Privacy in scientific data

Sébastien Gambs
Université de Rennes 1 - Inria / IRISA

sgambs@irisa.fr

7 November 2014

Introduction

Digital traces

- In the "Information Society", each individual constantly leaves digital traces of his actions that can be linked back to his identity.

- IP address ⇒ location, identifier, content.
- History of requests ⇒ interests.
- Knowledge of social network ⇒ inferences on political opinions, religion, hobbies, . . .

Privacy

- Privacy is one of the fundamental right of individuals:
  - Universal Declaration of the Human Rights at the assembly of the United Nations (Article 12), 1948.
  - European directive 95/46/EC on the protection of personal data (currently being revised towards a regulation).
- Risk: collect and use of digital traces for malicious purposes.
- Examples: targeted spam, identity theft, profiling, (unfair) discrimination.

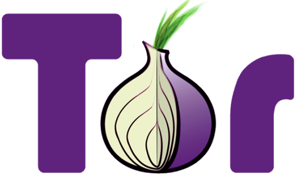
Personally identifiable information

- Personally identifiable information : ensemble of information that can be used to uniquely identified an individual.
- Examples : first and last name, social security number, place and date of birth, physical and email address, phone number, credit card number, biometric data (such as fingerprint and DNA), . . .
- Sensitive because they identify uniquely an individual and can be used to easily cross-referenced databases.
- Main limits of the definition :
  - does not take into account some attributes or patterns in the data that can seem innocuous individually but can identified an individual when combined together (quasi-identifiers).
  - does not take into account the inference potential of the data considered (*e.g.*, queries, social network).

Privacy enhancing technologies

Privacy Enhancing Technologies (*PETs*) : ensemble of techniques and applications for protecting the personal data of an individual while he is online.

Example of PET : anonymous communication network.



Two fundamental principles behind the PETs :

- Data minimization : only the information necessary for completing a particular purpose should be collected/revealed.
- Data sovereignty : enable a user to keep the control on his personal data and how they are collected and disseminated.

Research domains with strong privacy issues

- Genomic/medical data.
- Possible risks : inference on genetic diseases or tendency to develop particular health problems, leakage of information about ethnic origin and genomics of relatives, genetic discrimination, . . .
- Social data.
- Possible risks : reconstruction of the social graph, inferences on political opinions, religion, sexual orientations, hobbies, . . .
- Location data.
- Possible risks : later in this presentation.

Introduction

Inference attacks and privacy models

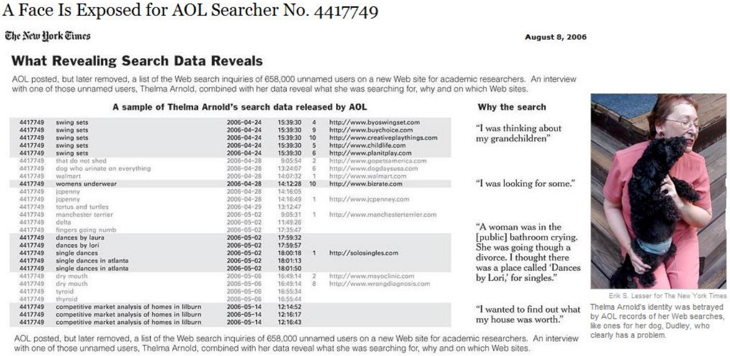
Use case: Location privacy

Conclusion

Inference attacks and privacy models

Pseudonymization is not an alternative to anonymization

Replacing the name of a person by a pseudonym $\nRightarrow$ preservation of the privacy of this individual



A Face Is Exposed for AOL Searcher No. 4417749

The New York Times

August 8, 2006

What Revealing Search Data Reveals

AOL posted, but later removed, a list of the Web search inquiries of 658,000 unnamed users on a new Web site for academic researchers. An interview with one of those unnamed users, Thelma Arnold, combined with her data reveal what she was searching for, why and on which Web sites.

A sample of Thelma Arnold's search data released by AOL

4417749	swing sets	2006-04-24	15:39:30	4	http://www.byoswingset.com
4417749	swing sets	2006-04-24	15:39:30	9	http://www.buychoice.com
4417749	swing sets	2006-04-24	15:39:30	10	http://www.creativeplaythings.com
4417749	swing sets	2006-04-24	15:39:30	5	http://www.childlife.com
4417749	swing sets	2006-04-24	15:39:30	6	http://www.planitplay.com
4417749	that do not shed	2006-04-28	9:05:54	2	http://www.gopetsamerica.com
4417749	dog who urinate on everything	2006-04-28	13:24:07	6	http://www.dogdaysusa.com
4417749	walmart	2006-04-28	14:07:32	1	http://www.walmart.com
4417749	womens underwear	2006-04-28	14:12:28	10	http://www.bizrate.com
4417749	jcpenny	2006-04-28	14:16:05		
4417749	jcpenny	2006-04-28	14:16:49	1	http://www.jcpenney.com
4417749	tortus and turtles	2006-04-29	13:12:47		
4417749	manchester terrier	2006-05-02	9:05:31	1	http://www.manchesterterrier.com
4417749	delta	2006-05-02	11:49:26		
4417749	fingers going numb	2006-05-02	17:35:47		
4417749	dances by laura	2006-05-02	17:59:32		
4417749	dances by lori	2006-05-02	17:59:57		
4417749	single dances	2006-05-02	18:00:18	1	http://solosingles.com
4417749	single dances in atlanta	2006-05-02	18:01:13		
4417749	single dances in atlanta	2006-05-02	18:01:50		
4417749	dry mouth	2006-05-06	16:49:14	2	http://www.mayoclinic.com
4417749	dry mouth	2006-05-06	16:49:14	8	http://www.wrongdiagnosis.com
4417749	tyroid	2006-05-06	16:55:34		
4417749	thyroid	2006-05-06	16:55:44		
4417749	competitive market analysis of homes in lilburn	2006-05-14	12:14:52		
4417749	competitive market analysis of homes in lilburn	2006-05-14	12:16:17		
4417749	competitive market analysis of homes in lilburn	2006-05-14	12:16:43		

Why the search

"I was thinking about my grandchildren"

"I was looking for some."

