(Big) Data Anonymization

Claude Castelluccia

Inria, Privatics

BIG DATA: The Risks

- Singling-out/ Re-Identification:
  - ADV is able to identify the target's record in the published dataset… from some know information
- Attribute Inference
  - ADV can infer private/sensitive attributes from released dataset
  - Because of cross-attributes and cross-users correlation!
  - Example:
    - a dataset reveals that all users who went to points A, B, C, also went to D (for example an hospital).
    - I know that Target was at A, B, C… i can then infer that Target was also in D!
  - Note: Target does not even have to be part of the published datasets (in this case this is a guess).

BIG DATA

The Risks of Re-identification: The AOL Case

- Possible Privacy Breach
  - Examples: AOL, Netflix, ..
- In 2006, AOL released 20 million search queries for 650.000 users
- « Anonymized » by removing AOL id and IP address
- Easily de-anonymized in a couple of days by looking at queries

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to her.

AOL removed the search data from its site over the

Why is is sensitive?

- AOL user 17556639:
- *how to kill your wife*
- *pictures of dead people*
- *photo of dead people*
- *car crash photo*

BIG DATA
The Risks of Inference: The Target Case

- Target identified about 25 products that, when analyzed together, allowed him to assign each shopper a “pregnancy prediction” score.
- More important, he could also estimate her due date to within a small window
- Target could (and does) send coupons timed to very specific stages of her pregnancy.

TARGET

Source: How Companies Learn Your Secrets, NYTimes, Feb. 2012

Data Anonymization/De-Identification

- Anonymisation is NOT pseudo-anonymization!
- What is Pseudo-Anonymization?
  - *Personal information contains identifiers, such as a name, date of birth, sex and address. When personal information is pseudonymised, the identifiers are replaced by one pseudonym. Pseudonymisation is achieved, for instance, by encryption of the identifiers in personal data.*
- What is Anonymization?
  - *Data are anonymised if all identifying elements have been eliminated from a set of personal data (all quasi-identifiers). No element may be left in the information which could, by exercising reasonable effort, serve to re-identify the person(s) concerned. Where data have been successfully anonymised, they are no longer personal data.*

Source: **Handbook on European data protection law,**

Pseudo-Ano. vs Anonymization (2)

- Why does Pseudo-Anonymization does not work?
  - It does not compose, i.e. several Pseudo-Anonymized data can be combined to de-anonymize…
  - External Information can also be exploited.
  - See Recent example with NY Taxi
    - **173 million** individual trips de-anonymized*
- Why is Data Anonymization Difficult?
  - Difficult (impossible) to identify all quasi-identifiers!

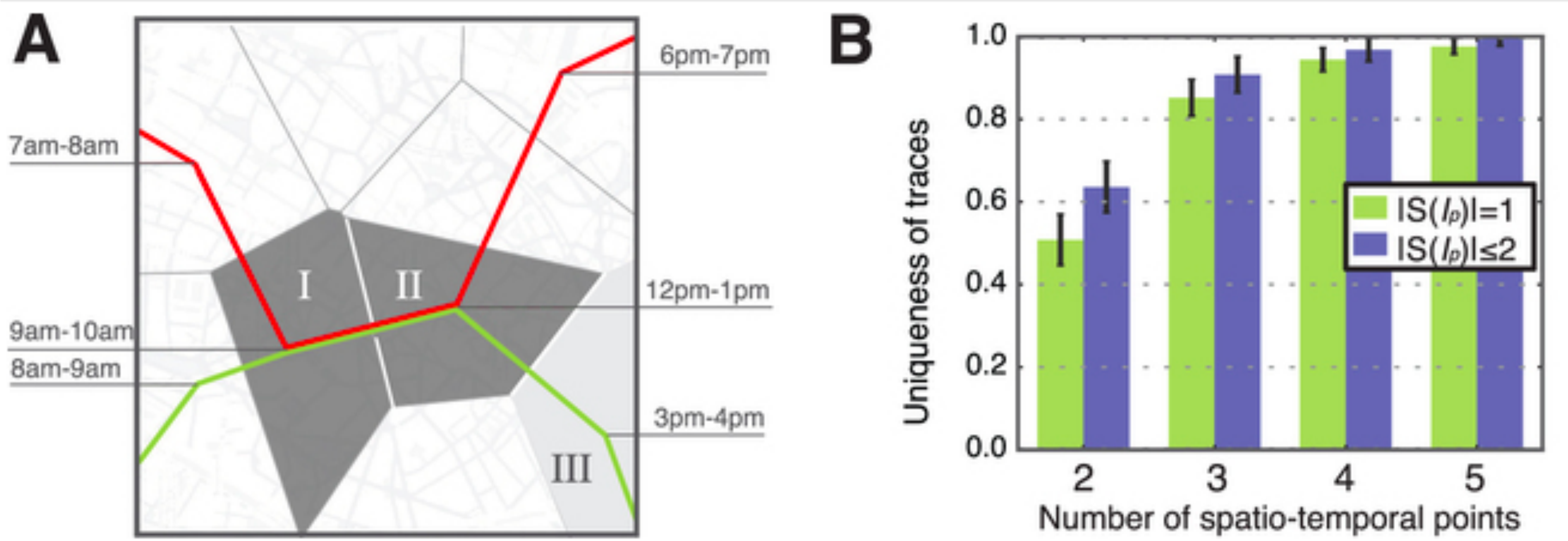
*source: On Taxis and Rainbows https://medium.com/@vijayp/of-taxis-and-rainbows-f6bc289679a1

