De-anonymization of Mobility Trajectories: Dissecting the Gaps between Theory and Practice

Huandong Wang¹, Chen Gao¹, Yong Li¹, Gang Wang², Depeng Jin¹, Jingbo Sun³

¹Tsinghua University, China
²Virginia Tech
³China Telecom Beijing Research Institute
Increasing Concern on Privacy/Security

- Anonymized user trajectories are increasingly collected by ISPs
  - High research and business value

- Growing privacy concern
  - ISPs are motivated to monetize or share user trajectory data

- De-anonymization attack
  - How likely users can be de-anonymized in the shared ISP trajectory dataset?
De-anonymization Attack: Theory and Practice

- **Appalling Theoretical Privacy Bound**
  - 4 location points uniquely re-identify 95% users [Scientific Report 2013]

  
  
  
  Is this true in practice?

- **Practical Challenge: Lack of large real-world ground-truth datasets**
  - Small datasets
    - 1717 users in [WWW 2016]
  - Synthetized datasets
    - Parts of the same dataset [TON 2011]
Our Approach: **Collect Three Real-world Ground-truth Datasets**

**Ground-Truth: Traces from the same set of users**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Total# Users</th>
<th>Total# Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISP</td>
<td>2,161,500</td>
<td>134,033,750</td>
</tr>
<tr>
<td>Weibo App-level</td>
<td>56,683</td>
<td>239,289</td>
</tr>
<tr>
<td>Weibo Check-in (Historical)</td>
<td>10,750</td>
<td>141,131</td>
</tr>
<tr>
<td>Weibo Check-in (One-week)</td>
<td>506</td>
<td>873</td>
</tr>
<tr>
<td>Dianping App-level</td>
<td>45,790</td>
<td>107,543</td>
</tr>
</tbody>
</table>

**ISP Dataset**
- Shanghai, 4/19-4/26, 2016 (victim dataset)
- 2 million users
- Access logs to cellular tower → Location traces

**Weibo Dataset**: One of the largest social networks in China (external information)

**Dianping Dataset**: “Chinese Yelp” (external information)
How to Obtain the Ground-Truth?

Ethical approval obtained from Weibo and Dianping

- ISP Traces
  - Weibo ID in HTTP Request
    - Weibo → Check-ins → GPS in ULR parameter
  - Dianping ID in HTTP Request
    - Dianping → GPS in ULR parameter
De-anonymization Attack: Threat Model

- **Anonymized Trajectory Data Published by ISP**
  - Anonymization: Replace user identity with the pseudonym

- **Adversary**
  - Match the anonymized traces (e.g., ISP traces) and external traces (e.g., Weibo/Dianping traces)
  - Social network has PII $\rightarrow$ real-world identifier

![Diagram showing comparison between external trajectories and anonymized trajectories, followed by similarity score function, candidate trajectories, and performance function.]
De-anonymization: Theoretical Bound based on Uniqueness

- Number of points sufficient to uniquely identify a trajectory
- $T\downarrow p$: Randomly sampled $p$ points
- $A(T\downarrow p)$: find all trajectories containing the $p$ points of $T\downarrow p$
- **Uniqueness**: $|A(T\downarrow p)|=1$?

5 points are sufficient to uniquely identify 75% trajectories! High potential risk of trajectories to be de-anonymized!
De-anonymization Attack: Actual Performance

Implement 7 state-of-the-art algorithms

- “Encountering” event
  - POIS [WWW 2016]
  - ME [AIHC 2016]

- Individual user’s mobility patterns
  - HMM [IEEE SP 2011]
  - WYCI [WOSN 2014]
  - HIST [TIFS 2016]

- Tolerating temporal/spatial mismatches
  - NFLX [IEEE SP 2008]
  - MSQ [TON 2013]

Maximum hit-precision is only 25%!
Far from the privacy bound!
Reasons Behind Underperformance

Existing algorithms tolerating spatio-temporal mismatches have the best performance

Algorithms with best performance

**NFLX [IEEE SP 2008]**
- Similarity function
  - Minimum time gap between users’ visits to the same location
- Tolerate temporal mismatches

**MSQ [TON 2013]**
- Similarity function
  - Square root of distance between trajectories
- Tolerate spatial mismatches
Reasons Behind Underperformance: Large Spatio-Temporal Mismatches

Spatial mismatches of over 40% records $\geq 2$km

Temporal mismatches of over 30% records

Significant Time and location Mismatches between Different Datasets!
Potential Reasons behind the Mismatches

- **GPS errors**
  - GPS unreachable locations (Indoor, underground)
  - Lazy GPS updating mechanisms [UbiComp 2007]

- **Deployment of base stations**
  - Lower density → larger mismatches

- **User behavior**
  - 39.9% remote (fake) check-ins [ICWSM 2016]
  - Earn virtual rewards, compete with their friends
The vast majority of users have sparse location records!

