User Privacy Is Not Preserved with ID-removed Anonymous Cellular Data

Zhen Tu
Tsinghua University
Beijing, China
tuzhen16@gmail.com

Contents

Motivation

Massive Mobile Data

- extensive use of mobile phones
- explosive mobile traffic

Great Potential Value

- academic research
- commercial application
- city management

Both academic and industrial communities are calling for mobile data publishing and sharing.

Motivation

Publishing and Sharing

- Potential risks of leaking mobile user privacy
- Anonymization before data publishing
  - Hashing of user identifiers (week attack resistance[9])
  - Generalization or suppression (low data utility[17])

[9] Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3, 2013.
[17] Hiding mobile traffic fingerprints with glove. CoNext, 2015.

New Way to Open Cellular Data

- Open the meta-data with all the ID or part of ID removed[2,4]
  - Only fine-grained spatio-temporal information remains.
- Publishers' belief: sufficient to protect user privacy & high data utility

[2] China telecom' big data products. *http://www.dtbig.com/*
[4] A case study: privacy preserving release of spatio-temporal density in Paris. SIGKDD, 2014.

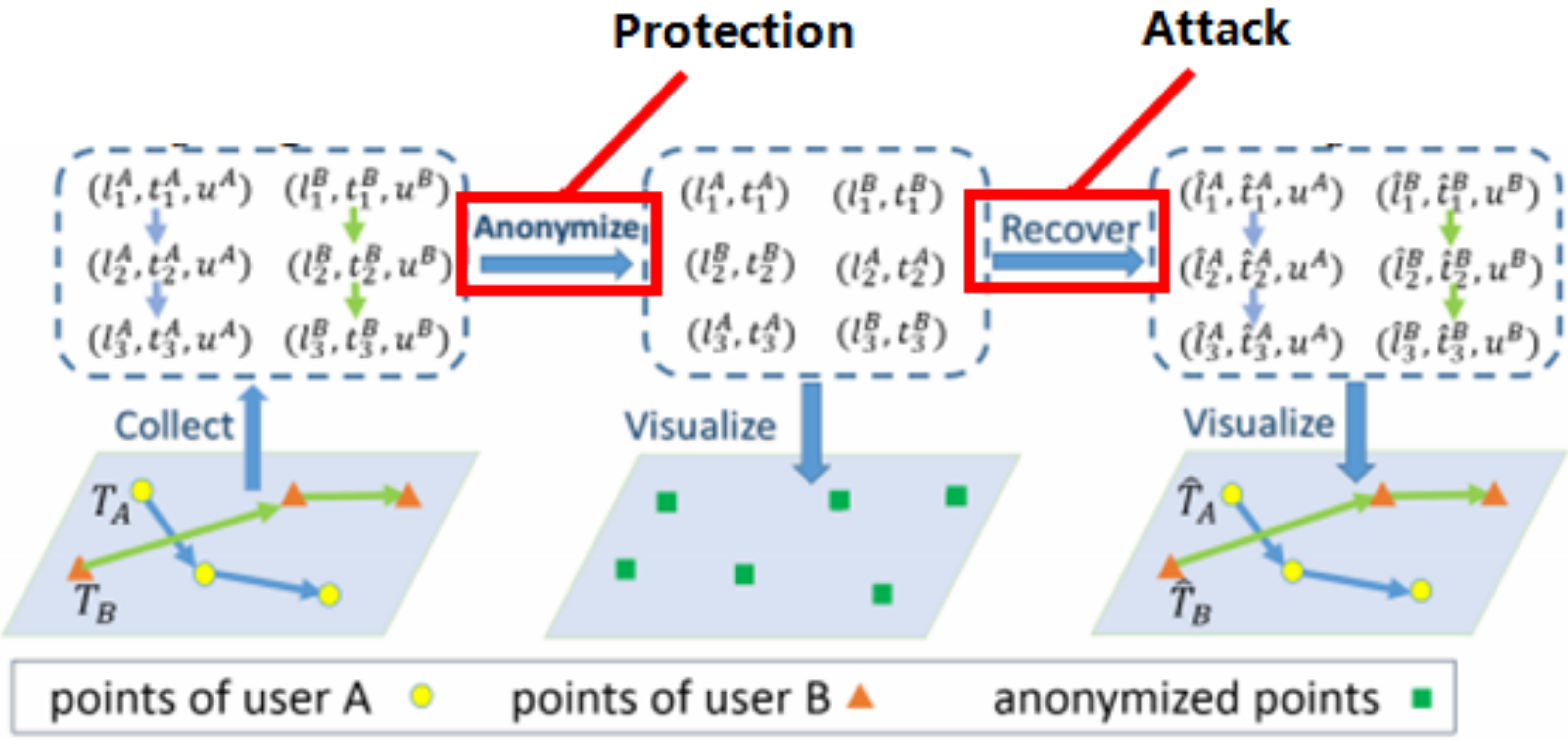
ID-removed Anonymous Data: Is that Really Safe?