Quasi-Identifiers

- Quasi-identifiers are difficult to identify exhaustively
- Many combination of attributes can be used to « single-out » a user
- We are all unique by different ways, we are full of Q.I.
  - See « Unicity me! *»
  - Mobility pattern, webhistory, .
  - Data (content) and meta-data
    - i.e. timing can betray you!
    - Google search timing pattern can tell when you were away!

*Unicity Me! American Scientific, http://www.americanscientist.org/libraries/documents/20142614253010209-2014-03CompSciHayes.pdf

Unique in the Crowd [Nature2013]



- Only 4 spatio-temporal points are necessary to uniquely identify a user with a probability > 95% !

Some Data anonymization methods…

- Random perturbation
  - Input perturbation
  - Output perturbation
- Generalization
  - The data domain has a natural hierarchical structure.
- Suppression
- Permutation
  - Destroying the link between identifying and sensitive attributes that could lead to a privacy leakage.

K-anonymity

- Privacy guarantee : in each group of the sanitized dataset, each invidivual will be identical to a least $k - 1$ others.
- Reach by a combination of generalization and suppression.
- Example of use : sanitization of medical data.

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Figure 1. Inpatient Microdata

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Figure 2. 4-anonymous Inpatient Microdata

But K-Ano. does not compose ☹!

- **Question** : suppose that Alice's employer knows that she is 28 years old, she lives in ZIP code 13012 and she visits both hospitals. What does he learn?

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	130**	≥ 40	*	Cancer
6	130**	≥ 40	*	Heart Disease
7	130**	≥ 40	*	Viral Infection
8	130**	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

(a)

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<35	*	AIDS
2	130**	<35	*	Tuberculosis
3	130**	<35	*	Flu
4	130**	<35	*	Tuberculosis
5	130**	<35	*	Cancer
6	130**	<35	*	Cancer
7	130**	≥ 35	*	Cancer
8	130**	≥ 35	*	Cancer
9	130**	≥ 35	*	Cancer
10	130**	≥ 35	*	Tuberculosis
11	130**	≥ 35	*	Viral Infection
12	130**	≥ 35	*	Viral Infection

(b)

But K-ANO does not compose ☹!

- **Question** : suppose that Alice's employer knows that she is 28 years old, she lives in ZIP code 13012 and she visits both hospitals. What does he learn?

