Privacy: re-identification and Inference Risks

Paris—26 Feb. 2018
Innocent looking data may pose risk

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Possible Inferences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instantaneous electricity consumption</td>
<td>Awake from 3am to 5am every day? Cooking at 11pm?</td>
</tr>
<tr>
<td>Water consumption</td>
<td>Frequent bathroom use?</td>
</tr>
<tr>
<td>Phone metadata</td>
<td>Reveals social network + habits</td>
</tr>
<tr>
<td>Google searches</td>
<td>Reveals +/- everything?</td>
</tr>
<tr>
<td>Phone/car localization</td>
<td></td>
</tr>
<tr>
<td>Credit card use</td>
<td></td>
</tr>
</tbody>
</table>

➡️ Risk of inference, deduction etc.

+ wrong side of big data:
Ex: buying coffee at McDonalds every day at around lunch time.
Take home message on privacy risk

Even « safe-looking » datasets may pose privacy issues

In particular,

- Removing names ≠ anonymizing
- Anonymity ≠ Privacy (inference risk)
- Combination of « safe » datasets may lead to privacy risk
Example: «anonymous» medical data

During the 90’s GIC, health insurance organism for Massachusetts state employees collected data on medical treatments

<table>
<thead>
<tr>
<th>Birth-date</th>
<th>ZIP</th>
<th>gender</th>
<th>Date visit</th>
<th>Diagnostic</th>
<th>(...)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31 Jul 45</td>
<td>02141*</td>
<td>male</td>
<td>(...)</td>
<td>Cancer</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data «anonymous»

→ GIC shared it with researchers, sold it to companies

**Problem:** publically available voter register contains names, birth date and zip-code of majority of americans

* Only one male in 02141 born on 31 Jul 45, William Weld, governor of Massachusetts!*
Inference: anonymity not sufficient

Hospital record for a day on which an employee born in 1976 seen by boss at the hospital:

<table>
<thead>
<tr>
<th>Birth</th>
<th>condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>76-80</td>
<td>Alcoholism</td>
</tr>
<tr>
<td>76-80</td>
<td>Severe psychiatric issue</td>
</tr>
<tr>
<td>76-80</td>
<td>Serious memory loss</td>
</tr>
<tr>
<td>76-80</td>
<td>Minor sport injury</td>
</tr>
<tr>
<td>76-80</td>
<td>Terminal illness</td>
</tr>
<tr>
<td>81-85</td>
<td>...</td>
</tr>
<tr>
<td>81-85</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Re-Identification impossible,
But employee might not get new responsibilities soon

No particularly serious condition, but on the "wrong side of big data"

can happen in various contexts!
Netflix dataset

<table>
<thead>
<tr>
<th></th>
<th>Titanic</th>
<th>Starwars</th>
<th>Primer</th>
<th>Lion King</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>User 1</td>
<td>7-11-04, 2*</td>
<td></td>
<td>8-4-03, 5*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>User 2</td>
<td>4-6-05, 4*</td>
<td></td>
<td>3-2-04, 2*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Private sensitive information? YES!*
- Correlation with sensitive info (sexual orientation, religion, mood...)
- Movies not consistent with “external image”
- Unusual watching behavior

*Publicly available information? YES!*
- Chatting about movie seen recently
- Rating/comments on certain movies on other websites (IMDB...)

But, no clear separation sensitive / non-sensitive, public / private...
Combination of safe datasets may be unsafe!

- **Netflix data**: «safe» because anonymous, even if contains sensitive information
- **IMDB dataset**: «safe» because no sensitive information, even if not anonymous

*But Netflix + IMDB unsafe:*

Public information in IMDB data also in Netflix
→ link between IMDB profile and anonymous Netflix profile
→ Link between IMDB identity and Netflix sensitive information

Anonymization more challenging! (except if movies anonymous)
Need to take *other existing and future datasets* into account
Solutions?

• Precision degradation
  But, for « wide » datasets (lot of info about every person), need to remove almost all useful information to guarantee privacy

• Binning: grouping of people
  But, either destroy correlations, or risk of re-identification/inference
Solutions?

Queries on confidential datasets
(OPen ALgorihtm initiative)

- Dataset remains in trusted hands, no public access
- Specific questions can be sent using publically available code
  → control of privacy risk and potential abuse

Issues

- Who to trust with dataset? Leak risk? Should it be decentralized?
- How to check privacy risk based on questions?
Questions – Food for Thought

• Is this fear a generation thing?
  Younger people often seem not to care.
  ➔ are they naive? Do they have a different view on privacy?

• Is the risk really new?
  – Discrimination was always present,
    (gossips, looks, habits, jobs etc.)
  – Is this new form worse?

• Big data offer countless opportunities in all domains.
  – Do we want to decline them?
  – How much privacy are we willing to sacrifice for personal comfort?
  – For common good?