"A woman was in the [public] bathroom crying. She was going though a divorce. I thought there was a place called 'Dances by Lori,' for singles."

"I wanted to find out what my house was worth."

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.

AOL posted, but later removed, a list of the Web search inquiries of 658,000 unnamed users on a new Web site for academic researchers. An interview with one of those unnamed users, Thelma Arnold, combined with her data reveal what she was searching for, why and on which Web sites.

(Extract from an article from the New York Times, 6 August 2006)

What the directive 95/46/EC says about anonymized data

> *"Whereas the principles of protection must apply to any information concerning an identified or identifiable person; whereas, to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person; whereas the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable; . . . "*

Main challenge : quantifying the risk and difficulty of de-anonymizing data.

What the draft of the data protection regulation says

> *"To ascertain whether means are reasonably likely to be used to identify the individual, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration both available technology at the time of the processing and technological development."*

Consequence : evaluation of risk of de-anonymization should take into account the ressources needed to conduct the re-identification and should be done on a regular basis.

Inference attack

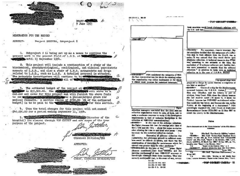
- Inference attack : the adversary takes as input a published dataset (and possibly some background knowledge) and tries to infer some personal information regarding individuals contained in the dataset.
- Main challenge : to be able to give some privacy guarantees even against an adversary having some auxiliary knowledge.
- We may not even be able to model this *a priori* knowledge.
- Remark: maybe my data is private today but it may not be so in the future due to the public release of some other data.

Sanitization

Sanitization : *process increasing the uncertainty in the data in order to preserve privacy.*

⇒ Inherent trade-off between the desired level of privacy and the utility of the sanitized data.

Typical application : public release of data (offline or online context).



[redacted]

DRAFT-[redacted]
9 June 1953

MEMORANDUM FOR THE RECORD

SUBJECT: Project MKULTRA, Subproject 8

1. Subproject 8 is being set up as a means to continue the present work in the general field of L.S.D. at [redacted] until 11 September 1954.

2. This project will include a continuation of a study of the biochemical, neurophysiological, sociological, and clinical psychiatric aspects of L.S.D., and also a study of L.S.D. antagonists and drugs related to L.S.D., such as L.A.E. A detailed proposal is attached. The principle investigators will continue to be [redacted] all of [redacted]

3. The estimated budget of the project at [redacted] is $39,500.00. The [redacted] will serve as a cut-out and cover for this project and will furnish the above funds to the [redacted] as a philanthropic grant for medical research. A service charge of $790.00 (2% of the estimated budget) is to be paid to the [redacted] for this service.

4. Thus the total charges for this project will not exceed $40,290.00 for a period ending September 11, 1954.

5. [redacted] (Director of the hospital) are cleared through TOP SECRET and are aware of the true purpose of the project.

Sidney Gottlieb
for [redacted]
Chemical Division/TSS

APPROVED:

Sidney Gottlieb
Chief, Chemical Division/TSS

[redacted]

[redacted] PROGRAM [redacted]

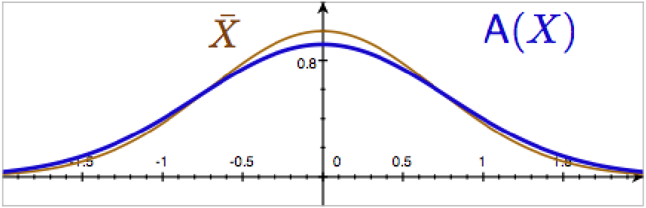
[illegible]

Examples drawn from the “sanitization” entry on Wikipedia

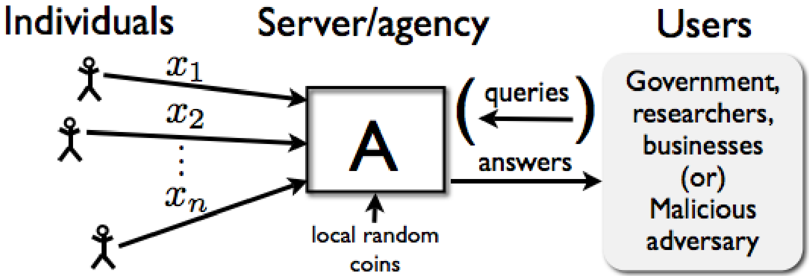
Remark : utility can be defined in terms of global properties of the data or depend on the application considered.

Randomization methods

Randomization : add independent noise (such as Gaussian or uniform) to the values transmitted.
Goal : hide the specific values of attributes while preserving the joint distribution of the data.



Possible model for the randomization methods



Extract from a tutorial of Adam Smith on the protection of privacy in databases (March 2008)

De-anonymization attack

- De-anonymization attack : the adversary takes as input a sanitized dataset and some background knowledge and tries to infer the identities of the individuals contained in the dataset.
- The *re-identification risk* measures the success probability of this attack.
- Other dimensions :
  - The attack can be *passive* (if the adversary simply observes the result of the anonymization) or *active* (if he can influence the system or the anonymization process).
  - Robustness of the attack against perturbation of the data.
  - Possibility of repeated de-anonymization or only one-shot.

Sweeney's original linking attack

Linking attack : the adversary tries to link together the records of two different datasets which contains a common fraction of individuals.