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	130**	≥40	*	Cancer
6	130**	≥40	*	Heart Disease
7	130**	≥40	*	Viral Infection
8	130**	≥40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

(a)

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<35	*	AIDS
2	130**	<35	*	Tuberculosis
3	130**	<35	*	Flu
4	130**	<35	*	Tuberculosis
5	130**	<35	*	Cancer
6	130**	<35	*	Cancer
7	130**	≥35	*	Cancer
8	130**	≥35	*	Cancer
9	130**	≥35	*	Cancer
10	130**	≥35	*	Tuberculosis
11	130**	≥35	*	Viral Infection
12	130**	≥35	*	Viral Infection

(b)

Other Attacks on k-Anonymity

- k-Anonymity does not provide privacy if
  - Sensitive values in an equivalence class lack diversity
  - The attacker has background knowledge

Homogeneity attack

Bob	
Zipcode	***Age***
47678	27

Background knowledge attack

Carl	
Zipcode	***Age***
47673	36

A 3-anonymous patient table

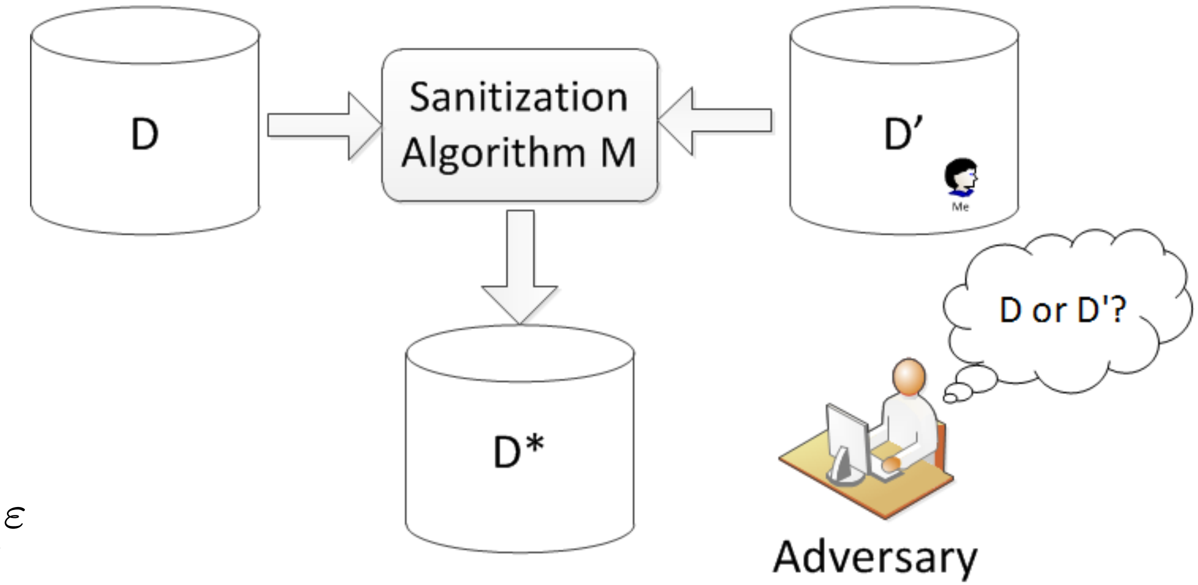
Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	⩾40	Flu
4790*	⩾40	Heart Disease
4790*	⩾40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

Approche X-DATA

- Développer des Algorithmes d’Anonymisation plus « surs »
- Développer une méthodologie d’analyse de risques des données anonymisées
- Rencontrer régulièrement et Collaborer avec la CNIL

