Foundations of Privacy

Class 1

The teachers of the course

Kostas Chatzikokolakis
CNRS & Ecole Polytechnique

Catuscia Palamidessi
INRIA & Ecole Polytechnique

Logistic Information

- The course will be in English
- We will put the slides on line before every class
- There will be a written exam at the end of the course (on November 28)
- We will give exercises during the course, leave you some time to solve them, and then show the solution. You should try to solve them, as they will help to prepare for the exam
- Please feel free to ask questions any time. We are very happy when people ask questions, as they help to make the class more interactive and lively

Plan of the lectures

- Motivations, a bit of history, main problems, research directions (3 hours)
- Quantitative Information Flow (9 hours)
- Differential Privacy and Extensions (9 hours)
- Location Privacy (3 hours)

Motivations

In the "Information Society", each individual constantly leaves digital traces of his actions that may allow to infer a lot of information about himself

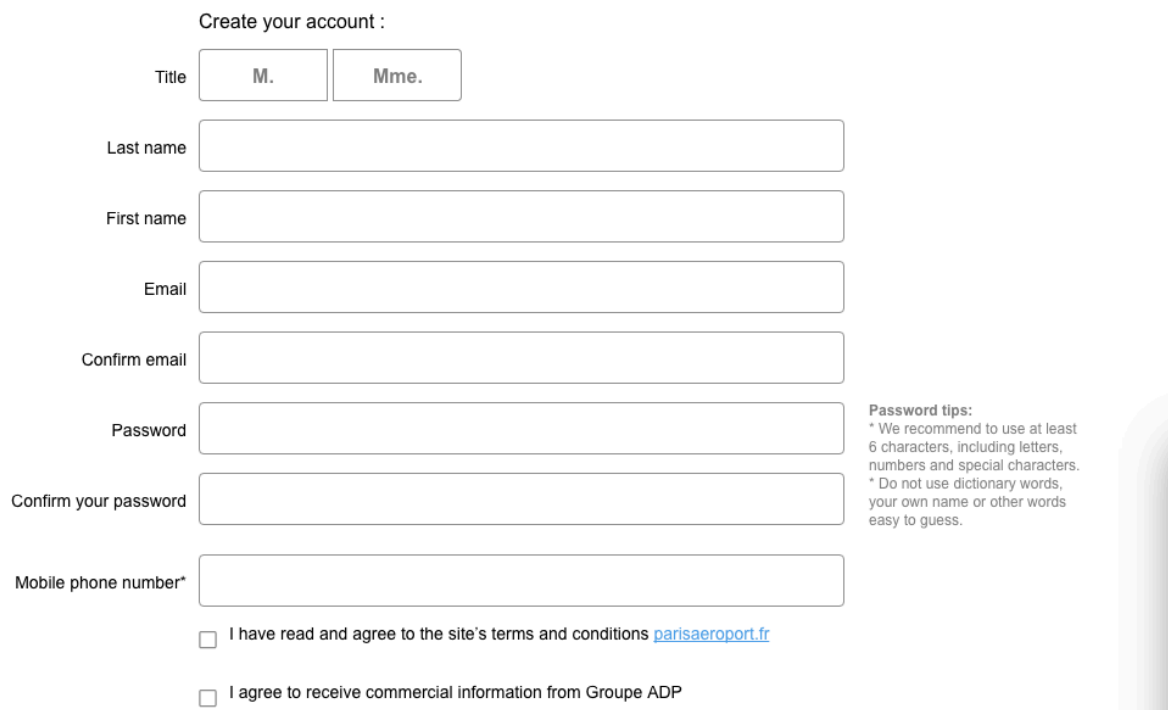
Request to a LBS ⇒ location.

History of requests ⇒ interests.

Activity in social networks ⇒ political opinions, religion, hobbies, . . .

Power consumption (smart meters) ⇒ activities at home.

Example:
Personal information in exchange of a service



We don't know how our information will be used

Concerns about privacy

Risk: collect and use of digital traces for fraudulent purposes.

Examples: targeted spam, identity theft, profiling, discrimination, …

The news are full of problems caused by privacy breaches

The need for privacy is intrinsic to the human nature, although it varies a lot from individual to individual, between cultures, and it evolves with time

Privacy is recognized as 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 stricter regulation).
- Japanese Act on the Protection of Personal Information from 2003 (current discussions to amend it and make stricter).

The new European regulation (will be enforced starting from 2018)

What will be the key changes?

- A **'right to be forgotten'** will help you manage data protection risks online. When you no longer want your data to be processed and there are no legitimate grounds for retaining it, the data will be deleted. The rules are about empowering individuals, not about erasing past events, re-writing history or restricting the freedom of the press.
- **Easier access to your own personal data.**
- **A right to transfer personal data** from one service provider to another.
- When your **consent is required, you must be asked to give it by means of a clear affirmative action.**
- More transparency about how your data is handled, with **easy-to-understand information**, especially for **children**.
- Businesses and organisations will need to **inform you about data breaches** that could adversely affect you **without undue delay**. They will also have to notify the relevant data protection supervisory authority.
- Better enforcement of data protection rights through improved **administrative and judicial remedies** in cases of violations
- Increased **responsibility and accountability** for those processing personal data – through **data protection risk assessments, data protection officers**, and the principles of **'data protection by design'** and **'data protection by default'**.