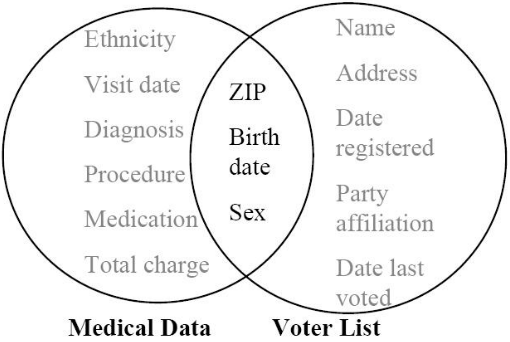
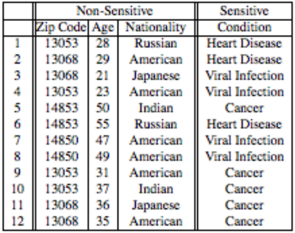


Figure 1 Linking to re-identify data

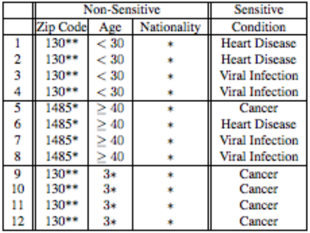
k-anonymity (Sweeney 02)

- 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

- Main challenge : extracting useful knowledge while preserving the confidentiality of individual sensitive data.

Extensions of k-anonymity

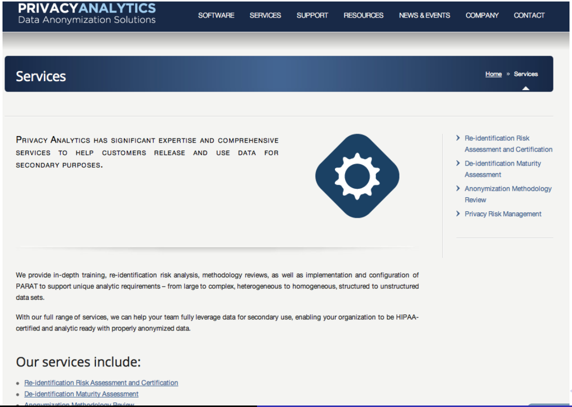
- l-diversity (MKGV[1] 07) : maintain the diversity for each group with respect to the possible values of the sensible attributes.
- Can be instancied by a metric based on *entropy*.
- Prevent against attacks based on homogeneity and some other attacks.
- t-closeness (LLV[2] 07) : the distribution of the attributes in each group must be close to that on the global population.
- t is a threshold that should not be exceed and which represents the proximity between distributions.
- Main objective of extensions : prevent the possibility of inferring the value of a sensitive attribute (but not to protect against re-identification).

[1] Machanavajjhala, Gehrke, Kifer and Venkitasubramaniam.

[2] Li, Li et Venkatasubramanian.

Privacy analytics

- Privacy analytics : Canadian company founded in 2007 specialized in anonymization solutions and re-identification assessment for medical data.



Services

PRIVACY ANALYTICS HAS SIGNIFICANT EXPERTISE AND COMPREHENSIVE SERVICES TO HELP CUSTOMERS RELEASE AND USE DATA FOR SECONDARY PURPOSES.

- Re-identification Risk Assessment and Certification
- De-identification Maturity Assessment
- Anonymization Methodology Review
- Privacy Risk Management

We provide in-depth training, re-identification risk analysis, methodology reviews, as well as implementation and configuration of PARAT to support unique analytic requirements – from large to complex, heterogeneous to homogeneous, structured to unstructured data sets.

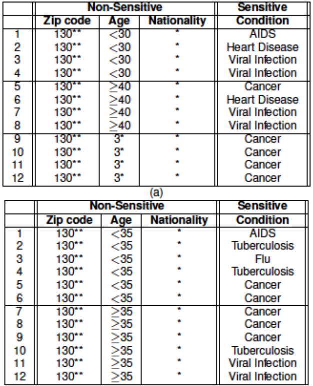
With our full range of services, we can help your team fully leverage data for secondary use, enabling your organization to be HIPAA-certified and analytic ready with properly anonymized data.

Our services include:

- Re-identification Risk Assessment and Certification
- De-identification Maturity Assessment
- Anonymization Methodology Review

Composition attack

- 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

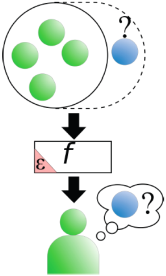
(a)

	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)

Differential privacy: principle (Dwork 06)

- Recent privacy notion developed within the community of private data analysis.



- Basically ensures that whether or not an item is in the profile of an individual does not influence too much the output.
- Give strong privacy guarantees that hold independently of the auxiliary knowledge of the adversary.

Differential privacy: definition

- Differential privacy (Dwork 06): A randomized function K gives $\epsilon-$*differential privacy* if for all possible inputs X_1 and X_2 differing in a most one element, and all $S \subseteq Range(K)$,

  $$\Pr[K(X_1) \in S)] \leq \exp(\epsilon) \times \Pr[K(X_2) \in S)] \qquad (1)$$

  The probability is taken over all the coin tosses of K.
- ϵ is a public privacy parameter.
- Typical value: 0.01, 0.1 or even 3.
- Properties:
  - Composition: the application of k ϵ-differentially private mechanisms leads to a $k\epsilon$-differentially private mechanism.
  - Postprocessing does not hurt privacy.

Sensitivity

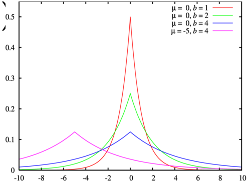
- The sensitivity measures how much the output of a function can change with respect to a small change in the input.
- Global sensitivity (Dwork 06): For $f : D^n \to R$, the (global)*sensitivity* of f is

  $$\mathrm{GS}(f) = max_{X_1,X_2} \|f(X_1) - f(X_2)\|_1 \tag{2}$$

  for all X_1,X_2 differing in at most one element.
- Example: two profiles S_1 and S_2 are *neighbours* if they are the same up to a particular item.
- The sensitivity of the Hamming distance (computed between two binary vectors) is one.