Toward Secure Privacy Model

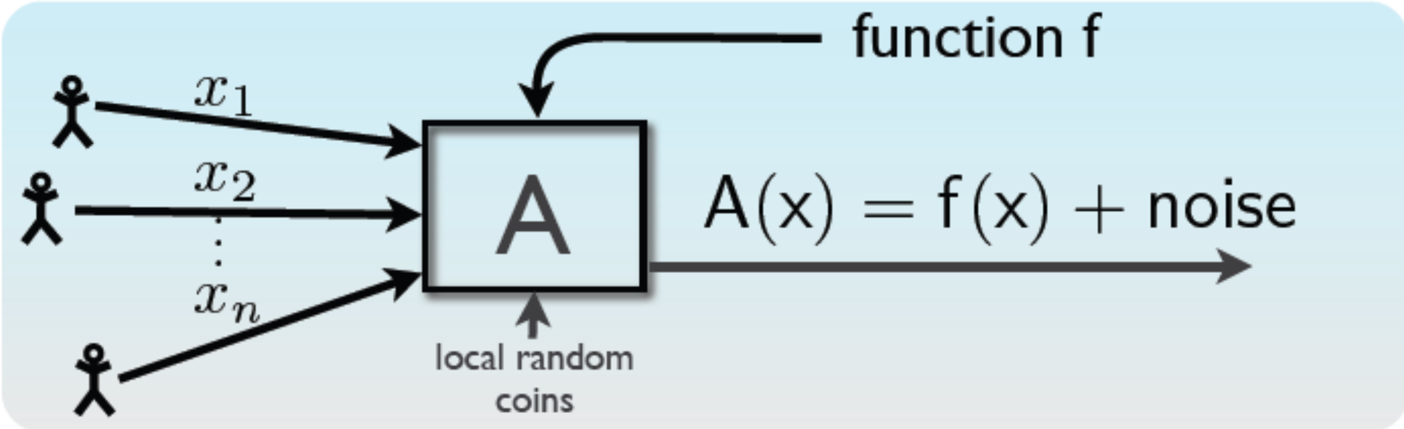
- **Differential privacy**



$$e^{-\varepsilon} \leq \frac{\Pr(M(D) = D^*)}{\Pr(M(D') = D^*)} \leq e^{\varepsilon}$$

- Secure even with arbitrary external knowledge!
- Compose
- Intuition of DP: Output is "independent" of my data

Differential Privacy

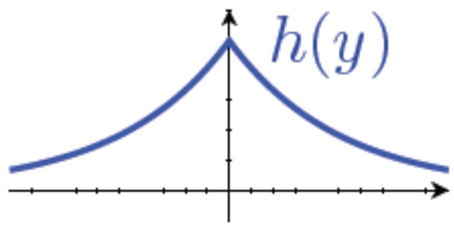


- **Global Sensitivity:** $GS_f = \max_{\text{neighbors } x, x'} \|f(x) - f(x')\|_1$
  - Example: $GS_{\text{proportion}} = \frac{1}{n}$

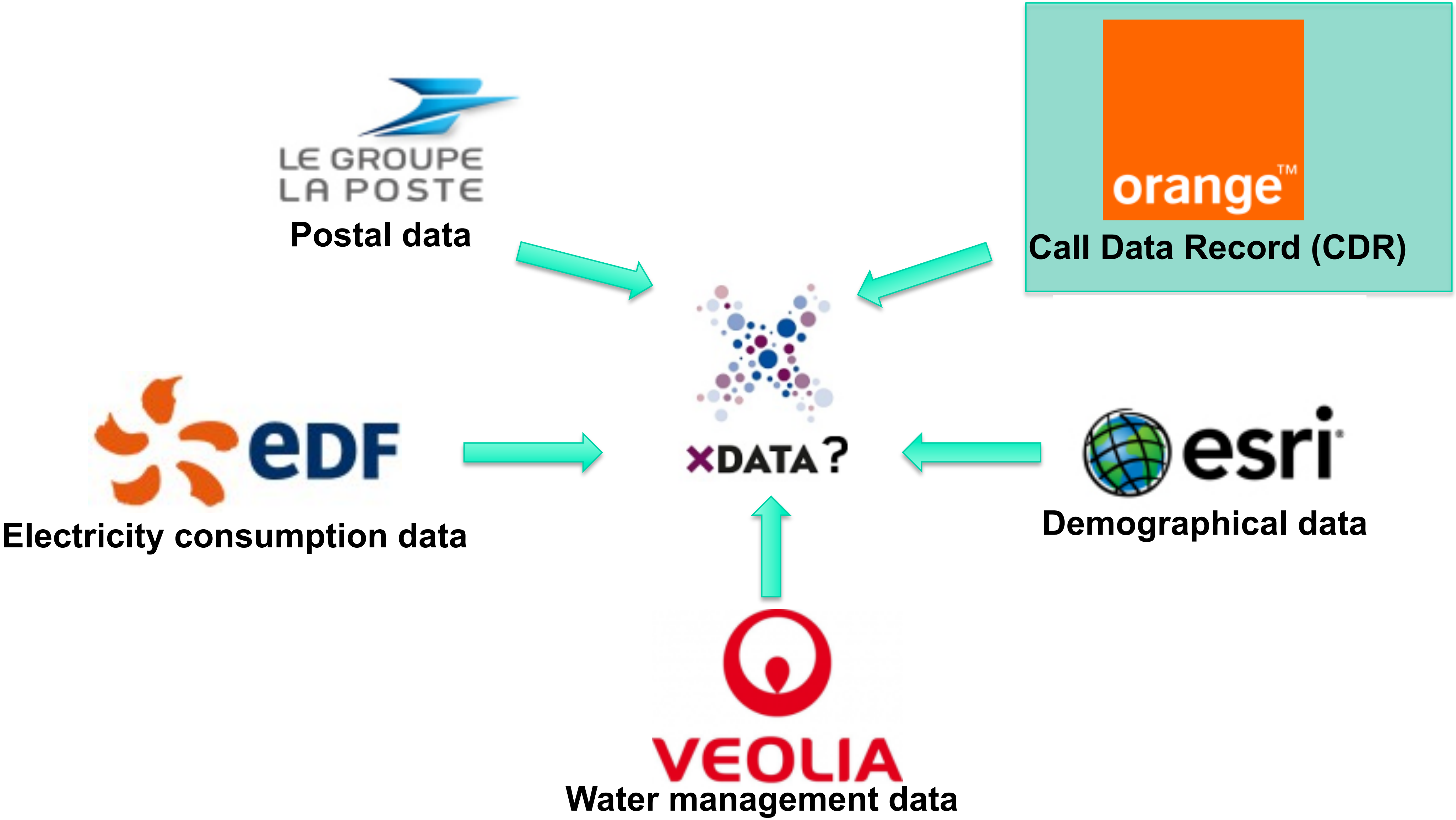
Theorem: If $A(x) = f(x) + \mathrm{Lap}\left(\frac{GS_f}{\epsilon}\right)$, then A is ϵ-differentially private.

