icteam

UCL Université catholique de Louvain

Data Analytics and the Re-identification and Inference Risks

Julien Hendrickx

Brussels – 13 May 2016

Large datasets more and more available

- Open data initiatives:
  data.gov, G8 open data charter,

- Commercial agreement with third parties

- Leaks: customer data (verizon, SNCB, Ashley Madison, PlayStation …), wikileaks, panama leaks, etc.

  → *caution advised even if dataset never meant to be disclosed*

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	(…)	(condition)	
…					

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

* approximately

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	(…)	(condition)	
…					

D

P irth

d

Only one male in 02141 born on 31 Jul 45, William Weld, governor of Massachusetts!

→ Latanya Sweeney, PhD student at MIT, sent the governor his whole medical record

* approximately

Image Source: wikipedia

Identifier

Birth date, ZIP code and gender

- Typically available (especially if you know the person)
- uniquely identify most Americans,

→ Removing name but keeping zip, birth date and gender ***does not anonymize data***

To improve privacy while keeping data useful: be more vague

Birth date	ZIP	gender	Date visit	Diagnostic	(…)
…					
31 Jul 45	02141	male	(…)	(condition)	
…					

Use of side information

To improve privacy while keeping data useful: be more vague

Birth date	ZIP	gender	Date visit	Diagnostic	(…)
…					
31 Jul 45	02141	male	(…)	(condition)	
…					

Probably 150-300 people matching birth and zip information
Should be anonymous, but

- Only state employees appear in datasets
- If known that governor has been in hospital in a certain period…, or is in the dataset…

Measure of anonymity (example)

k-anonymity: For every person appearing in the datasets, at least k-1 others share the same « public information »

→ identification of at best a k-person group

Obtained e.g. by making data less precise, and removing outliers

Birth date	condition
Jan 1984	
Feb 1988	
Oct 1975	
Dec 1972	
Feb 1914	
Mar 1989	
Oct 1976	

Birth date	condition
Jan **198***	
Feb **198***	
Oct **197***	
Dec **197***	
Feb 1914	
Mar **198***	
Oct **197***	

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 will probably not get new responsibilities soon.
From the boss point of view, 80% chance the employee has a serious problem

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 will probably not get new responsibilities soon

No particularly serious condition, but on the ***« wrong side of big data »***

can happen in various contexts!

Summary so far

Dataset consisting of two sets of columns

public non-sensitive info: **Birth, zip code, etc…**	***Private, potentially sensitive info:*** health condition, shopping habits, political opinion …

Privacy challenge: knowing public information should not allow

- ***re-identification***
- ***Too much inference*** about individual private info (l-diversity etc.)

Even if limited private info available (ex: arm not broken),

while allowing *as much general inferences as possible*
(otherwise dataset useless)

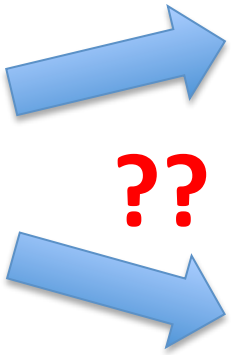
Tools: binning, coarsening, "noise", removal of outliers ….

Trade-off privacy vs information

Improving privacy destroys information

→How to keep as much as possible?, What is « information » ?
→Which information to keep?
→What will be relevant for end-use ? Unknown future research?

Birth date	sex	condition
Jan 92		
Feb 88		
Oct 88		
Dec 92		



Birth date	sex	condition
Jan 92		
Feb 88		
Oct 88		
Dec 92		

Birth date	sex	condition
Jan-Mar		
Jan-Mar		
Oct-Dec		
Oct-Dec		

Homogeneous data: Netflix

Not all datasets can be separated in public – private information

Netflix Prize up to 1M$ in contest to improve recommandation algorithm, in 2006-2008

Dataset: 500k « anonymous » users, 17k movies, 100M ratings (1-5)
(= 10% of dataset for 1999-2005)

Goal: best prediction of some other ratings

FAQ : Is there any customer information in the dataset that should be kept private ?

