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

**Huandong Wang**\(^1\), Chen Gao\(^1\), Yong Li\(^1\), Gang Wang\(^2\), Depeng Jin\(^1\), Jingbo Sun\(^3\)

\(^1\)Tsinghua University, China  
\(^2\)Virginia Tech  
\(^3\)China 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]

- **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]

Is this true in practice?
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 → real-world identifier

![Diagram showing the process of de-anonymization attack with external trajectories vs. anonymized trajectories, similarity score function, candidate trajectories, performance function, and attack performance.](image)
De-anonymization: Theoretical Bound based on Uniqueness

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

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]

Hit-precision

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

Actual Performance Based on Weibo’s App-level Trajectories

Maximum hit-precision is only 25%!
Far from the privacy bound!
Existing algorithms tolerating spatio-temporal mismatches have the best performance

**Reasons Behind Underperformance**

**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 ≥ 2km

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!

Reasons Behind Underperformance: Data Sparsity

Cumulative distribution function (CDF)

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

Sparsier location records → Worse performance

Data Sp

The vast majority of users have sparse location records!
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)

\[ P(S(t)|L) = \sum_{p=-H_t}^{H_u} \pi(p) \cdot \mathcal{N}(S(t)|L(t-p), \sigma^2(p)) \]

- 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
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

Impact of Maximum Tolerant Delay

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

Impact of Parameters in GMM

- Comparable Performance
  - Empirical vs. Estimated
  - Robust to parameter settings.