- Laplace distribution $\mathrm{Lap}(\lambda)$ has density

$$h(y) \propto e^{-|y|/\lambda}$$

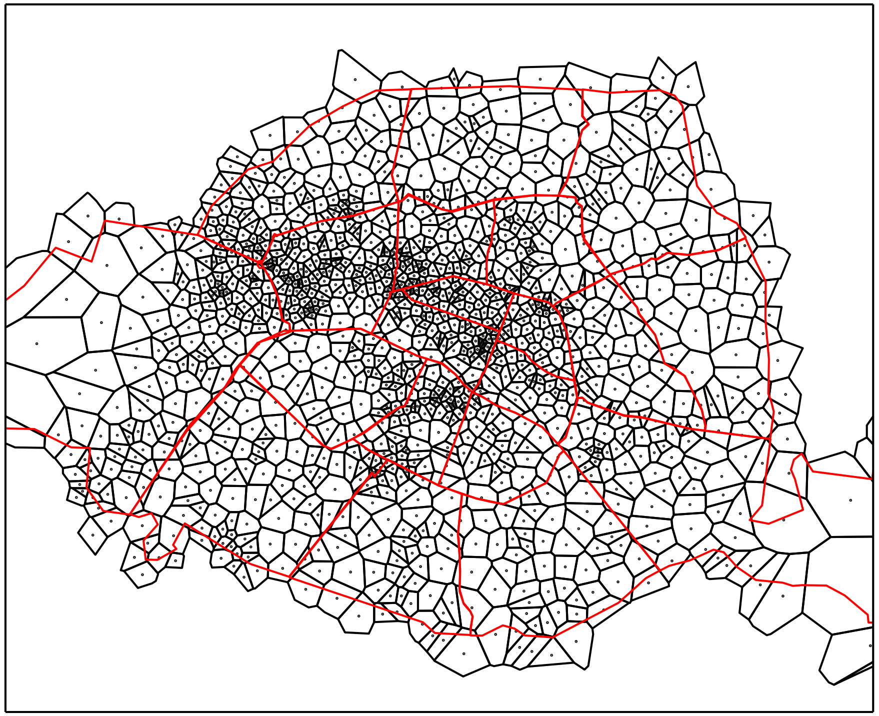


Example: Spatio-temporal density from CDR



Paris CDR (provided by Orange™)