Answer: *No, all customer identifying information has been removed; all that remains are ratings and dates. (…)*

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 movies seen recently
- Rating/comments on certain movies on other websites (IMDB…)

But, no clear separation sensitive / non-sensitive, public / private…

Netflix de-anonymization

[Narayanan & Shmatikov 08]. **Robust de-anonymization of large sparse datasets**, 2008 (preprint 2006):

- **99%** users identifiable with **8 ratings** (even if errors) and dates up to **14 days errors**
- **68%** users identifiable with **2 ratings** and dates up to **3 days errors**
- Significant part can be identified without dates, especially if movies not blockbuster

+ re-identification by exploiting public ratings on IMDB

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:

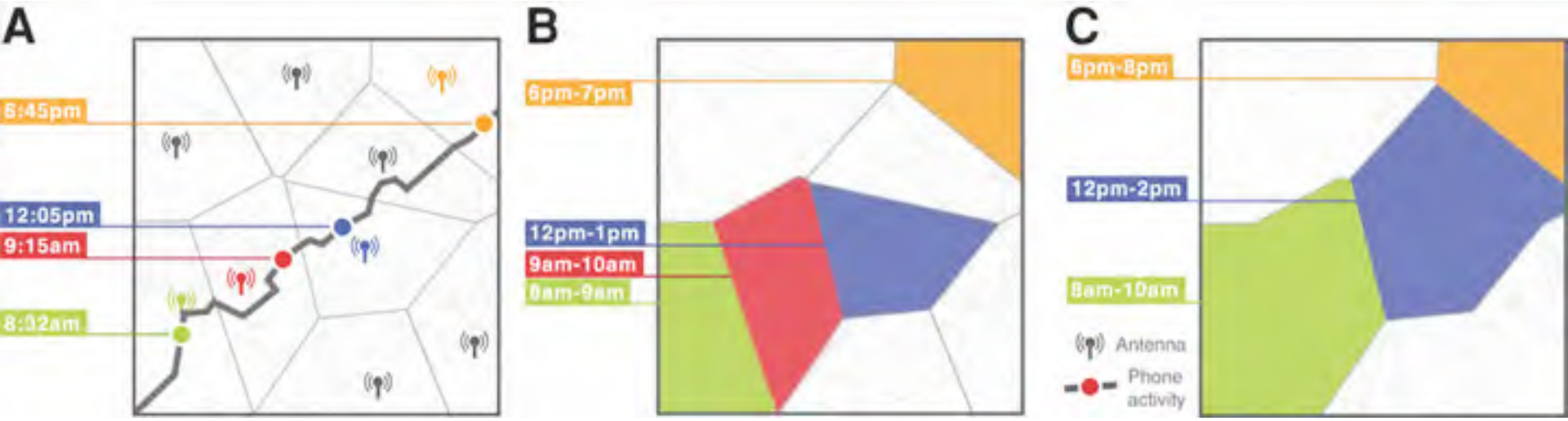
Nonsensitive information in IMDB data also in netflix
→ link between IMDB profile and anonymous netflix profile
→ Link between IMDB identity and netflix sensitive information

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

Phone localization data

Phone operators record information about call made, including tower to which cellphone is connected

→ ***Data about people localization***



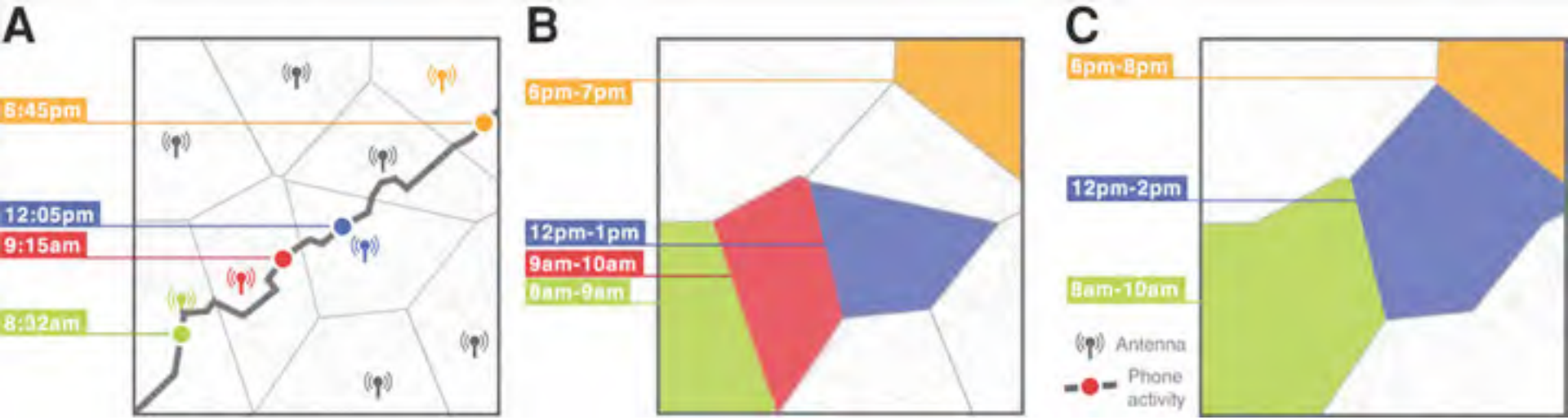
Lot to be learned: commuting patterns, census-like information etc.

