Privacy: re-identification and Inference Risks

Paris– 26 Feb. 2018

Innocent looking data may pose risk

Instantaneous electricity consumption	Awake from 3am to 5am every day? Cooking at 11pm?
Water consumption	Frequent bathroom use?
Phone metadata	Reveals social network + habits
Google searches	Reveals +/- everything?
Phone/car localization	
Credit card use	

➔ 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

Birth-date	ZIP	gender	Date visit	Diagnostic	(…)
…					
31 Jul 45	02141*	male	(…)	Cancer	
…					

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!

* approximately

Image Source: wikipedia

Inference: anonymity not sufficient

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

Birth	condition
76-80	Alcoholism
76-80	Severe psychiatric issue
76-80	Serious memory loss
76-80	Minor sport injury
76-80	Terminal illness
81-85	…
81-85	
…	

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

	Titanic	Starwars	Primer	Lion King	…
User 1	7-11-04, 2*			8-4-03, 5*	
User 2		4-6-05, 4*		3-2-04, 2*	

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 ALgorithm 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 Tought

- 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?