Laplacian mechanism

- Achieves ϵ-differential privacy by adding noise directly proportional to GS(f)



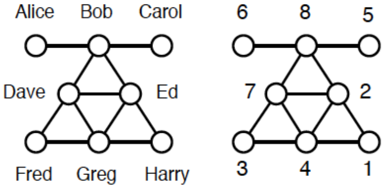
- Theorem (Dwork 06): For $f : D^n \rightarrow R$, a randomized function K achieves ϵ-differential privacy if it releases on input x

$$K(x) = f(x) + \text{Lap}(\frac{\text{GS}(f)}{\epsilon}) \qquad (3)$$

for GS(f) the sensitivity of the function f and Lap is a randomly generated noise according to the Laplacian distribution parametrized by $\frac{\text{GS}(f)}{\epsilon}$.

The difficulty of anonymizing structured data

- Anonymizing a (social) graph can be a very difficult task because some patterns in the graph may be unique.
- Example: you are the only one that has 47 friends and which has 3 friends each having 52 friends.
- More structured example:



Alice	6
Bob	8
Carol	5
Dave	7
Ed	2
Fred	3
Greg	4
Harry	1

- Consequence: anonymizing the graph by removing the labels on the nodes and edges is not sufficient.

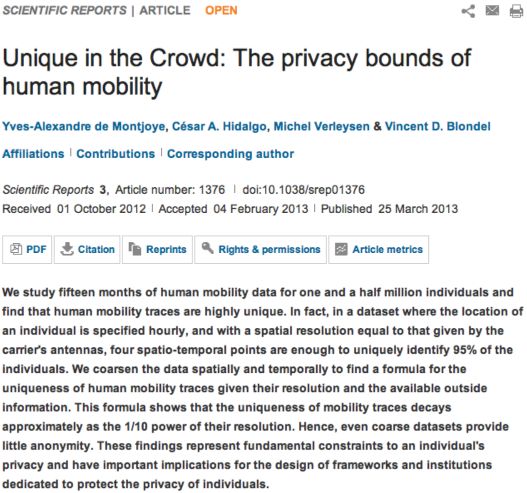
Use case: Location privacy

Location-based services (LBSs)

- Personalize the service provided to the user according to his current position.
- Example :

- Main types of LBS :
  1. LBS depending only from the individual position of the user.
  2. Collaborative LBS whose global output is a function of the locations of many users.
- Non-interactive scenario : sanitization of location data.

Unique in the crowd



SCIENTIFIC REPORTS | ARTICLE OPEN

Unique in the Crowd: The privacy bounds of human mobility

Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen & Vincent D. Blondel

Scientific Reports **3**, Article number: 1376 | doi:10.1038/srep01376
Received 01 October 2012 | Accepted 04 February 2013 | Published 25 March 2013

We study fifteen months of human mobility data for one and a half million individuals and find that human mobility traces are highly unique. In fact, in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals. We coarsen the data spatially and temporally to find a formula for the uniqueness of human mobility traces given their resolution and the available outside information. This formula shows that the uniqueness of mobility traces decays approximately as the 1/10 power of their resolution. Hence, even coarse datasets provide little anonymity. These findings represent fundamental constraints to an individual's privacy and have important implications for the design of frameworks and institutions dedicated to protect the privacy of individuals.

Inference attack on location data

- Location privacy seeks to *prevent an unauthorized entity from learning the past, current and future location of an individual* (Beresford et Stajano 03).
- Inference attack on location data : the adversary takes as input a geolocated dataset (and possibly some background knowledge) and tries to infer some personal information regarding individuals contained in the dataset.
- Main objective of this work : quantify the privacy risks of disclosing location data.
- Joint work with Marc-Olivier Killijian (LAAS-CNRS) and Miguel Núñez del Prado (now Intersec, previously LAAS-CNRS).

Possibles objectives of an inference attack

1. Identification of important places, called *Point of Interests* (POI), characterizing the interests of an individual.

- Example: home, place of work, gymnasium, political headquarters, medical center, . . .

2. Prediction of the movement patterns of an individual, such as his past, present and future locations.
3. Linking the records of the same individual contained in the same dataset or in different datasets (either anonymized or under different pseudonyms).

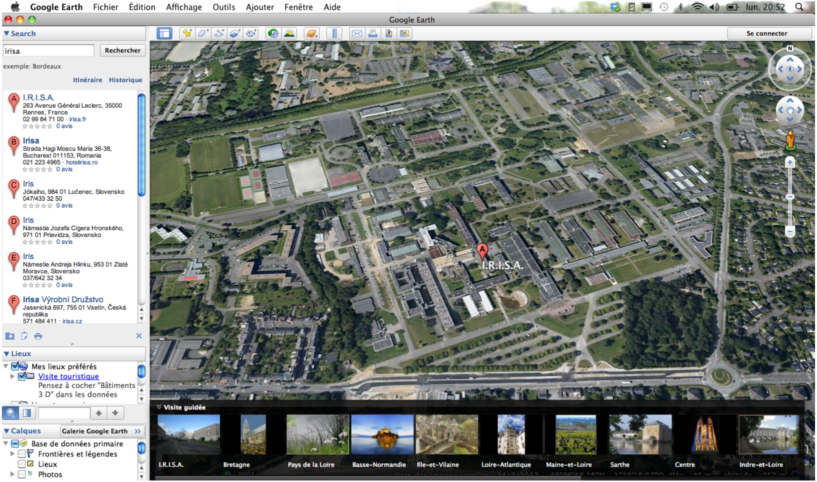
Auxiliary knowledge

The adversary can have *auxiliary* knowledge that may help him in conducting a privacy breach.

Examples of auxiliary knowledge :

- presence of an individual within an anonymized dataset,
- partial knowledge of its attributes (such as home address or place of work),
- a model of his habits,
- knowledge of his social network,
- knowledge of the distribution of the attributes within the population,
- geographical knowledge of roads and relief,
- . . .