→ Improved policy/commercial decisions,(including in real time)

But how about privacy?

Figure from *[de Montjoye et al, Nature sc report 13]*

Localization and privacy



(1 antenna: 0.15km^2 → 15km^2)

Some information public, some information private and sensitive
Again, no simple way to make the distinction

Figure from *[de Montjoye et al, Nature sc report 13]*

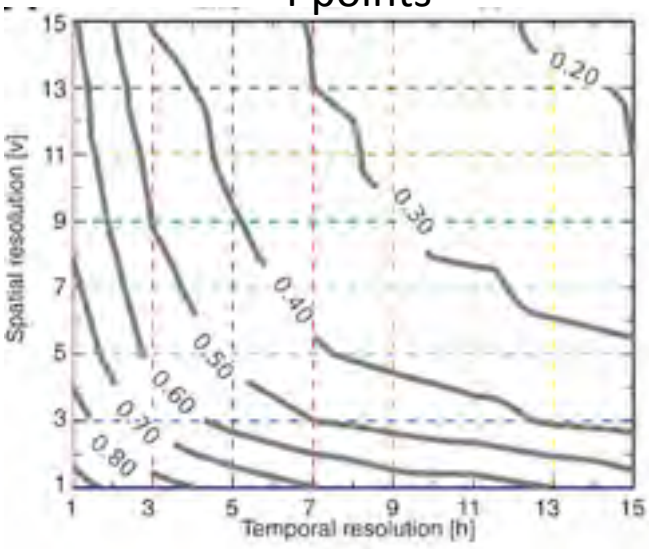
Localization and privacy

[de Montjoye, et al 13] **Unique in the crowd: The privacy bounds of human mobility**. *Scientific reports*, *3*.

Dataset with 1.5M users, in some European country, 15 months.

- Precision 1hr -1 antenna → ***95% identification with 4 datapoints***
- Decreasing identification probability requires strong precision degradation (explicit formula provided)

4 points



10 points

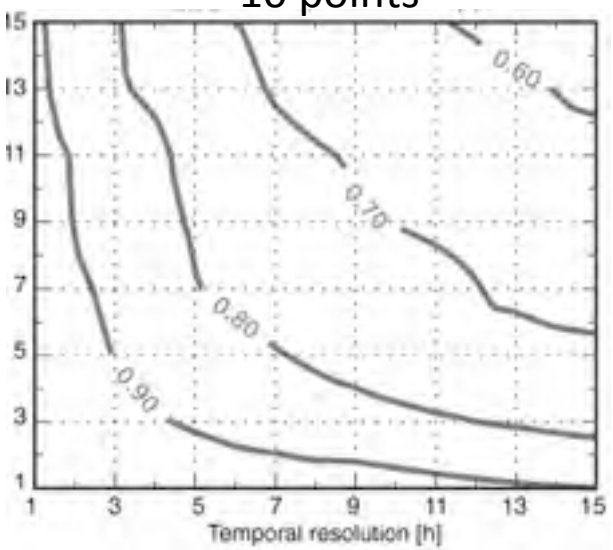


Figure from *[de Montjoye et al, Nature sc report 13]*

Conclusions

Risks of large datasets:

- Removing names ≠ anonymizing
- Anonymity ≠ Privacy (inference risk)
- Combination of « safe » datasets may lead to privacy risk
  → *need to take other (existing and future) datasets into account*!

A few examples presented, problems arise in many other situations

Conclusions

But ***huge potential (and actual) benefits*** (research, commercial interest, government…), and ***large datasets unlikely to disappear***

➔ Need to raise awareness about

- Risk
- Privacy enhancing processing (coarsening etc…) *when applicable*
- Otherwise, enhanced computer security and/or facing consequences of data release/privacy breach

Conclusions

- Trade-off ***privacy vs information content***

Challenge: measuring “info content”, a priori,

In open data context: no way of knowing which information will be useful

➔ Maybe need for a different model where dataset is not released, but tests can be ran.

icteam

UCL
Université
catholique
de Louvain

Questions?

julien.hendrickx@uclouvain.be