Anonymisation of localisation data

Arnaud Le Guilcher
IGN, COGIT

8 septembre 2016
Motivations for algorithms protecting privacy

The notion of quasi-identifier

Adaptation of $k$-anonymity to trajectories
Increased availability of data

- Large quantities of data gathered
  - with traditional forms
  - automatically with connected devices
- Data sold or made available
  - for business purposes
  - for research (health, social sciences)
A threat to privacy

- Data contain sensitive information
  - health, family, religion, purchases, travelling habits
- A threat to individual privacy
- Inference of more information is possible
Sensitivity of geo-localised data

- Geo-localised data may point to sensitive places
  - doctors, hospitals
  - religious buildings
  - schools
  - shops, restaurants, clubs
How linking can cause privacy breaches

- Linking different databases reveals new information
- The example of de-identified medical records

![Diagram showing linking of medical data and voter list](image)

How linking can lead to privacy breaches (Sweeney 2002).
Quasi-identifiers (QIDs)

- Specific values of some tuples lead to the re-identification of a record owner
  - (ZIP code, birth date, gender)
- Quasi identifier: a set of attributes for which some values point to a unique individual
- Quasi identifiers defined by the database owner
  - depend on other available data
**Principle of $k$-anonymity**

- $k$-anonymity: for a determined QID, each record must share its values with at least $k-1$ other records

<table>
<thead>
<tr>
<th>Race</th>
<th>Birth</th>
<th>Gender</th>
<th>ZIP</th>
<th>Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>Black</td>
<td>1965</td>
<td>m</td>
<td>0214*</td>
</tr>
<tr>
<td>t2</td>
<td>Black</td>
<td>1965</td>
<td>m</td>
<td>0214*</td>
</tr>
<tr>
<td>t3</td>
<td>Black</td>
<td>1965</td>
<td>f</td>
<td>0213*</td>
</tr>
<tr>
<td>t4</td>
<td>Black</td>
<td>1965</td>
<td>f</td>
<td>0213*</td>
</tr>
<tr>
<td>t5</td>
<td>Black</td>
<td>1964</td>
<td>f</td>
<td>0213*</td>
</tr>
<tr>
<td>t6</td>
<td>Black</td>
<td>1964</td>
<td>f</td>
<td>0213*</td>
</tr>
<tr>
<td>t7</td>
<td>White</td>
<td>1964</td>
<td>m</td>
<td>0213*</td>
</tr>
<tr>
<td>t8</td>
<td>White</td>
<td>1964</td>
<td>m</td>
<td>0213*</td>
</tr>
<tr>
<td>t9</td>
<td>White</td>
<td>1964</td>
<td>m</td>
<td>0213*</td>
</tr>
<tr>
<td>t10</td>
<td>White</td>
<td>1967</td>
<td>m</td>
<td>0213*</td>
</tr>
<tr>
<td>t11</td>
<td>White</td>
<td>1967</td>
<td>m</td>
<td>0213*</td>
</tr>
</tbody>
</table>

*Figure 2 Example of $k$-anonymity, where $k=2$ and QI={$Race, Birth, Gender, ZIP$}*

A $k$-anonymous table (Sweeney 2002).

- Anonymisation through generalisation and suppression
A simple model for trajectories

- A trajectory as a sequence of couples (time, position)
- Anonymisation of a database of trajectories (offline anonymisation)
- Trajectories with no labels
- Operation of generalisation: \((t, x) \rightarrow (T, X)\)
  - where \(T\) and \(X\) are sets (intervals or discrete sets)
Adapting the notion of QID

- Homogeneity of the data
- Any couple (time, position) may be a QID or part of a QID
- Any couple may contain sensitive information
- Anonymisation strategies depend on the modelling of adversary knowledge
A few anonymisation examples

Anonymisation against an attacker with local knowledge (Terrovitis et al. 2008).

Anonymisation when each trajectory has a distinct QID (Yarovoy et al. 2009).

An attack graph against unsafe $k$-anonymisation (Yarovoy et al. 2009).
A few anonymisation examples

Anonymisation using data uncertainty (Abul et al. 2008).

Anonymisation with clusters of trajectories (Nergiz et al. 2008).
Conclusion

- No unified framework to handle geo-localised data
- Anonymisation strategies depend on
  - privacy criteria
  - models for adversary knowledge
  - purposes of the released data
Abul, O., Bonchi, F., and Nanni, M. Never Walk Alone : Uncertainty for anonymity in moving objects databases. In Proc. of the 24nd IEEE Int. Conf. on Data Engineering (ICDE’08).


Terrovitis, M., and Mamoulis, N. Privacy preservation in the publication of trajectories. In Proc. of the 9th Int. Conf. on Mobile Data Management (MDM’08).