Example of background knowledge : Google Earth



irisa

exemple: Bordeaux

A I.R.I.S.A.
263 Avenue Général Leclerc, 35000 Rennes, France
02 99 84 71 00 · irisa.fr
☆☆☆☆☆ 0 avis

B Irisa
Strada Hagi Moscu Maria 36-38, Bucharest 011153, Romania
021 223 4965 · hotelirisa.ro
☆☆☆☆☆ 0 avis

C Iris
Jókaiho, 984 01 Lučenec, Slovensko
047/433 32 50
☆☆☆☆☆ 0 avis

D Iris
Námestie Jozefa Cígera Hronského, 971 01 Prievidza, Slovensko
☆☆☆☆☆ 0 avis

E Iris
Námestie Andreja Hlinku, 953 01 Zlaté Moravce, Slovensko
037/642 32 34
☆☆☆☆☆ 0 avis

F Irisa Výrobní Družstvo
Jasenická 697, 755 01 Vsetín, Česká republika
571 484 411 · irisa.cz

☑ Mes lieux préférés
☑ Visite touristique
Pensez à cocher "Bâtiments 3 D" dans les données

Base de données primaire
□ Frontières et légendes
□ Lieux
□ Photos

Identification of home and place of work (SPRINGL'10)

Heuristic to identify the home :

- Choose the last stop before midnight.

Heuristic to identify the place of work :

- Choose the most "stable" location during the day.

Reverse geocoding : maps the coordinates of a location to a physical address.

⇒

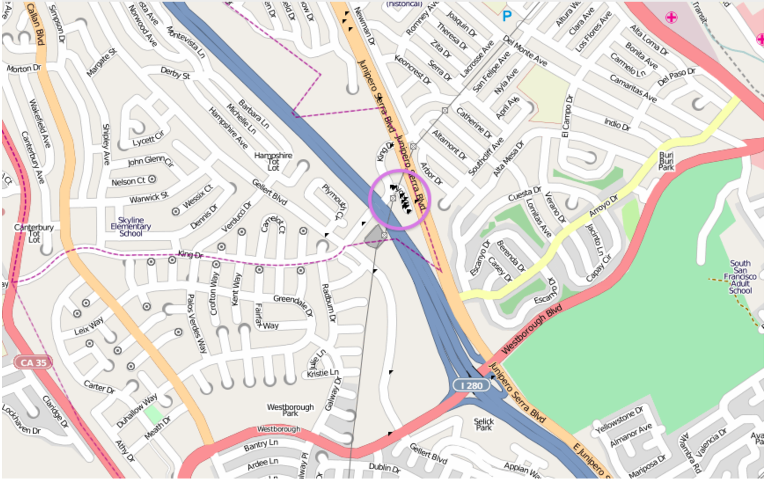
Yellow Pages : associate a physical address with a list of possibles candidates.

Identification of POIs through clustering algorithm

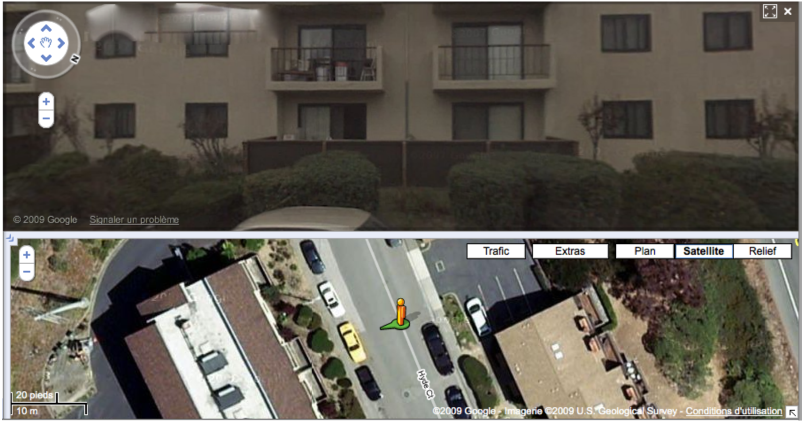
- Clustering : form of unsupervised learning that aims at grouping together objects that are similar (*intra-similarity*) while putting in separate clusters objects that are different (*inter-dissimilarity*).
- Inference attack :
  1. Delete all mobility traces in which the person is in movement.
  2. Run a clustering algorithm on the remaining traces in order to discover significant clusters.
  3. Return as POI the median of each cluster.

  Validation issue : how to evaluate the quality of the POIs returned if we do not have access to a “ground truth”?

Identification of the house of a taxi (AINA'10)



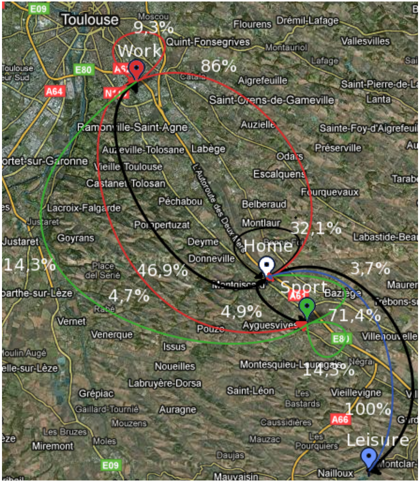
Identification of the house of a taxi (view from GoogleMaps and StreetView)



Mobility Markov chain (TDP'11)

- Objective : to represent in a compact way the mobility behaviour of an individual.
- The states of the chain are POIs and a transitions represents the probability from moving from one POI to another.
- Construction :
  - Remove all moving traces.
  - From the resulting traces, extract the POIs by running a clustering algorithm.
  - Label each trace with the corresponding POI and compute the transitions probabilities.
- Temporal variant of the model (DYNAM'11): decompose the time into slices, the label of a stage corresponds to POI/time slice.