Different types of sensitive data

- Sensitive information about an individual :
  - credit card / bank information, home access code, passwords, …
    - sensitive because it can bring to attacks to the person or his properties
  - ethnicity, religious beliefs, political opinions, medical status, intimate videos ..
    - Sensitive because it can lead to discrimination.
- Identification information : information that can uniquely identify an individual.
  - First and last name, social security number, physical and email address, phone number, biometric data (such as fingerprint and DNA), …
    - Sensitive because it can be used to cross-reference databases, or to identify him as the subject of certain actions
- Sensitive information for organizations
  - Industries: production plans, research, strategies,…
  - Governments. Police. Armies…
- In this course, we will try to encompass the various scenario. We will abstract from the nature of the sensitive information whenever possible, and present the common principles of information protection, but we will also show that the kind of information (and of adversary) induces differences in the approach.

Why it is difficult to protect privacy

- Traditionally, privacy is protected via:
  - Anonymization
  - Encryption
  - Access control

- However, these methods often fail:
  - encryption and access control cannot protect against the inference of private information from public information
  - anonymization has been proved highly ineffective

Privacy via anonymity

Nowadays, many institutions and companies that collect data use anonymization, i.e., they remove all personal identifiers: name, address, SSN, …

"We don't have any raw data on the identifiable individual. Everything is anonymous"
(CEO of NebuAd, a U.S. company that offers targeted advertising based on browsing histories)

Similar practices are used by Facebook, MySpace, Twitter, …

Privacy via anonymity

However, anonymity-based sanitization has been shown to be **highly ineffective:** Several de-anonymization attacks have been carried out in the last decade

- The quasi-identifiers allow to retrieve the identity in a large number of cases.
- More sophisticated methods (k-anonymity, ℓ-diversity, …) take care of the quasi-identifiers, but they are still prone to composition attacks

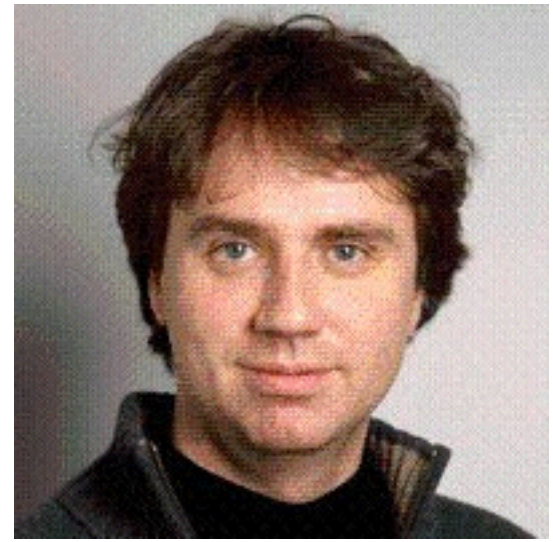
Differential Privacy at Google

RAPPOR

ABSTRACT

Randomized Aggregatable Privacy-Preserving Ordinal Response, or RAPPOR, is a technology for crowdsourcing statistics from end-user client software, anonymously, with strong privacy guarantees. In short, RAPPORs allow the forest of client data to be studied, without permitting the possibility of looking at individual trees. By applying randomized response in a novel manner, RAPPOR provides the mechanisms for such collection as well as for efficient, high-utility analysis of the collected data. In particular, RAPPOR permits statistics to be collected on the population of client-side strings with strong privacy guarantees for each client, and without linkability of their reports.

This paper describes and motivates RAPPOR, details its differential-privacy and utility guarantees, discusses its practical deployment and properties in the face of different attack models, and, finally, gives results of its application to both synthetic and real-world data.



Úlfar Erlingsson

Head of the team
on data security
and privacy at Google

Differential Privacy at Apple

Differential privacy is the statistical science of trying to learn as much as possible about a group while learning as little as possible about any individual in it.

