A Case Study: Privacy Preserving Release of Spatio-temporal Density in Paris

Gergely Acs (INRIA)
gergerely.acs@inria.fr

Claude Castelluccia (INRIA)
claude.castelluccia@inria.fr
Motivation: XData project

- Postal data
- Call Data Record (CDR)
- Electricity consumption data
- Demographical data
- Water management data
Problem: European Data Protection law (Directive 95/46/EC)

- All datasets have to be anonymized such that data subjects are no longer identifiable.

- Who is identifiable?
  - “account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person”

  ➔ Anonymization must be done before cross-processing datasets.

- In practice, CNIL (French Data Protection Office) checks if data releases are compliant with the rule.
Privacy Model

- **Differential privacy**

  $$e^{-\varepsilon} \leq \frac{\Pr(M(D) = D^*)}{\Pr(M(D') = D^*)} \leq e^\varepsilon$$

- **composes securely**: retain privacy guarantees in the presence of independent releases\(^1\)

- even with arbitrary external knowledge!

---

\(^1\) S.R. Ganta, S. Kasiviswanathan, A. Smith. *Composition Attacks and Auxiliary Information in Data Privacy*. KDD’08
Focus: Anonymization of CDR

Postal data

Electricity consumption data

Call Data Record (CDR)

Demographical data

Water management data
(Simplified) Call Data Record

<table>
<thead>
<tr>
<th>Rec #</th>
<th>Phone</th>
<th>Lat</th>
<th>Lon</th>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0644536701</td>
<td>46.345</td>
<td>2.32</td>
<td>13:34:12</td>
<td>Incoming SMS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>01/09/2007</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0634556702</td>
<td>47.123</td>
<td>1.65</td>
<td>14:31:02</td>
<td>Outgoing Call</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>02/09/2007</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 4 types of events:
  - Incoming SMS/Call
  - Outgoing SMS/Call
  - Phone numbers are scrambled (No Personal Data in the dataset)
Paris CDR (provided by Orange™)

- 1,992,846 users
- 1303 towers
- Mean trace length: 13.55 (std.dev: 18)
- Max. trace length: 732
Data to be anonymized

- Number of individuals at a given hour at any IRIS cell in Paris

IRIS cells

- Challenge: Large dimensional data
Overview of our approach

1. Sample $x \approx 30$ visits per user uniformly at random

2. Create time-series: map tower cell counts to IRIS cell counts

3. Perturb these time-series to guarantee differential privacy
From CDR to Spatio-temporal Density

1. Create the Voronoi-tessellation of the towers
2. Map each Voronoi cell to Iris cells
3. Compute the IRIS cell count at any time from the count of the overlapping voronoi cells
Overview of our approach

1. Sample $x \approx 30$ visits per user uniformly at random

2. Create time-series: map tower cell counts to IRIS cell counts

3. Perturb these time-series to guarantee differential privacy
Perturbation of time series

- Naïve solution: add properly calibrated Laplace noise to each count of the IRIS cell (one count per hour over 1 week)

**Problem:** Counts are much smaller than the noise!

- Our approach:
  1. cluster nearby less populated cells until their aggregated counts become sufficiently large to resist noise.
  2. perturb the aggregated time series by adding noise to their largest Fourier coefficients
  3. scale back with the (noisy) total number of visits of individual cells to get the individual time series
Performance evaluation 1: Mean Relative Error

\[
\text{MRE}(X, \hat{X}) = (1/168) \sum_{i=1}^{168} \frac{\left| \hat{X}_i - X_i \right|}{\max(\gamma, X_i)}
\]

Naïve approach (\(\varepsilon=0.3\))

Average MRE: 1.01

Our scheme (\(\varepsilon=0.3\))

Average MRE: 0.17
Performance evaluation 2: Pearson Correlation

\[
PC(X, \hat{X}) = \frac{\sum_{i=1}^{168} (X_i - \bar{X}/168)(\hat{X}_i - \bar{\hat{X}}/168)}{\sqrt{\sum_{i=1}^{168} (X_i - \bar{X}/168)^2} \sqrt{\sum_{i=1}^{168} (\hat{X}_i - \bar{\hat{X}}/168)^2}}
\]

Naive approach (\(\varepsilon=0.3\))
Average PC: 0.47

Our scheme (\(\varepsilon=0.3\))
Average PC: 0.96
Conclusions

- secure composability is an implicit requirement in European Data Protection laws
  - this favours randomization-based notions of privacy (such as differential privacy)

- we obtained accurate results for large dimensional data within the differential privacy model

- there are no “universal” anonymization solutions that fit all applications
  - in order to get the best accuracy, they have to be customized to the application and the public characteristics of the dataset
Performance evaluation 3: Error depending on time

Relative error ($\varepsilon = 0.3$)

Earth Mover’s Distance
“Meters of errors”
($\varepsilon = 0.3$)

Average: 0.18
Average: 188 meters
European Data Protection law

- Personal data is any information relating to an identified or identifiable individual.
  - Can be used to identify him or her, and to know his/her habits.
  - Account must be taken of all the means available [...] to determine whether a person is identifiable.

- Any processing of any personal data must be (1) transparent (to the individual), (2) for specified explicit purpose(s), (3) relevant and not excessive in relation to these purposes.

- Legally nonbinding: all member states have enacted their own data protection legislation.

- Anonymized data is considered to be non-personal data, and as such, the directive does not apply to that.
“personal data is any information relating to an individual, whether it relates to his or her private, professional or public life. It can be anything from a name, a photo, an email address, bank details, posts on social networking websites, medical information, or a computer’s IP address.”

- applies if personal data of EU residents are processed (even by non-EU companies)

- **Risk assessment and mitigation is required** (checked by a Single Data Protection Authority)

- a single set of rules applies to all EU member states
American vs. European Data Protection Directive (95/46/EC)

- US has no single data protection law comparable to the EU's Data Protection law
  - ad-hoc legalisation: certain sectors partially satisfy the EU Directive, however most do not
  - this is probably due to the American “lassiez-faire” economics
- HIPAA’s Privacy Rule mainly regulates the use of medical data (PHI), which is (in theory) similarly strict as EU’s current data protection law