- Not safe at all! Indeed, our study shows that it has severe potential user privacy leakage.

Problem Statement

Privacy Concerns in ID-removed Data Publishing Scenario

Is it possible to recover user identifications with no prior information even for such ID-removed meta-data?



Attack System

How to build a feasible attack system?

- The aim of our attack system is to recover user identifications from the ID-removed anonymous cellular data. In other words, we need to identify those spatio-temporal points that belong to a single user. So we have to answer the following questions:

1) Does the trajectory of a single user have his or her own characteristics?

2) Is there any difference between trajectories generated by different users?

Attack System

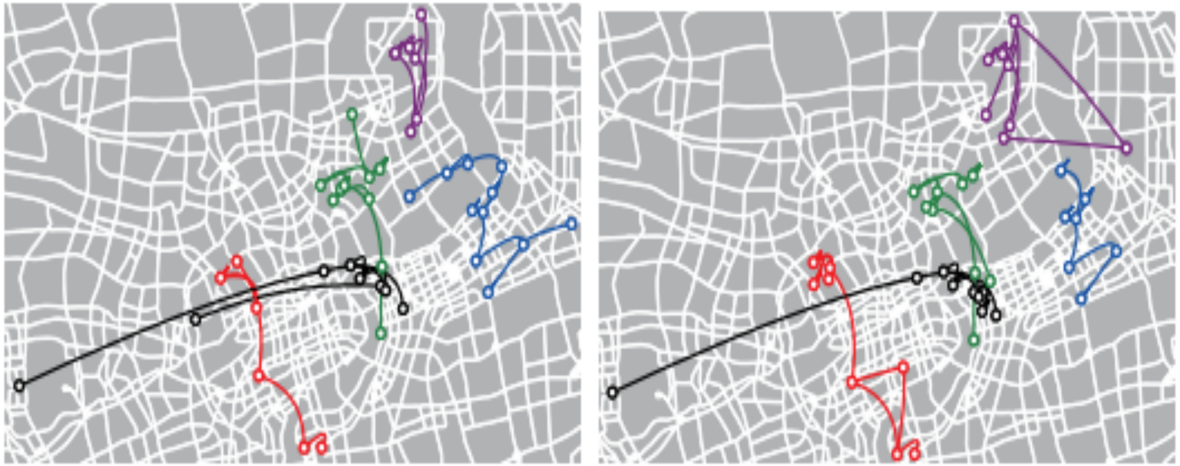
Datasets

Datasets& Metrics	Operator Dataset	Application Dataset
Source	Cellular network	Mobile application
Location	Shanghai, China	Shanghai, China
Time	Apr. 2016	Nov. 2015
Duration	1 week	2 days
User number	5.90 millions	15.50 thousands
Record number	1.54 billions	7.69 millions
Records/user	261	496

Diverse & Representative:

- ◆ cellular network & mobile devices
- ◆ spatial and temporal resolutions
- ◆ total number of records
- ◆ average number of records