Example of mobility Markov chain

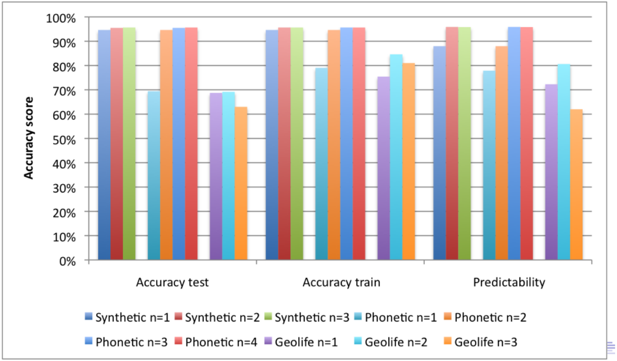


Predicting the next location (MPM'12)

- Prediction technique : from the actual location, choose te transition leaving from this POI that has the highest probability and predicts the corresponding POI.
- Evaluation method : splitting of the mobility traces between a training set and a testing set (50%-50%).
- The mobility Markov chain is learnt from the training set and his prediction rate is evaluated in the testing set.
- Variant of the method : to remember the *n* last visited states (instead of simply the current one).
- Example : a user has visited "work" and then "supermarket", which POI is the one visited next by the user?

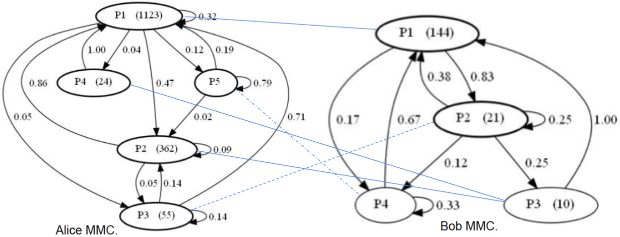
Experimental results

- The prediction method was tested on 3 mobility datasets (synthetic, Phonetic, Geolife) with n varying between 1 and 3 (best prediction rate obtained for $n = 2$).
- Results : success rate of the prediction between 70 and 95%.

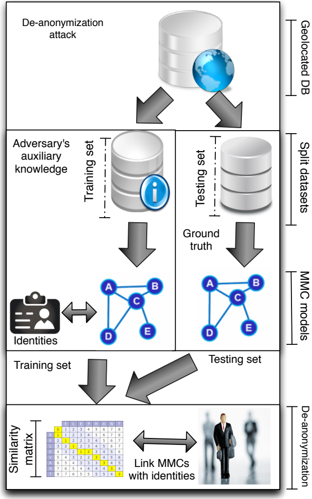


De-anonymization attack via MMC (Trustcom'13)

- Objective : find an individual in an anonymous geolocated dataset.
- Assumption : the adversary has been able to observe in the past the mobility of the some individuals present in the dataset.
- Main idea : to compute a distance metric between 2 MMCs quantifying the difference between two mobility behaviours.



Overview of the de-anonymization attack



De-anonymization attack

Geolocated DB

Adversary's auxiliary knowledge

Training set

Testing set

Split datasets

Ground truth

Identities

MMC models

Training set

Testing set

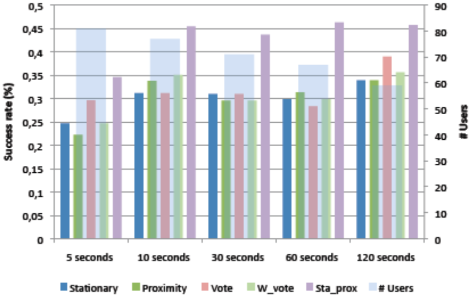
Similarity matrix

Link MMCs with identities

De-anonymization

De-anonymization attack via MMC

- Design of different distance metrics (geometrical, topological, logical) between MMCs and different way to combine the predictors.



- Best de-anonymization rate : 45% (obtained by combining 2 predictors).

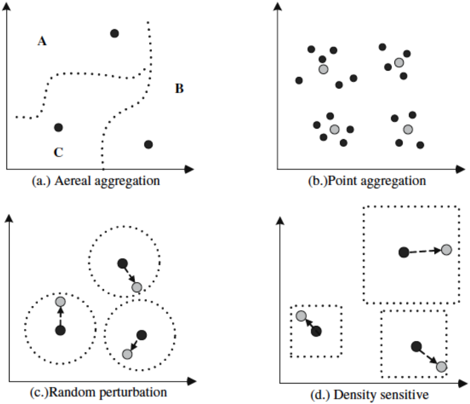
Geographical masks

Main idea : *modify the geographical location to preserve the privacy of an individual.*

Possible modifications :

- Aggregate the location of several individuals into one spatial area or a single location.
  Example : choose the average or median location within a group of locations.
- Randomly perturb the location.
  Example : choose a direction at random and apply some noise (for instance uniform or Gaussian).
- Randomize the location by taking into account the density within the neighborhood.
  dense area ⇒ weak perturbation
  area with low density ⇒ large perturbation

Examples of geographical masks



(a.) Aereal aggregation

(b.)Point aggregation

(c.)Random perturbation

(d.) Density sensitive

(Taken from Amstrong, Rushton and Zimmerman 99)

Other possible transformations

- Sample the data with a lower frequency.
  Example : store only the location every 15 minutes instead of every 15 seconds.
- Remove the recordings that are deemed too sensible.
- Add dummy records.

Remark : even if the location is perturbed, the range of the query can be such that the mobile device can locally compute the true answer to the query.

k-anonymity and spatial cloaking

- Main idea : protect the privacy of a user by *blending him into the crowd*.
- k-anonymization : sanitization process of a dataset (by suppression and generalization) in which each record is indistinguishable from at least $k - 1$ other records.
- Spatial cloaking (Gruteser and Grunwald 03) : extension of the concept of k-anonymity to spatio-temporal data.
- Main idea : ensure that at each time step, each user is located within an area that is shared by at least $k - 1$ other users.
- Possible method : recursively split the space in areas of different sizes such that each area respects the property of k-anonymity.

Illustration of spatial cloaking

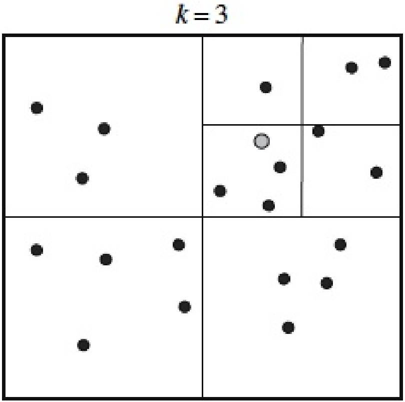


Illustration of spatial cloacking for $k = 3$
(extrait de Gruteser et Grunwald 03).