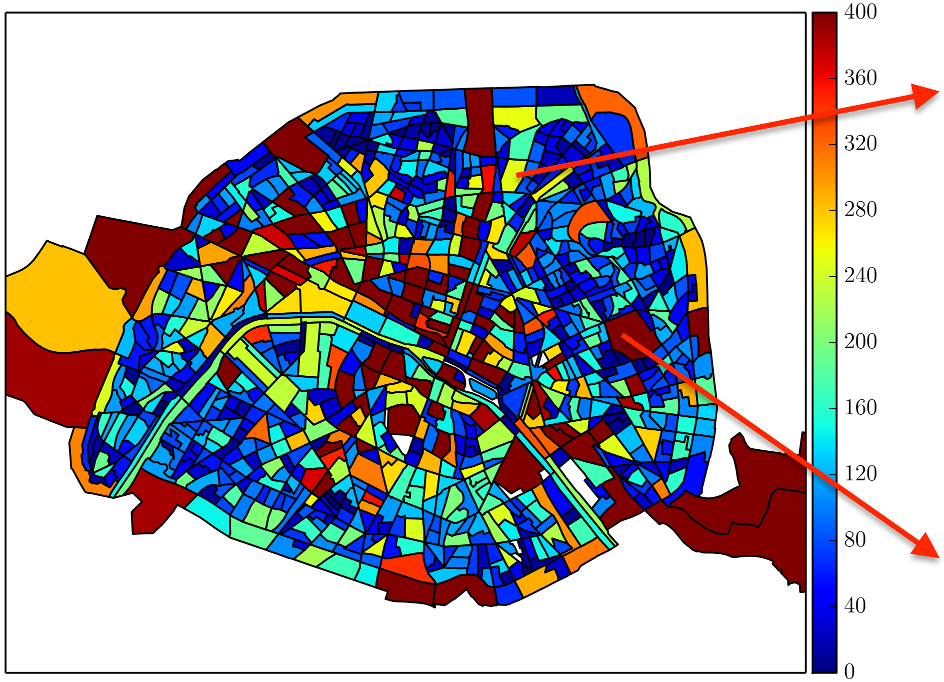
- 1,992,846 users
- 1303 towers
- 10/09/2007 – 17/09/2007
- Mean trace length: 13.55 (std.dev: 18)
- Max. trace length: 732

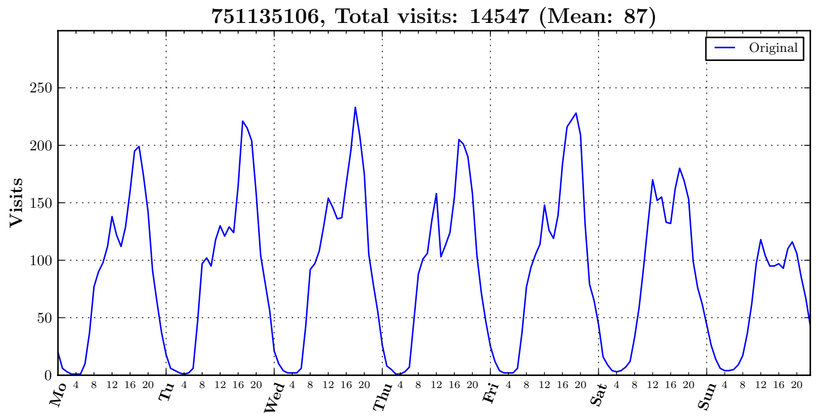


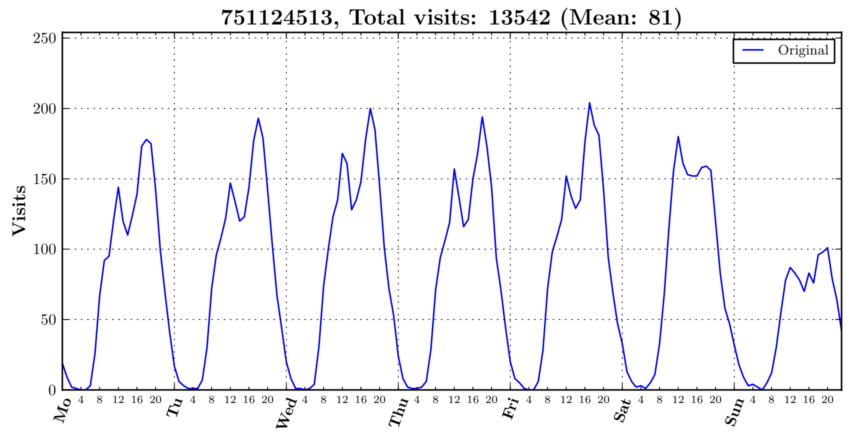
Goal: Release spatio-temporal density (and not CDR)