Characteristics of Mobile User Trajectories



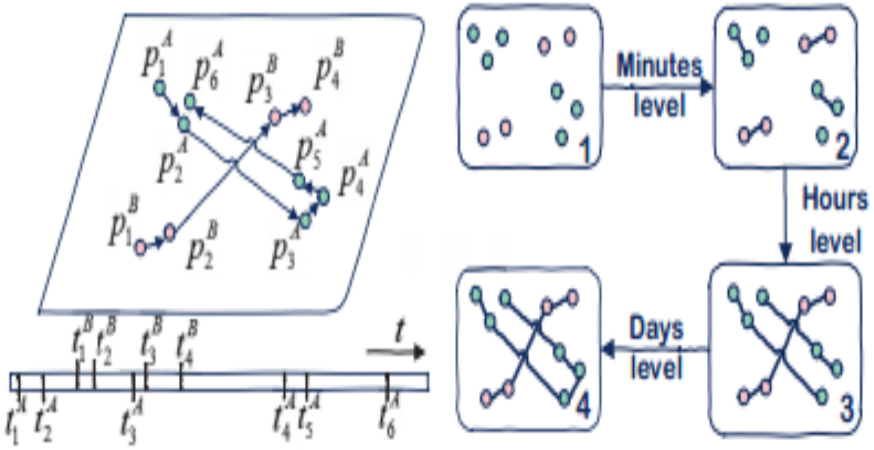
(a) Day 1 (b) Day 2

Figure 2: The locations of cellular towers visited by five randomly selected mobile users.

- ◆ same user: similar traces on day 1 and day 2
- ◆ different users: different mobility traces

Attack System

ID-recovered System



(a) Visualization of original trajectory (b) System pipeline

Figure 5: An overview of the trajectory recovery attack system.

1

- **Minutes-level Identification**
- identify the spatio-temporal records contributed by the same user within several minutes

2

- **Hours-level Identification**
- identify the spatio-temporal records contributed by the same user with a timestamp gap of several hours

3

- **Days-level Identification**
- identify the spatio-temporal records contributed by the same user across different days

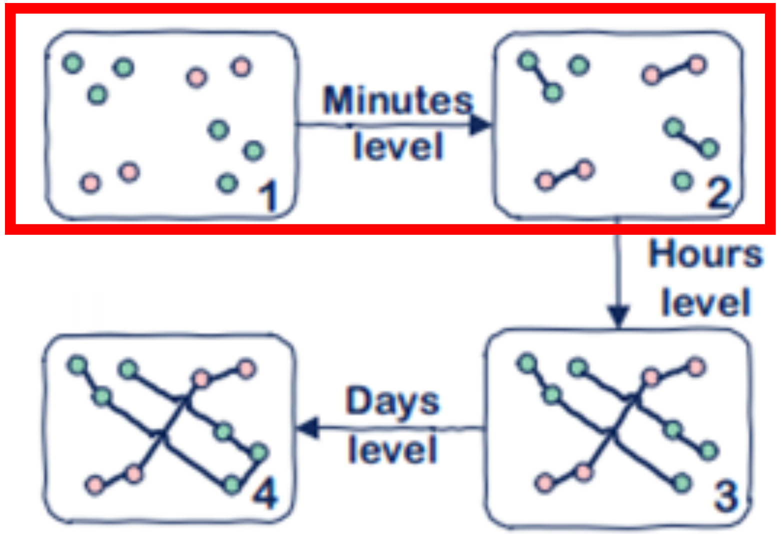
Input : ID-removed spatio-temporal points

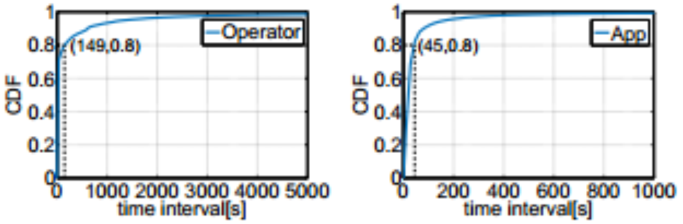
Output: ID-recovered trajectories

Attack System

Minutes-level Identification

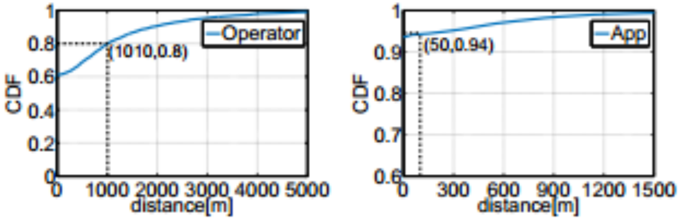
- A single user's trajectory recorded by the cellular network is bursty in both temporal and spatial domain.





