Measuring Membership Privacy on Aggregate Location Time-Series

ACM SIGMETRICS 2020

Apostolos Pyrgelis\(^1\), Carmela Troncoso\(^1\), and Emiliano De Cristofaro\(^2\)

\(^1\) EPFL \hspace{1cm} \(^2\) UCL & Alan Turing Institute
Introduction

Mobility analytics are useful in modern cities for journey planning, etc.

Large-scale collection and usage of individual users’ location data prompts privacy concerns

Pseudonymization / anonymization of location traces is ineffective
Let There Be Aggregation

Analysts are given access to aggregate location statistics, e.g., time-series

Privacy-Friendly: Individual user data is hidden in the crowd!

Utility: Forecasting Traffic
Anomaly Detection
Hotspot Discovery
Map Inference

Real World Use Cases: Uber Movement, Waze, Telefonica Smart Steps
But, Location Aggregates Leak Privacy

• Recent research has shown that location aggregates can be exploited for:
  
  • User Profiling / Localization (PETS’17)
  
  • Trajectory Extraction (WWW’17)
  
  • Membership Inference (NDSS’18)

1) Important privacy implications if the aggregates relate to a group sharing a sensitive characteristic, e.g., disease, income, etc.

2) A first step to other more invasive attacks
In This Work

• Measurement study to understand Membership Inference attacks (MIAs) on aggregate location time-series
  
  • Which spatio-temporal factors contribute to the inference?
  
  • Which users are more vulnerable than others?
  
  • How well defense strategies based on generalization, hiding, and perturbation protect against MIAs?
  
  • How do these defenses perform wrt. mobility analytics tasks? e.g., traffic forecasting, hotspot discovery, etc.
Real-world Mobility Datasets

**Transport for London (TFL)**
- Oyster Card trips of London commuters
- Monday, March 1 to Sunday March 28, 2010 (4 weeks)
- 60M trips / 4M users / 582 stations (ROIs)
- Sparse / Regular

**San Francisco Cabs (SFC)**
- GPS mobility traces of taxis in SF
- May 19 - June 8, 2008 (3 weeks)
- 11M coordinates / 534 cabs / 10x10 downtown grid (ROIs)
- Dense / Irregular

Generate hourly time-series, # of users in a ROI
Outline

- Understanding MIAs
- Evaluating Defenses against MIAs
- Studying Privacy-Utility Tradeoffs
Methodology

**Target Users**
Randomly pick 150 users from 3 mobility groups and run MIA

**Sample & Aggregate**
Balanced dataset of groups that include / exclude the target and aggregate their locations

**Classification**
Use of a Logistic Regression classifier

**Dimensionality Reduction**
Use of Principal Component Analysis (PCA)

**Adversarial Prior Knowledge:**
- Target’s Location Data
- Target’s Past Location Patterns
Spatio-temporal Factors

Prior: Target’s Past Location Patterns
Group Size: 9,5K

Prior: Target’s Location Data
Group Size: 100

Commuter Regularity!

Dense GPS trajectories – large attack surface

Sparse Locations / Times
## Mobility Characteristics

<table>
<thead>
<tr>
<th>Feature</th>
<th>TFL</th>
<th>SFC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Events</td>
<td>0.03</td>
<td>0.17</td>
</tr>
<tr>
<td>Unique Locations</td>
<td>0.39</td>
<td>0.01</td>
</tr>
<tr>
<td>Active Timeslots</td>
<td>0.06</td>
<td>0.23</td>
</tr>
<tr>
<td>Locations per Timeslot</td>
<td>0.05</td>
<td>0.30</td>
</tr>
<tr>
<td>Active Timeslots / Weekday</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Active Timeslots / Weekend</td>
<td>0.11</td>
<td>0.01</td>
</tr>
<tr>
<td>Events / Weekday</td>
<td>0.01</td>
<td>0.07</td>
</tr>
<tr>
<td>Events / Weekend</td>
<td>0.13</td>
<td>0.03</td>
</tr>
<tr>
<td>Spatial Entropy</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>Temporal Entropy</td>
<td>0.06</td>
<td>0.01</td>
</tr>
<tr>
<td>Unicity</td>
<td>0.16</td>
<td>0.17</td>
</tr>
</tbody>
</table>