Number of individuals at a given hour at any IRIS cell in Paris

IRIS cells







Overview of our approach

1. Sample *x* *(≈ 30)* visits per user uniformly at random (to decrease sensitivity)
2. Create time-series: map tower cell counts to IRIS cell counts
3. Perturb these time-series to guarantee differential privacy

Overview of our approach

1. Sample *x (≈ 30)* visits per user uniformly at random
2. Create time-series: map tower cell counts to IRIS cell counts
3. Perturb these time-series to guarantee differential privacy

Perturbation of time series

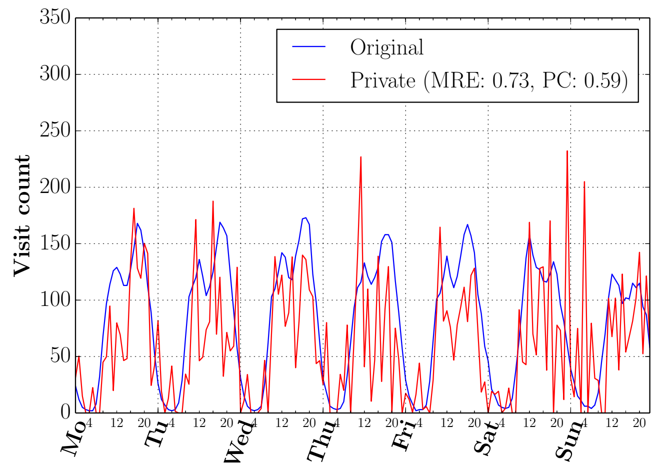
Naïve solution: add properly calibrated Laplace noise to each count of the IRIS cell (one count per hour over 1 week)

Problem: *Counts are much smaller than the noise!*

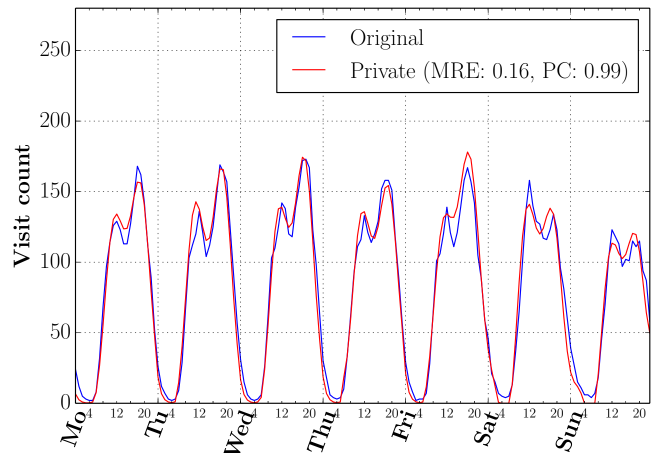
Our approach:

1. cluster nearby less populated cells until their aggregated counts become sufficiently large to resist noise.
2. perturb the aggregated time series by adding noise to their largest Fourier coefficients
3. scale back with the (noisy) total number of visits of individual cells to get the individual time series

Naïve approach (ε=0.3)



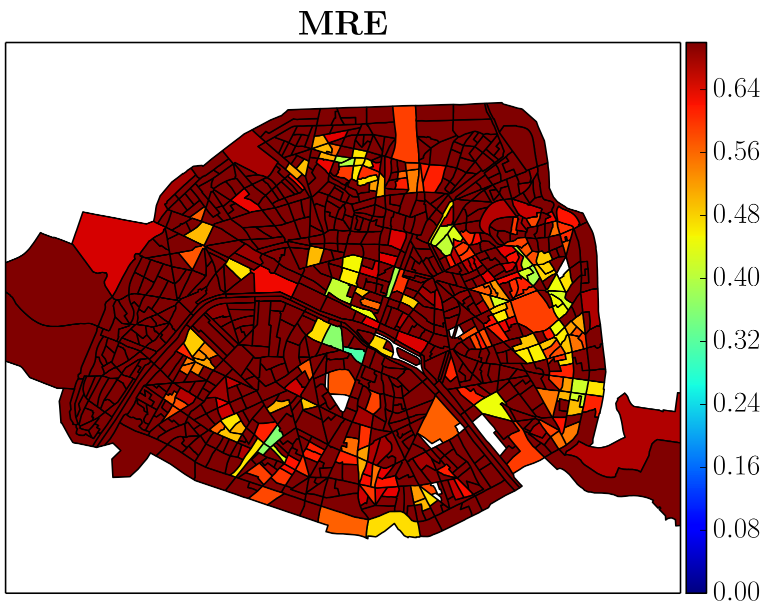
Our approach (ε=0.3)