(a) Operator dataset (b) App dataset

Figure 6: The CDF of interval time between two sequential records.



(a) Operator dataset (b) App dataset

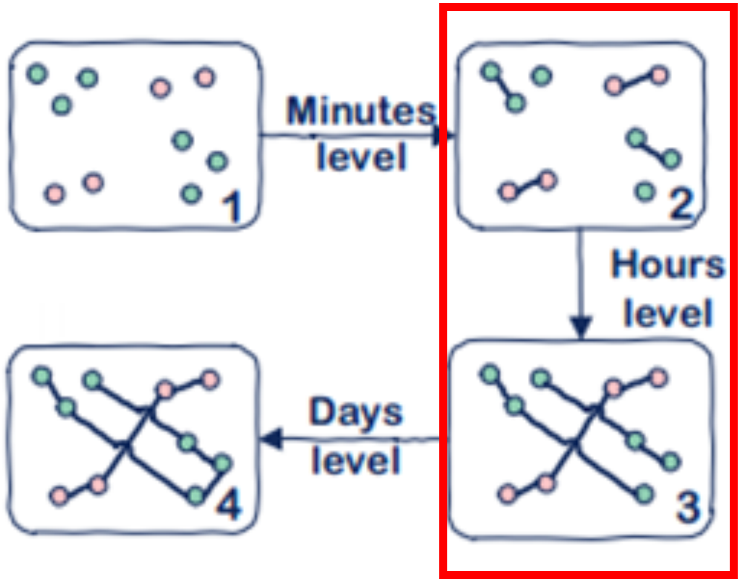
Figure 7: The CDF of the distance between two sequential records.

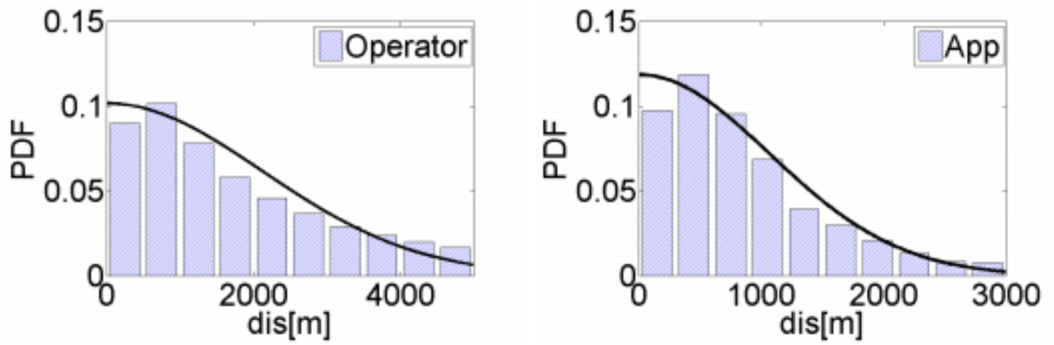
- Bursty records, which have a short time interval and a near distance, have a high probability to be generated by the same person.

Attack System

Hours-level identification

- A single user’s mobility has a continuous feature, thus we can estimate a user’s next location using the current location and velocity.