Cumulative distribution function (CDF)

Reasons Behind Underperformance: Data Sparsity

Data Sparsity => Rare "Encountering" Event! => Inaccurate Mobility Modelling!

Sparser location records → Worse performance
Can we bridge this gap?
Our De-anonymization Method

\[ D_{GM}(S, L) = \log p(S|L) = \prod_{S(t) \neq \emptyset} p(S(t)|L). \]

1) Modelling Spatio-Temporal Mismatches: Gaussian Mixture Model (GMM)

\[ PS(t) \quad L = \sum_{p=H↓l \uparrow H↓u} \pi(p) \cdot \mathcal{N}(S(t)|L(t-p), \sigma^2) (p) \]

- Parameters chosen by empirical values or estimated by EM algorithm

2) Modelling Users’ Mobility Pattern: Markov Model

- Solving the data sparsity issue: rare “encountering” event
- Missing locations are estimated by Markov Model
Our De-anonymization Method

3) Use Location Context
- Solve the data sparsity issue
- Use aggregated user behavior at locations
- To infer individual user behavior (location transition probability)

4) Use Time Context
- “Whether the user is active” is helpful
- Modelling user inactive period (previously ignored feature)
Performance Evaluation

- 7 state-of-the-art algorithms
- Our proposed algorithm: GM-B, GM
- Transferred parameters: GM-B (Trans.)

Our proposed algorithms outperform baselines by over 17%
Summary

- **Large-scale Ground-truth Datasets**
  - ISP trajectories with over 2 million users
  - 2 different social networks, 2 different types of external information

- **Demonstrate the Gaps between Theory and Practice**
  - High theoretical bound
  - Low actual performance

- **Bridge the Gaps between Theory and Practice**
  - Considering spatio-temporal mismatches, data sparsity, location/time context
  - Improve the performance → confirm our observations
Thanks you!

For Data Sample and Code, Please Contact
whd14@mails.tsinghua.edu.cn
liyong07@tsinghua.edu.cn
Reference


Metric of the ranking

- Hit-precision:

\[
h(x) = \begin{cases} 
\frac{k-(x-1)}{k}, & \text{if } k \geq x \geq 1, \\
0, & \text{if } x > k.
\end{cases}
\]

- If the right one rank 1 in candidate trajectories, \( h(x)=1 \).
- If the right one rank 3 in candidate trajectories, \( h(x)=(k-2)/k \).
Performance Evaluation: Parameter Study

- **Larger Tolerant Delay** => Better Performance
  - 0->1: Significant improvement
  - 12->24: Little improvement

- **Comparable Performance**
  - Empirical vs. Estimated
  - Robust to parameter settings.
Our De-anonymization Method

Use Location Context:
- Solve the **sparsity** issue (inaccurate mobility modelling)

Use the aggregate user behavior at locations!

### Marginal distribution

\[
E(r) := p(L(t) = r) = \frac{\sum_{t \in T} I(L(t) = r) + \alpha(r)}{\sum_{t \in T} I(L(t) \neq \emptyset) + \sum_{r \in R} \alpha(r)}.
\]

\[
\alpha(r) = \alpha_0 \cdot \sum_{v \in V} \sum_{t \in T} I(L_v(T) = r),
\]

### Transition matrix

\[
T(r_1, r_2) := p(L(t) = r, L(t + 1) = r),
\]

\[
= \frac{\sum_{t \in T} I(L(t) = r_1)I(L(t + 1) = r_2) + \beta(r_1, r_2)}{\sum_{t \in T} I(L(t) \neq \emptyset)I(L(t + 1) \neq \emptyset) + \sum_{r_2, r_2 \in R} \beta(r_1, r_2)}.
\]

\[
\beta(r_1, r_2) = \beta_0 \cdot \sum_{v \in V} \sum_{t \in T} I(L_v(t) = r_1) \cdot I(L_v(t + 1) = r_2).
\]
**Our De-anonymization Method**

- **Use Time Context**
  - Whether there is record in each time bin is also an important information (previously ignored feature).

\[
D_B(S, L) := \log \prod_{t \in T} P(I_{S(t)}|I_{L(t)})
\]

\[
= \log \prod_{t \in T} P_{1|1}^{I_{S(t)}} P_{0|0}^{I_{L(t)}} P_{0|1}^{1-I_{S(t)}} P_{1|0}^{1-I_{L(t)}}
\]

\[
D_{GM-B} = D_{GM} + D_B
\]

\[
p(S|L) = \prod_{S(t) \neq \emptyset} p(S(t)|L).
\]