Privacy-preserving publication of complex data

Grigorios Loukides
Dept. of Informatics
King’s College London
grigorios.loukides@kcl.ac.uk

Sensing and Privacy workshop, QMUL, London
20 June, 2017
Big Data and complex data

- **Characteristics of Big Data**
  - Volume, Velocity, **Variety**, Veracity, **Value**

- **Big Data in data mining**
  - Explores complex, evolving relationships among data
    (HACE Theorem [Wu et al. TKDE’13])

- **Complex data**
  - Movement data (sensors, social-network checkins)
  - Marketing data (purchases over time)
  - Health data (activities, diagnoses)
Complex data publishing

Goal: Share individuals’ data for analysis and mining

- **Utility**
  - Location-based services with movement data
  - Personalized services, advertising with marketing data
  - Medical services (ADL) with health data

- **Privacy**
  - Prevent re-identification and/or sensitive location inference
  - Prevent mining of sensitive knowledge
  - Prevent inference of health profile

- **Maximize utility subject to privacy. But,**
  - Data are high dimensional (classic distance becomes meaningless)
  - Data are large (need efficient processing)
  - Problems are difficult (computationally hard and inapproximable)
Focus of the presentation

- Trajectory data anonymization\(^1,2\)
- Event sequence sanitization\(^3\)
- IoT data privacy monitoring

\(^1\)https://doi.org/10.1109/ICDMW.2013.136
\(^2\)http://dl.acm.org/citation.cfm?id=2870625
\(^3\)https://doi.org/10.1137/1.9781611974010.87
Trajectory data anonymization: Setting (1/2)

- **Trajectory dataset**
  - sequence of locations per user, thousands of users
  - collected from sensors, phones, social apps
  - shared for analysis (queries, mining, visualization)

<table>
<thead>
<tr>
<th>Id</th>
<th>Name</th>
<th>trajectory</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>Mary</td>
<td>$(d, a, c, e)$</td>
</tr>
<tr>
<td>$t_2$</td>
<td>Jim</td>
<td>$(b, a, e, c)$</td>
</tr>
<tr>
<td>$t_3$</td>
<td>Anne</td>
<td>$(a, d, e, f)$</td>
</tr>
<tr>
<td>$t_4$</td>
<td>Nick</td>
<td>$(b, d, e, c)$</td>
</tr>
<tr>
<td>$t_5$</td>
<td>Mark</td>
<td>$(d, g, c)$</td>
</tr>
<tr>
<td>$t_6$</td>
<td>Helen</td>
<td>$(d, e)$</td>
</tr>
</tbody>
</table>
Privacy threats

- Re-identification: users linked to their anonymous profiles
- Inference of sensitive locations
  - Revealing health issues, religion & political orientation

Example

<table>
<thead>
<tr>
<th>Id</th>
<th>Name</th>
<th>trajectory</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>Mary</td>
<td>$(d, a, c, e)$</td>
</tr>
<tr>
<td>$t_2$</td>
<td>Jim</td>
<td>$(b, a, e, c)$</td>
</tr>
<tr>
<td>$t_3$</td>
<td>Anne</td>
<td>$(a, d, e, f)$</td>
</tr>
<tr>
<td>$t_4$</td>
<td>Nick</td>
<td>$(b, d, e, c)$</td>
</tr>
<tr>
<td>$t_5$</td>
<td>Mark</td>
<td>$(d, g, c)$</td>
</tr>
<tr>
<td>$t_6$</td>
<td>Helen</td>
<td>$(d, e)$</td>
</tr>
</tbody>
</table>

- John knows that Anne visited a cinema ($a$) and then a restaurant ($d$)
- Anne linked to $t_3$ and visited clinic ($f$) with prob. 1
Trajectory data anonymization: Goal

Problem (at a very high level)

Given a trajectory dataset, set of sensitive locations, and a distortion measure, transform the trajectory dataset, so that

- sufficiently low re-identification probability
- sufficiently low prob. of inferring any subsequence of sensitive locations
- minimum distortion (maximum usefulness)

Original dataset $T$

<table>
<thead>
<tr>
<th>Id</th>
<th>trajectory</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>($d, a, c, e$)</td>
</tr>
<tr>
<td>$t_2$</td>
<td>($b, a, e, c$)</td>
</tr>
<tr>
<td>$t_3$</td>
<td>($a, d, e, f$)</td>
</tr>
<tr>
<td>$t_4$</td>
<td>($b, d, e, c$)</td>
</tr>
<tr>
<td>$t_5$</td>
<td>($d, g, c$)</td>
</tr>
<tr>
<td>$t_6$</td>
<td>($d, e$)</td>
</tr>
</tbody>
</table>