(a) Operator dataset (b) App dataset

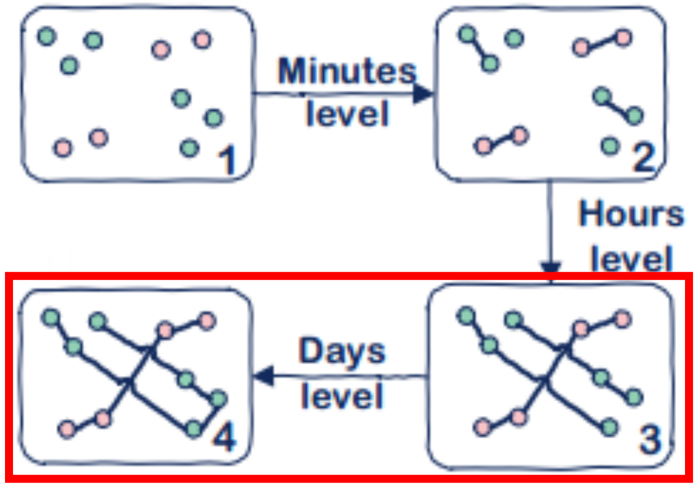
Figure 8: The PDF of the errors between predicted location and the ground truth.

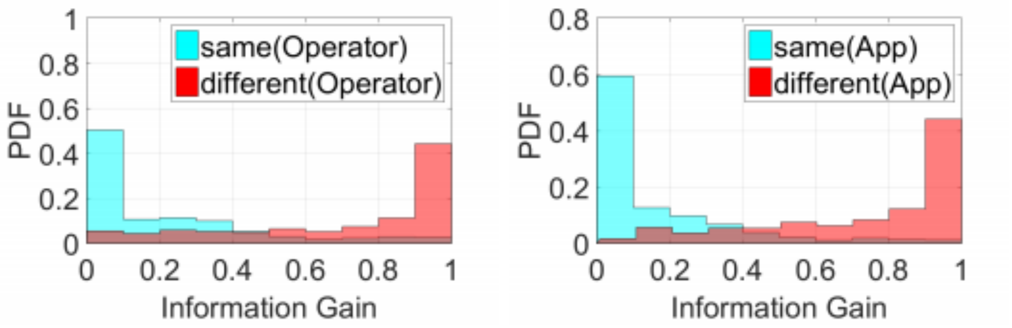
- Continuous traces, when connected the error between predicted location and actual location is small , have a high probability to be generated by the same person.

Attack System

Days-level identification

- A single user’s mobility pattern is regular across days and different users have different mobility patterns.





(a) Operator dataset (b) App dataset

Figure 9: The PDF of information gain in grouping hours-level records contributed by a single user or different users.

The information gain measures the difference between two traces’ location distributions.

- Similar traces, when connected the information gain is small, have a high probability to be generated by the same person.

Contents

Performance Evaluation

Recovery Results

1)App dataset

1000 original trajectories

331,036 ID-removed spatio-temporal points

1255 ID-recovered trajectories

Recall Rate: 76.7%
Precision Rate: 84.3 %
F1 Score: 80.3%

2) Operator Dataset

5000 original trajectories

1135,838 ID-removed spatio-temporal points

5780 ID-recovered trajectories

Recall Rate: 71.7%
Precision Rate:73.3%
F1 Score: 72.2%

We have recovered the ID-removed cellular data with high accuracy!

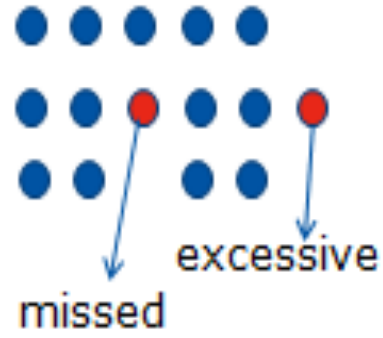
Accuracy of each ID-recovered trajectory

☐Metric

Original trajectory(N=5)

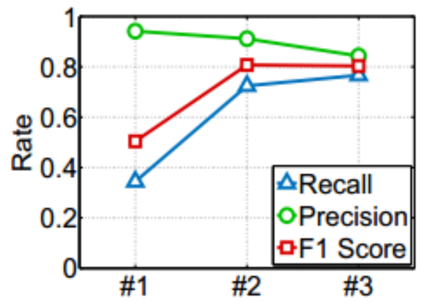
ID-recovered trajectory(M=6)

Accurate trace points(L=4)