**Prior:** Target’s Location Data

A. Pyrgelis, C. Troncoso, and E. De Cristofaro - Measuring Membership Privacy on Aggregate Location Time-Series
Take Aways

- Various spatio-temporal factors (e.g., commuting patterns, dense GPS trajectories) contribute to the attack

- Users contributing more data points to the aggregates are more susceptible to MIA

- Movements in sparse locations/times ease MIA

- Unique mobility patterns are identifiable in the aggregates

- Regular mobility patterns reveal users’ membership to the aggregates
Outline

• Understanding MIAs

• Evaluating Defenses against MIAs

• Studying Privacy-Utility Tradeoffs
Defenses Evaluation

- **Generalization**: Spatial, Temporal, Data

<table>
<thead>
<tr>
<th>Space</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>45 65</td>
<td>23 76</td>
</tr>
<tr>
<td>13 12</td>
<td>24 7</td>
</tr>
<tr>
<td>123 256</td>
<td>276 164</td>
</tr>
<tr>
<td>56 52</td>
<td>55 59</td>
</tr>
</tbody>
</table>

- **Spatial Generalization**, e.g., 58
- **Temporal Generalization**, e.g., 108
- **Data Generalization**, e.g., 160 < # < 170
Defenses Evaluation

- **Generalization**: Spatial, Temporal, Data

- **Hiding**: Sampling, Suppression
Defenses Evaluation

- **Generalization**: Spatial, Temporal, Data

- **Hiding**: Sampling, Suppression

- **Perturbation**: Differential privacy, Crowd-blending privacy
Defenses Evaluation

- **Generalization:** Spatial, Temporal, Data

- **Hiding:** Sampling, Suppression

- **Perturbation:** Differential privacy, Crowd-blending privacy

**Privacy Gain:** Normalized decrease in the attack’s performance given the *defended vs raw* aggregates
Take Aways

• Spatio-temporal generalization does not protect against MIA - data generalization can be configured to do so

• Hiding techniques work better when the input signal is sparse

• Perturbation techniques that achieve DP yield high privacy – similar protection levels can be reached with less noise

• Combining defenses can improve privacy
Outline

• Understanding MIAs

• Evaluating Defenses against MIAs

• Studying Privacy-Utility Tradeoffs
## Mobility Analytics

<table>
<thead>
<tr>
<th>Task</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forecasting Traffic</td>
<td>Mean Relative Error</td>
</tr>
<tr>
<td>Anomaly Detection</td>
<td>(Pearson’s) Correlation</td>
</tr>
<tr>
<td>Hotspot Discovery</td>
<td>F1 Score</td>
</tr>
<tr>
<td>Map Inference</td>
<td>Distribution Similarity (Jensen Shannon)</td>
</tr>
</tbody>
</table>

**Utility Loss:** Decrease in utility compared to performing the same task on *raw* aggregate location time-series
Privacy-Utility Tradeoffs

Forecasting Traffic

Hotspot Discovery

Transport for London

A. Pyrgelis, C. Troncoso, and E. De Cristofaro - Measuring Membership Privacy on Aggregate Location Time-Series
Take Aways

• Different defenses yield variable tradeoffs for various analytics

• No single defense preserves the utility of the analytics for arbitrary applications

• Spatio-temporal generalization yields poor privacy and utility

• Other defenses can achieve reasonable tradeoffs for specific tasks:
  • Data generalization – forecasting traffic
  • Hiding – map inference
  • Perturbation – hotspot discovery
  • Combining hiding + perturbation - anomaly detection
Conclusion

• Measurement study to understand Membership Inference Attacks (MIAs) on aggregate location time-series

  • Regular/uncommon mobility patterns are easy to recognize

  • **Size matters**: users contributing more data to the aggregates are easier to attack

  • There is no single characteristic that can be singled out and thwart the attack

  • There does not exist a single defense that protects against MIA while enabling arbitrary mobility analytics

  • Different defenses yield variable privacy-utility tradeoffs for different settings and analytics

  • Some defenses yield reasonable tradeoffs for specific tasks

There is need for work on the design of novel defenses!
The end...

Thank you for your attention!

For more details, see our full paper: https://arxiv.org/abs/1902.07456

Contact Details: apostolos.pyrgelis@epfl.ch