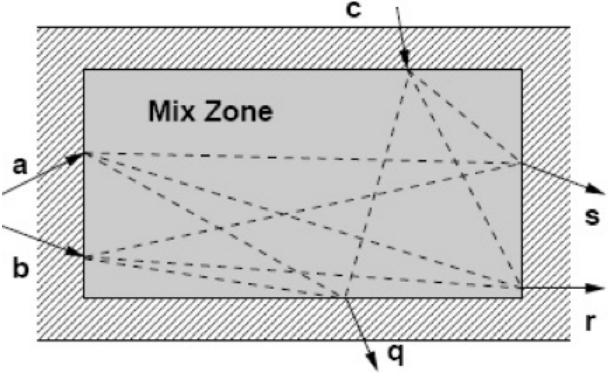
Limits of geographical masks and spatial cloaking

- If the adversary has some geographical knowledge about the sanitized area, he can discard some unrealistic hypotheses.
- Example : if after a random perturbation, the returned location is situated within an area difficult to reach such as river or a mountain ⇒ the adversary can reject this hypothesis by considering instead the closest accessible area.
- Linkability risk : even if it is impossible to identify exactly an individual, it is sometimes possible to link the actions of a group of individuals.
- Example of inference : at each time step, the adversary can follow the movement of one group from one area to another.

Addressing the linkability risk

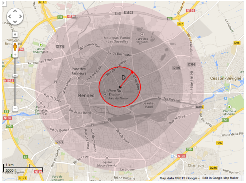
- Swap some traces within two different pseudonyms.
- Example : user A exchanges his pseudonym with user B to make his behavior more atypical and less predictable.
- Mix-zone (Beresford and Stajano 03) : area of space in which
  - no observations are recorded and
  - such that a user leaves the area with a different pseudonym that this person had when entering.
- Inspired from the *Mix-nets* of Chaum used for the anonymous communication of messages.
- Example : the university can be a mix-zone in which no measurement are performed while we are here.
  When we leave it, we will receive a different pseudonym from the one we had before entering.

Illustration of mix-zone



(D,ϵ)-location privacy (work in progress with Ehab Elsalamouny)

- Main idea: the adversary should not be able to distinguish the real location of the user to adjacent locations within a distance D (ϵ represents the desired level of indistinguishability).

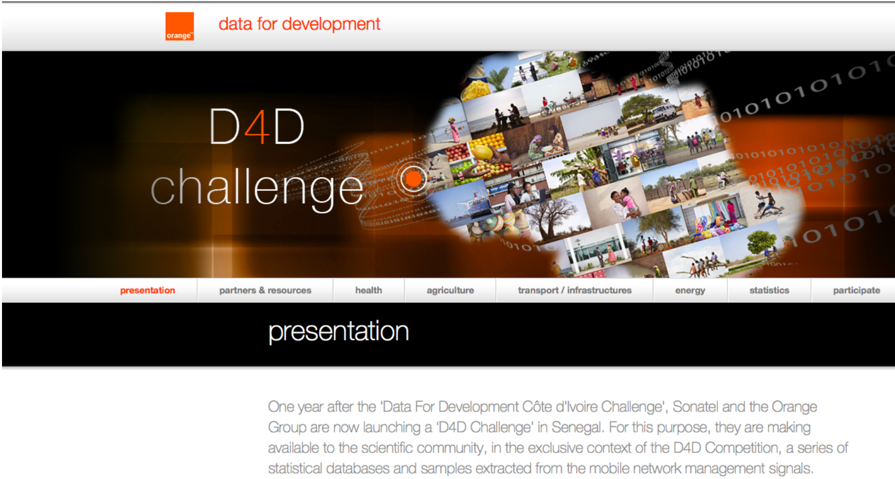


- Adaption of differential privacy (Dwork 06) to the context of location-based services.
- Related notion: geo-indistinguishability (Andrés, Bordenabe, Chatzikokolakis and Palamidessi 13)

Example of privacy challenge : opening of mobility data

- Objective : publishing of the mobility traces of users issued from a public transportation system (*e.g.*, subway, bus or bike) or a phone operator.
- Fundamental question : how to anonymize the data before publishing them to limit the privacy risks?
- By perturbing the traces? By aggregating them? By decreasing the granularity of the information revealed?
- Determining the "good manner" to anonymize a dataset is often a long and difficult process.

Example of Call Details Records



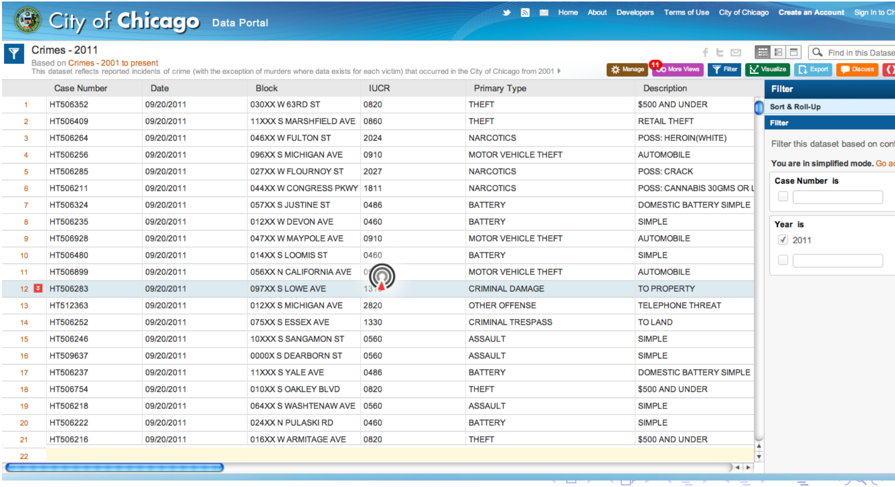
data for development

presentation | partners & resources | health | agriculture | transport / infrastructures | energy | statistics | participate

presentation

One year after the 'Data For Development Côte d'Ivoire Challenge', Sonatel and the Orange Group are now launching a 'D4D Challenge' in Senegal. For this purpose, they are making available to the scientific community, in the exclusive context of the D4D Competition, a series of statistical databases and samples extracted from the mobile network management signals.