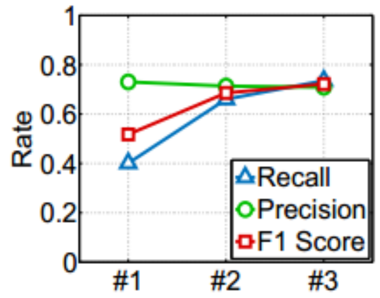


Recall Rate = L/M = 66.7%
Precision Rate= L/N = 80%
F1 score = 2x(Re x Pr)/(Re + Pr) = 72.7%

☐Result



(a) App dataset

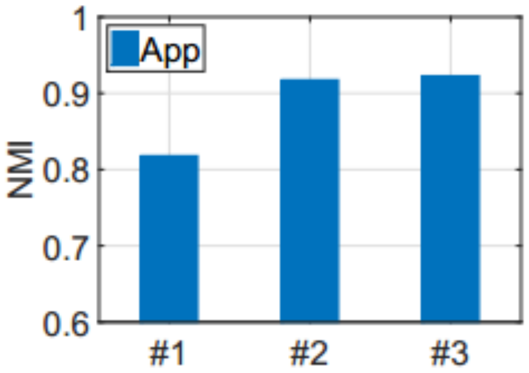


(b) Operator dataset

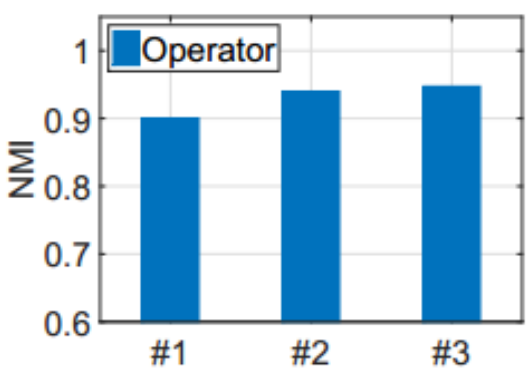
Performance Evaluation

Privacy Leakage Level

- **Normalized mutual information(NMI)**
  - An index to quantify the amount of information over the original trajectories that we can obtain from the recovered trajectories.
  - Higher the value is, more the user privacy leaks.



(a) App dataset



(b) Operator dataset

Our system is able to recover over 90% information of the original trajectories.

User privacy is not preserved with ID-removed anonymous cellular data!

Performance Evaluation

Key Factors to Reduce Privacy Leakage

- Dataset Scale

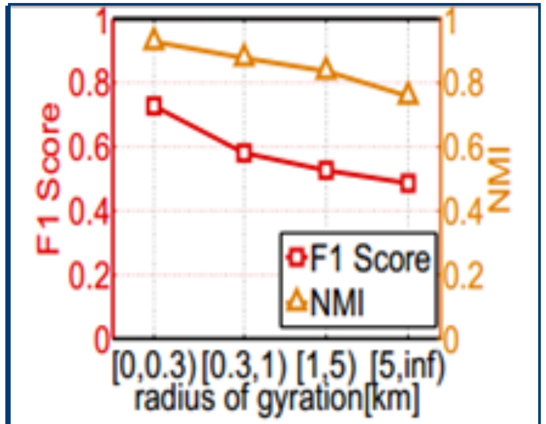
Tips: only publish and share large-scale datasets.

- Data Resolution

Tips: open datasets with low spatial granularity.

- Mobility Behavior
  - Radius of gyration is an index to measure the space covered by each user's trajectory, users of high mobility usually have large radius of gyration.

Tips: only share trajectories of high mobility and large active area.



(b) App: NMI and F1

Summary

Innovation

We are the first to identify and study the privacy problem about ID-removed anonymous cellular data.

Observations

◆ ID-removed anonymous cellular data has severe potential user privacy leakage.
◆ Dataset scale, data resolution and mobility behaviors are key factors to impact the extent of privacy leakage.

Guidelines

- only publish large-scale datasets
- open datasets with low spatial granularity
- only share trajectories of high mobility and large active area

Thanks you!
I’m happy to take questions.

For Data Sample, Please Contact

tuzhen16@gmail.com
liyong07@tsinghua.edu.cn
FIB-LAB: http://fi.ee.Tsinghua.edu.cn