Anonymized dataset

<table>
<thead>
<tr>
<th>Id</th>
<th>trajectory</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1'$</td>
<td>({d, a, c, e}, {d, a, c, e})</td>
</tr>
<tr>
<td>$t_3'$</td>
<td>({d, a, c, e}, {d, a, c, e}, {d, a, c, e}, $f$)</td>
</tr>
<tr>
<td>$t_6'$</td>
<td>({d, a, c, e}, {d, a, c, e}, {d, a, c, e}, {d, a, c, e})</td>
</tr>
<tr>
<td>$t_4'$</td>
<td>({a, b, d}, {a, b, d}, e, c)</td>
</tr>
<tr>
<td>$t_2'$</td>
<td>({a, b, d}, {a, b, d}, e, c)</td>
</tr>
<tr>
<td>$t_5'$</td>
<td>({a, b, d}, g, c)</td>
</tr>
</tbody>
</table>
Transformation operation
- Replace a location with a set of locations
  \[ a \rightarrow \{a, b, c\} \] (interpreted as any of the locations)

Utility measures
- Based on geographical and/or semantic distance, to find similar trajectories
Privacy principle

\((k, \ell)^m\)-anonymous set of trajectories

- Prob. of re-identification using any sequence \(s\) of at most \(m\) nonsensitive locations bound by \(\frac{1}{k}\)
- Prob. of inferring any sequence of sensitive locations \(s'\), given \(s\), bound by \(\frac{1}{\ell}\)

Efficient algorithms: Select similar trajectories, group them together, anonymize them

- ZGA: Z-order space filling curves
- SGA: Projection on frequent sequential patterns
- SeqAnon: Lattice-based search
Trajectory data anonymization: Results

- **Utility: Gowalla dataset**

- **Runtime (secs): Oldenburg dataset**
Data stream sanitization: Setting

- **Data stream: event sequence**
  - thousands of time points and a segmental structure
  - featured in applications such as marketing, web analysis, and medicine

\[
\begin{align*}
\{a.1, b.4, c.5\}, t_1 & \quad \{a.2, b.4, c.7, d.7\}, t_2 & \quad \{a.8, b.7, c.1, d.4\}, t_3
\end{align*}
\]

- **Frequent event mining**
  - find events with relative frequency \( \geq \delta \) in any sequence prefix
  - fundamental for analyzing event sequences

  - may expose **sensitive events** that represent confidential knowledge
Data stream sanitization: Example

**BigMart**
- Collects purchases from customers
- Transforms them into event sequence

\[
\{(a.1, b.4, c.5), t_1\} \quad \{(a.2, b.4, c.7, d.7), t_2\} \quad \{(a.8, b.7, c.1, d.4), t_3\} \quad \ldots
\]

**MarketResearch**
- Customer profiling
- Fraud or change analysis
- Trend analysis
- Mines frequent events that give competitive advantage and sells it to BigMart's competitors

- Financial and reputational loss to organizations & unwarranted public concern
Data stream sanitization: Contributions

The first approach to sanitizing event sequences.

- **Utility model** for sanitized event sequences

- Impact of deletion on the probability distribution of events

- **Definition** of the *Event Sequence Sanitization* (ESS) problem

---

**Event Sequence Sanitization**

Given an event sequence, a set of sensitive events, and threshold $\delta$, construct a sanitized event sequence $D'$ s.t.:

(I) no sensitive event has relative frequency $\geq \delta$, in *any prefix*

(II) $D'$ has optimal utility, according to the model.

---

**Optimal algorithms** for the ESS problem: *ODESA* and *ESSA*
Data stream sanitization: Overview

- **Dynamic programming:** for selecting occurrences to delete

- **Optimal error:** accurate frequent event mining
Privacy evaluation of IoT devices and applications

**Alarming findings:** out of 314 IoT devices
- 59% didn’t explain how personal inf. was collected, used, disclosed
- 68% didn’t explain how user information was stored
- 72% didn’t explain how users can delete their information off the device

**Standards and guides** (OWASP, GSM, OneM2M)
- How much personal information is collected?
- Are personal data encrypted at rest and/or in transit?
- Has data been de-identified and/or anonymized?
- Could user opt-out from the collection of unnecessary data for the device’s operation?

---

*GPEN privacy sweep 2016 https://www.privacyenforcement.net/node/717*
Different from the standard data publishing setting
- Data of different formats and come as streams
- Privacy threats need to be discovered, explained, prevented

Research issues
- How to discover privacy issues on the fly from Big Data Streams?
- How to do that without breaching the privacy of IoT device users?
- How to involve users in the loop?

How to deal with the issues?
- Project starts in August
Conclusions

- **Big Data are often complex and high-dimensional**
  - trajectories
  - event sequences
  - combinations of them (demographics and sequential data)
  - graphs

- **Protecting the privacy of complex data is not easy**

- **Overview of some of our current and future work in the area**