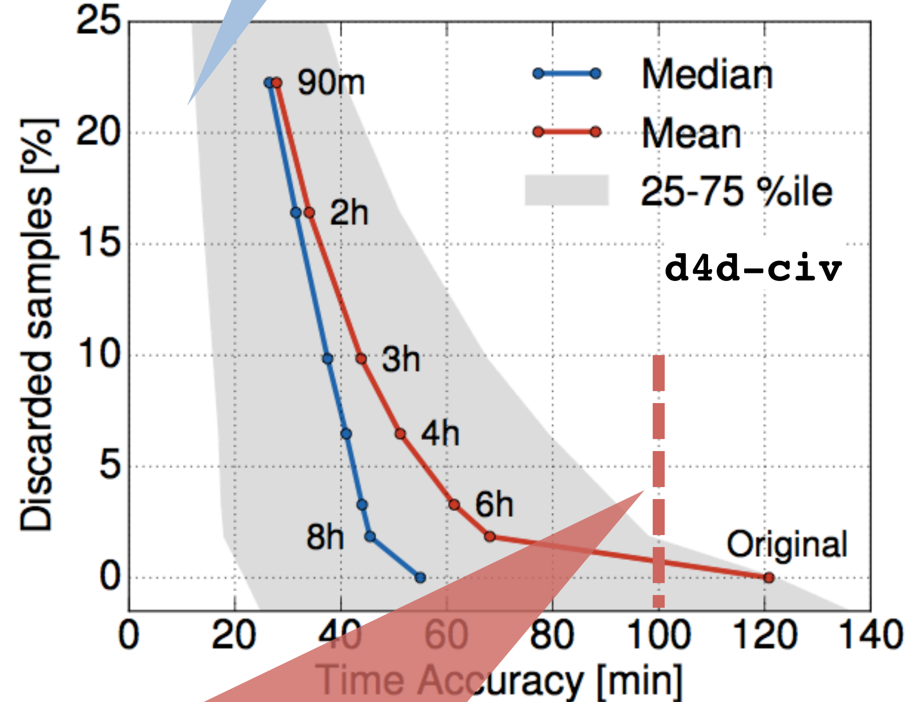
input : Anonymization level k
input : Mobile fingerprint dataset $\mathbb{M}$
output: Anonymized fingerprint dataset $\mathbb{M}$

foreach $a,b \in \mathbb{M}, a \neq b$ **do**
  S $[a,b]$=`calcStretch` (a,b) ;
end
while $\exists\ a,b \in \mathbb{M}$ s.t. $a.k < k,\ b.k < k$ **do**
  $a,b \leftarrow$ `leastStretch(S)` ;
  `remove`$(\mathbb{M}$,S$,a,b)$;
  $m \leftarrow$ `merge`(a,b) ;
  $m.k = a.k + b.k$;
  `add`$(\mathbb{M},m)$;
  if $m.k < k$ **then**
    foreach $c \in \mathbb{M}$ s.t. $c.k < k$ **do**
      S $[c,m]$=`calcStretch` (c,m) ;
    end
  end
end

More performance evaluation

- **Accuracy of samples** in *2*-anonymized datasets
  - over space (geographical span) and time (temporal span)





< 30% of users 2-anonymized with legacy spatiotemporal generalization!

4

Concluding remarks

and future work

Conclusions

- **Contributions**
  - Unveiled the **root cause behind the poor *k*-anonymizability** of mobile user fingerprints in mobile traffic datasets
  - Designed and evaluated a **first algorithm capable of *k*-anonymizing fingerprints** without (fully) disrupting utility
  - More results in the paper
    - impact of *k*, comparative evaluation, different dataset features

- **Open issues**
  - GLOVE is a **proof-of-concept:** large space for improvements
    - fundamental operators / computational efficiency / extensive testing
  - ***k*-anonymity is not a one-for-all solution!**
    - other criteria may be needed to cope with diverse attackers

?

Questions