Differentially Private Location Privacy in Practice

Vincent Primault
Sonia Ben Mokhtar
Cédric Lauradoux
Lionel Brunie

May 17th 2014
Location-based services
PLEASE ROB ME

Raising awareness about over-sharing

Check out our guest blog post on the CDT website.
Location privacy threats

Only 4 points are sufficient to uniquely identify you! [1]

Can a protection mechanism efficiently protect points of interest of a user?
Outline

• Introduction
• **About points of interest**
• Protection mechanisms
• Experimental settings
• Evaluation metrics & results
• Sum-up
A mobility trace
Points of interest
Outline

• Introduction
• About points of interest
• **Protection mechanisms**
• Experimental settings
• Evaluation metrics & results
• Sum-up
Location-privacy protection mechanisms

- Pseudonymity
- Mix-zones
- Noise-based solutions
- Spatial cloaking
- k-anonymity
- Cryptographic protocols
Geo-indistinguishability

Geo-indistinguishability

Level of privacy $l_i$ within $r_i$ proportional to an $\varepsilon$

Outline

• Introduction
• About points of interest
• Protection mechanisms
• **Experimental settings**
• Evaluation metrics & results
• Sum-up
## Two different data sets

<table>
<thead>
<tr>
<th>San Francisco cabs</th>
<th>Geolife</th>
<th>Reduced Geolife</th>
</tr>
</thead>
<tbody>
<tr>
<td>In the SF Bay Area</td>
<td>Around Beijing</td>
<td>Around Beijing</td>
</tr>
<tr>
<td>1 month in 2009</td>
<td>4 years (2007-2011)</td>
<td>1 continuous month</td>
</tr>
<tr>
<td>536 taxis</td>
<td>182 users</td>
<td>61 users</td>
</tr>
<tr>
<td>11 millions points</td>
<td>25 millions points</td>
<td>5 millions points</td>
</tr>
</tbody>
</table>
POIs extraction algorithm

1. Mobility trace
2. Extract stays [4]
4. Time-ordered list of locations
   - 1 hour
   - Centroids of areas where a user has spent at least \( \text{minTime} \) within a \( \text{maxDistance} \) radius
   - Stays within \( \frac{3}{4} \text{maxDistance} \) where a user passed through at least \( \text{minPts} \) times
5. POIs
   - 2 times
   - A set of important places for a user

Playing with distance threshold

<table>
<thead>
<tr>
<th></th>
<th>SF cabs</th>
<th>Geolife</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unobfuscated</td>
<td>250 m</td>
<td>250 m</td>
</tr>
<tr>
<td>Weak privacy</td>
<td>700 m</td>
<td>600 m</td>
</tr>
<tr>
<td>Medium privacy</td>
<td>1000 m</td>
<td>1200 m</td>
</tr>
<tr>
<td>Strong privacy</td>
<td>2000 m</td>
<td>2500 m</td>
</tr>
</tbody>
</table>

We must greatly increase the `maxDistance` threshold at highest privacy levels in order to retrieve an interesting number of POIs.
Outline

• Introduction
• About points of interest
• Protection mechanisms
• Experimental settings
• Evaluation metrics & results
• Sum-up
Measuring privacy
Recall rate is the proportion of real POIs successfully retrieved.
# Recall rate

<table>
<thead>
<tr>
<th>Privacy Level</th>
<th>SF cabs</th>
<th>Geolife</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weak privacy</td>
<td>73 %</td>
<td>72 %</td>
</tr>
<tr>
<td>Medium privacy</td>
<td>72 %</td>
<td>71 %</td>
</tr>
<tr>
<td>Strong privacy</td>
<td>71 %</td>
<td>61 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset Type</th>
<th>SF cabs</th>
<th>Geolife</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference (unobfuscated)</td>
<td><strong>1111 POIs</strong> (~ 2/user)</td>
<td><strong>258 POIs</strong> (~ 4/user)</td>
</tr>
</tbody>
</table>
Geographic distance

Geographic distance between an obfuscated POI and the nearest real POI
Cumulative geographic distance

SF cabs

Geolife
Re-identification rate

Scenario: I use a LBS without any protection and one day, I use a geo-indistinguishable mechanism.

Will my privacy be preserved or will the LBS be able to link my obfuscated trace with my original trace?
Re-identification rate

Distance = median(geographic distances)
Re-identification rate

Associate to each set of obfuscated POIs the set of real POIs with which it has the minimal distance.

Re-identification $= \frac{1}{3}$
Re-identification rate

<table>
<thead>
<tr>
<th></th>
<th>SF cabs</th>
<th>Geolife</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong privacy</td>
<td>6 %</td>
<td>63 %</td>
</tr>
<tr>
<td>Medium privacy</td>
<td>8 %</td>
<td>83 %</td>
</tr>
<tr>
<td>Weak privacy</td>
<td>10 %</td>
<td>90 %</td>
</tr>
</tbody>
</table>

- Few unique patterns in SF cabs data set, drivers are likely to have a similar behavior.
- Mobility patterns can be captured in Geolife and act like a fingerprint.
Measuring precision

Eviction rate is the ratio between the number of useless results and the total number of results.

Precision is 1 minus the eviction rate.

Eviction rate = 6 / 10
Precision = 4 / 10
Precision of results when querying LBS

• 100 points sampled from the SF cabs dataset

• Use a "find restaurants 500 meters around me" query against OpenStreetMap data
Outline

• Introduction
• About points of interest
• Protection mechanisms
• Experimental settings
• Evaluation metrics & results
• Sum-up
Conclusion

• Protection mechanisms improve privacy...
  – but still allow to infer a large quantity of sensitive information (> 60 %)
  – at the cost of degraded performance

• Difficult to achieve a trade-off between precision, utility and performance
Future work

• Study the exact impact of the temporal component
• Investigate if dynamically adapting the privacy parameter can help
• Propose counter-measures w.r.t. our framework and related work