A final practical case : crime records of Chicago



Crimes - 2011

Based on Crimes - 2001 to present

This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001

	Case Number	Date	Block	IUCR	Primary Type	Description
1	HT506352	09/20/2011	030XX W 63RD ST	0820	THEFT	$500 AND UNDER
2	HT506409	09/20/2011	11XXX S MARSHFIELD AVE	0860	THEFT	RETAIL THEFT
3	HT506264	09/20/2011	046XX W FULTON ST	2024	NARCOTICS	POSS: HEROIN(WHITE)
4	HT506256	09/20/2011	096XX S MICHIGAN AVE	0910	MOTOR VEHICLE THEFT	AUTOMOBILE
5	HT506285	09/20/2011	027XX W FLOURNOY ST	2027	NARCOTICS	POSS: CRACK
6	HT506211	09/20/2011	044XX W CONGRESS PKWY	1811	NARCOTICS	POSS: CANNABIS 30GMS OR L
7	HT506324	09/20/2011	057XX S JUSTINE ST	0486	BATTERY	DOMESTIC BATTERY SIMPLE
8	HT506235	09/20/2011	012XX W DEVON AVE	0460	BATTERY	SIMPLE
9	HT506928	09/20/2011	047XX W MAYPOLE AVE	0910	MOTOR VEHICLE THEFT	AUTOMOBILE
10	HT506480	09/20/2011	014XX S LOOMIS ST	0460	BATTERY	SIMPLE
11	HT506899	09/20/2011	056XX N CALIFORNIA AVE	09	MOTOR VEHICLE THEFT	AUTOMOBILE
12	HT506283	09/20/2011	097XX S LOWE AVE	131	CRIMINAL DAMAGE	TO PROPERTY
13	HT512363	09/20/2011	012XX S MICHIGAN AVE	2820	OTHER OFFENSE	TELEPHONE THREAT
14	HT506252	09/20/2011	075XX S ESSEX AVE	1330	CRIMINAL TRESPASS	TO LAND
15	HT506246	09/20/2011	10XXX S SANGAMON ST	0560	ASSAULT	SIMPLE
16	HT509637	09/20/2011	0000X S DEARBORN ST	0560	ASSAULT	SIMPLE
17	HT506237	09/20/2011	11XXX S YALE AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE
18	HT506754	09/20/2011	010XX S OAKLEY BLVD	0820	THEFT	$500 AND UNDER
19	HT506218	09/20/2011	064XX S WASHTENAW AVE	0560	ASSAULT	SIMPLE
20	HT506222	09/20/2011	024XX N PULASKI RD	0460	BATTERY	SIMPLE
21	HT506216	09/20/2011	016XX W ARMITAGE AVE	0820	THEFT	$500 AND UNDER
22						

Filter this dataset based on con

You are in simplified mode. Go a

Case Number is

☐

Year is

☑ 2011

☐

Crime register of Chicago

- Database accessible via the open data gateway of Chicago.
- Each entry contains information such as the number of the case, the date, the type of crime, a more detailed description but also . . .
- the location (accurate up to the level of the street but the two last digits among the possible 5 are removed).
- Fundamental question : is this anonymization method sufficient to prevent a crime to be associated with an individual or a small group of individuals?
- Example of risk : two neighboring addresses sometimes differ by several decades of numbers.

Defining and quantifying location privacy

- Back to the fundamental question : what does it mean to have a "good" location privacy?
- To be hidden inside a crowd gathered in a small area?

- To be alone in a desert?
- To have a behavior indistinguishable from those of a non-negligible number of other individuals?
- To be unlinkable between different positions?
- Proposed answer : to prevent the inference of sensitive information from the location data revealed (rather than focusing on protecting the location itself).

Conclusion

Privacy in the era of Big Data

- Observation 1 : the capacity to record and store personal data as increased rapidly these last years.
- Examples : activity trackers, smart meters, . . .
- Observation 2 : "Big Data" will result in more and more being available ⇒ increase of inference possibilities.
- Observation 3 : the "Open data" movement will lead to the release of a huge amount of dataset ⇒ worsen the privacy impact of Big Data (observation 2).
- Each new technology that collect and use personal data has to be investigated with respect to privacy.
- Other privacy challenge : implementing the right to be forgotten.

Risks identified by the working party of the article 29 against anonymization techniques

- Singling out: corresponds to the possibility to isolate some or all records which identify an individual in the dataset.
- Corresponds to a direct de-anonymization.
- Linkability: ability to link, at least, two records concerning the same data subject or a group of data subjects (either in the same database or in two different databases).
- Corresponds to a linking attack.
- Inference: possibility to deduce, with significant probability, the value of an attribute from a set of other attributes.
- Not directly related to re-identification risk unless . . . the attribute is directly identifying or it is possible to predict several attributes whose combination acts as a quasi-identifiers.

Conclusion

- Strong need to develop inference attacks that can assess the privacy level provided by a particular sanitization mechanism for a particular research domain.
- By default, all attributes should be consider as possible quasi-identifiers.
- Correlations between attributes can be exploited to increase the efficiency of a de-anonymization attack.
- Anonymization of structured data is even harder (needs to understand how the structure of the data can be used for de-anonymization and how to perturb it to avoid it).
- The possibility of repeatedly de-anonymizing users is much more damaging to privacy that showing the uniqueness of the characteristics of an individual at a particular occasion.

This is the end

Thanks for your attention
Questions?