Performance evaluation 1: Mean Relative Error

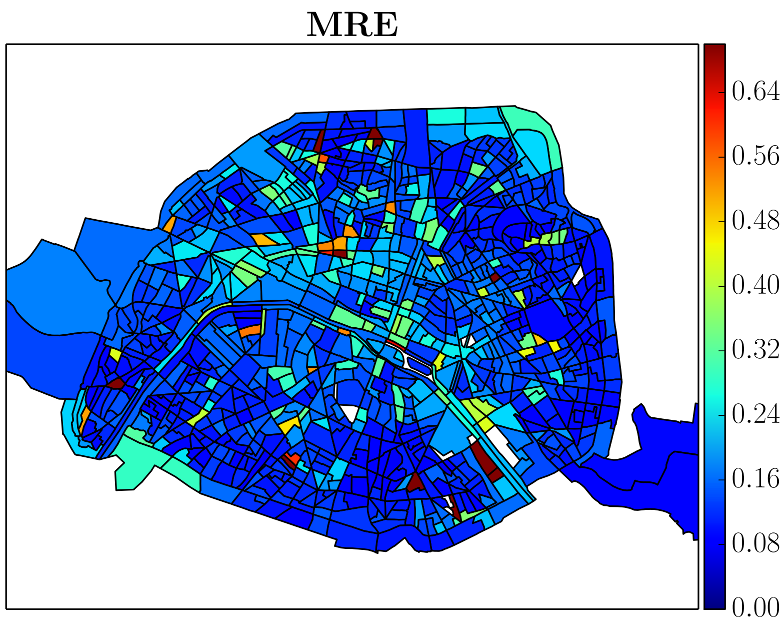
$$\mathrm{MRE}(\mathbf{X}, \hat{\mathbf{X}}) = (1/168) \sum_{i=1}^{168} \frac{|\hat{X}_i - X_i|}{\max(\gamma, X_i)}$$

Naïve approach (ε=0.3)



Average MRE: 1.01

Our scheme (ε=0.3)



Average MRE: 0.17

Conclusion 1: There are no universal solutions!

- **There are no “universal” anonymization solutions that fit all applications**
  - in order to get the best accuracy, they have to be customized to the application and the public characteristics of the dataset
    - specific context
    - specific utility/privacy tradeoff
    - Specific ADV models
    - Specific impacts..
- Anonymization is all about utility/efficiency trade-off!

Conclusion 2:
Anonymization does not solve everything!

- Anonymization schemes protect against re-identification, **not inference**!

- You can learn and infer a lot from data and meta-data
  - You can infer religion from Mobility data!
  - Interest from Google search requests

- It is up to the society to decide what is acceptable or not!
  - By balancing the benefits with the risks*.

**Benefit-Risk Analysis for Big Data Projects, http://www.futureofprivacy.org/wp-content/uploads/FPF_DataBenefitAnalysis_FINAL.pdf*

Conclusion 3: Data Anonymization is Hard!

- Most people don’t understand it or don’t want to understand it
  - “this is trivial”, “I need the data for my business”
- It is technically difficult
  - Requires real expertise
- Big Data Anonymization is even harder
  - Finality/utility is not known
- We need better tools to:
    - Anonymize datasets
    - To perform PRA (Privacy Risk Analysis) of Anonymized Data
    - To evaluate Anonymization solutions
  - Security by Obscurity does not work
    - Anonymization algorithms should be auditable

MERCI!

Claude.castelluccia@inria.fr