PULP: Achieving privacy and utility trade-off in user mobility data

Sophie Cerf, Bogdan Robu, Nicolas Marchand, gipsa-lab, Univ. Grenoble Alpes
Vincent Primault, Antoine Boutet, Sara Bouchenak, Sonia Ben Mokhtar, LIRIS, INSA Lyon
Robert Birke, Lydia Y. Chen, IBM Zurich Research Lab
Outline

- Context & Motivation
- Problem Statement & Objectives
- PULP framework
- Experimental Evaluation
- Discussion
- Perspectives
Context

Mobility data collection

Provide **useful** information to companies and researchers
Exploitation of location data raises Privacy issues

Some threats:

→ Re-identification attack
→ Mobility prediction
→ Social relationships inference
→ Extraction of Points Of Interest (POI)
Location Privacy Protection Mechanisms

Real Data → LPPM → Obfuscated Data
LPPM exemple

- **Geo-Indistinguishability**
  - Configuration parameter: $\varepsilon$ in meters$^{-1}$

Images showing raw data and data with different noise levels.

- Geo-I with few noise
- Geo-I with a lot of noise

Configuration parameter: $\varepsilon = 0.1$

Configuration parameter: $\varepsilon = 0.01$
LPPM exemple

- Promesse
  - Configuration parameter : $\varepsilon$ in meters
Obfuscation of a mobility dataset with a user granularity with a set of LPPMs, each one configurable by a single parameter.
Problem Statement

Privacy and utility of obfuscated data depends on:

- the LPPM used
- its configuration
- the user
For each user, given objectives of privacy and utility, how to choose a LPPM and configure it?

For each user:

user mobility trace
privacy and utility objectives

PULP

LPPM* & configuration $\epsilon*$
State of the Art

• (Agir, 2014) iteratively modify the configuration to meet the privacy objective
  
  **Limits:** Computing intensive, no utility

• (Chatzikokolakis, 2015) adapts Geo-I’s parameter to the density of the area
  
  **Limits:** not objective driven, no utility

• (Primault, 2016) iteratively evaluates the privacy and utility for refining configuration parameters
  
  **Limits:** Computing intensive
Outline

- Context & Motivation
- Problem Statement & Objectives
- PULP framework
- Experimental Evaluation
- Discussion
- Perspectives
PULP Principle

Not yet another LPPM

For each user, find adequate obfuscation to achieve privacy and utility objectives

Adaptation to:
- user
- LPPM
- metrics
PULP: Achieving privacy and utility trade-off in user mobility data – Sophie Cerf

**Used Metrics**

- **Privacy** $\rho$: proportion of hidden Points of Interest
- **Utility** $\mu$: proportion of areas rightly covered
I. PULP Profiler

- **Objective**: Characterize a user regarding a LPPM

→ Automated computation of LPPMs on the user data
I. PULP Profiler
II. PULP Modeler

- **Objective**: Model the impact of LPPMs on a user
II. PULP Modeler

Models:
\[ \rho = a_\rho \cdot \tan^{-1}(b_\rho (\ln(\epsilon) - c_\rho)) + d_\rho \]
\[ \mu = a_\mu \cdot \tan^{-1}(b_\mu (\ln(\epsilon) - c_\mu)) + d_\mu \]
III. PULP Configurator

- **Objective:** Find adequate LPPM & configuration to achieve privacy to utility trade-off

<table>
<thead>
<tr>
<th>LPPM 1</th>
<th>LPPM 2</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_p(\varepsilon)$</td>
<td>$f_p(\varepsilon)$</td>
<td></td>
</tr>
<tr>
<td>$f_\mu(\varepsilon)$</td>
<td>$f_\mu(\varepsilon)$</td>
<td></td>
</tr>
</tbody>
</table>
III. PULP Configurator

OBJECTIVES:

I. $\rho\mu$-ratio
   - $\rho = \omega_{pr/ut} \mu$, maximize $\rho$ and $\mu$

II. $\rho\mu$-thld
    - $\rho \geq \rho_{min}$, $\mu \geq \mu_{min}$

III. $\rho$-thld
    - $\rho \geq \rho_{min}$, maximize $\mu$

IV. $\mu$-thld
    - $\mu \geq \mu_{min}$, maximize $\rho$
III. PULP Configurator

For $\rho\mu$-ratio ($\rho=\omega_{pr/ut}\mu$):

1) Find the solution for each LPPM $j$

$$\varepsilon_j^* = \arg\min_{\varepsilon_j} |\rho(\varepsilon_j) - \omega_{pr/ut}\mu(\varepsilon_j)|$$

2) LPPM selection

$$LPPM^* = \arg\max_{j\in LPPM} (\rho(\varepsilon_j^*) + \omega_{pr/ut}\mu(\varepsilon_j^*))$$
Outline

• Context & Motivation
• Problem Statement & Objectives
• PULP framework
• Experimental Evaluation
• Discussion
• Perspectives
PULP Evaluation

Experimental Setup

- **770 users** from 4 datasets
  - **MDC** People in Geneva region
  - **CabSpotting** Taxi cab in San Francisco
  - **Geolife** People in Beijing
  - **Privamov** Students in Lyon Campus
PULP Evaluation

Modeler

- Very good model accuracy in median
- And still good for extreme users

![Graph showing model accuracy](image-url)
PULP Evaluation

Configurator

- 97% of users have $\omega_{pr/ut} = \omega_{pr/ut}^* \pm 1\%$
Outline

• Context & Motivation
• Problem Statement & Objectives
• PULP framework
• Experimental Evaluation
• Discussion
• Perspectives
PULP Analysis

- Not a single LPPM fits all users and all objectives
PULP Analysis

• Not a single LPPM *configuration* fits all users and all objectives
PULP: Achieving privacy and utility trade-off in user mobility data – Sophie Cerf

PULP Analysis

- Barely no user get really low privacy and utility
PULP Analysis

- **Metrics are parametrized**
  - Privacy: POI diameter, POI duration, matching distance
    - Defined by user
  - Utility: Cell size
    - Defined by the service
PULP Analysis

- All metrics are well modeled
  - Privacy: POI diameter

![Promesse Graph](image1)

![Geo-I Graph](image2)
Comparison with State of the Art

• **State of the Art**
  - ALP: heuristic solution that maximises privacy and utility

• **Setup**
  - $\omega_{pr/ut} = 1$
  - Dataset: Geolife

• **Results**
  For 80% of users:
  - ALP: $\rho \geq 0.8$ and $\mu \geq 0.4$
  - PULP: $\rho \geq 0.7$ and $\mu \geq 0.7$

• **Execution times**
  - *ALP*: 10 hours
  - *PULP*: 1 minute
Conclusion

For each user, PULP finds adequate obfuscation (LPPM and its configuration) to achieve privacy and utility objectives.

PULP is a automated tool that adapts to user, LPPM and metrics.
Perspectives

Explore **temporality** of the data

- **Off-line:**
  - Data quality
  - Patterns detection

- **On-line:**
  - Privacy metric
  - Real time obfuscation
  - Prediction
Thank you for your attention

This work has been partially supported by LabEx PERSYVAL-Lab funded by the French program Investissement d’Avenir, & the LabEx IMU of Université de Lyon in the frame of the MobiCampus-UdL project.
**Formally:** $i$ user index ($n$ in total), $\text{poi}(T)$ set of POIs of trace $T$, $\text{cell}(T)$ set of areas covered by trace $T$

**Privacy $\rho$:** matching of POI Between raw and obfuscated traces.

Precision:

$$P_\rho(i) = \frac{|\text{Matched}(\text{poi}(T'_i), \text{poi}(T_i))|}{|\text{poi}(T'_i)|}$$

Recall:

$$R_\rho(i) = \frac{|\text{Matched}(\text{poi}(T'_i), \text{poi}(T_i))|}{|\text{poi}(T_i)|}$$

Harmonic mean:

$$\rho = 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{2 \cdot P_\rho(i) \cdot R_\rho(i)}{P_\rho(i) + R_\rho(i)}$$

**Utility $\mu$:** Comparison of covered areas between raw and obfuscated traces.

Precision:

$$P_\mu(i) = \frac{|\text{cell}(T'_i) \cap \text{cell}(T_i)|}{|\text{cell}(T'_i)|}$$

Recall:

$$R_\mu(i) = \frac{|\text{cell}(T'_i) \cap \text{cell}(T_i)|}{|\text{cell}(T_i)|}$$

Harmonic mean:

$$\mu = 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{2 \cdot P_\mu(i) \cdot R_\mu(i)}{P_\mu(i) + R_\mu(i)}$$