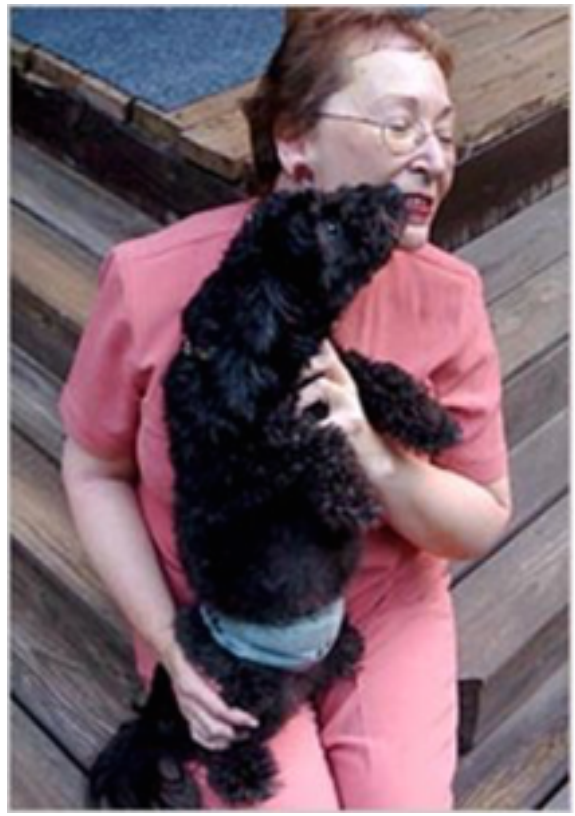
Apple has been doing some important work in this area to enable differential privacy to be deployed at scale."

Craig Federighi,
Vice president of
Software Engineering @Apple

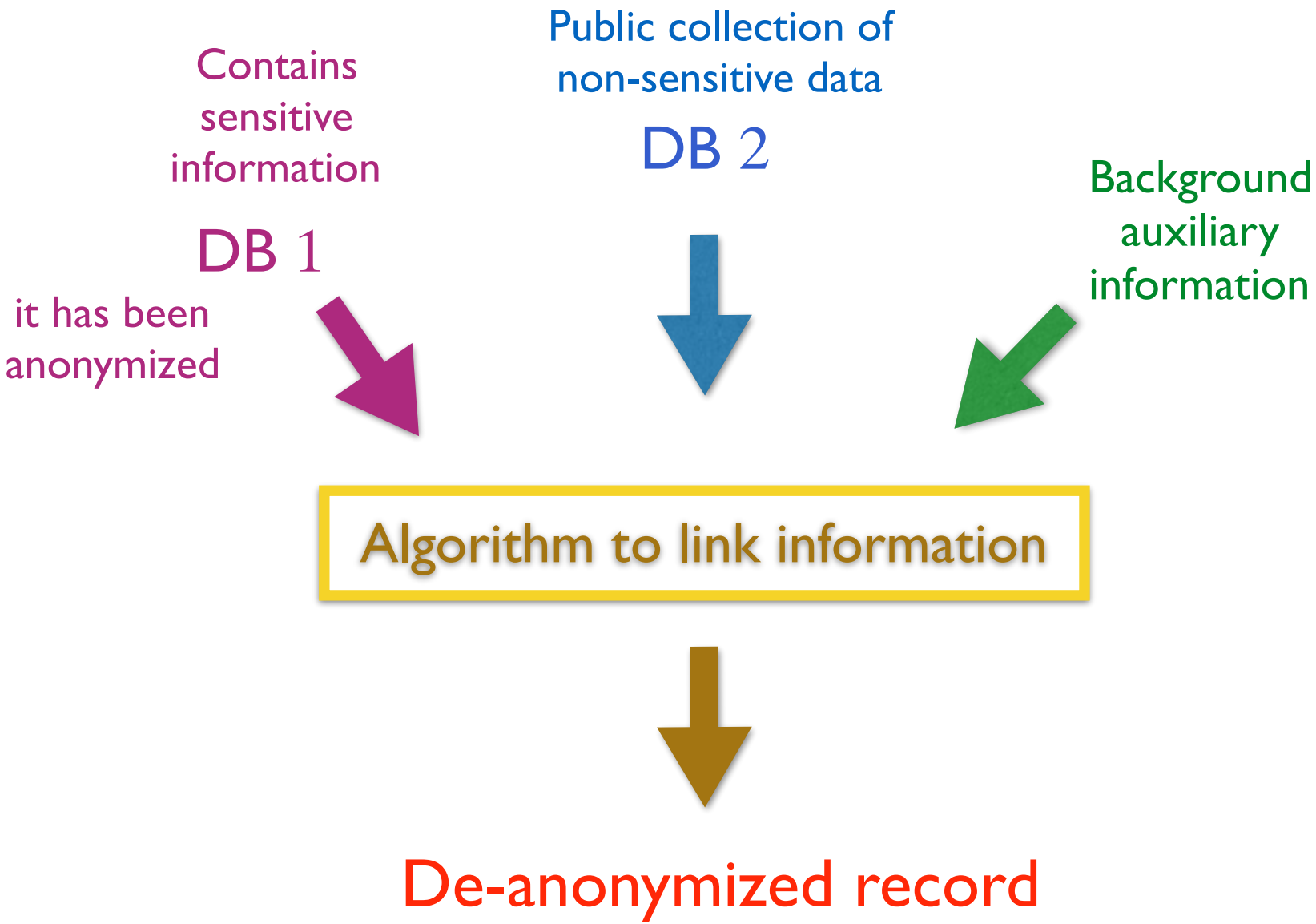
Keynote speech
Annual conference 2016
Apple software developers

Deanonymization attacks (I)

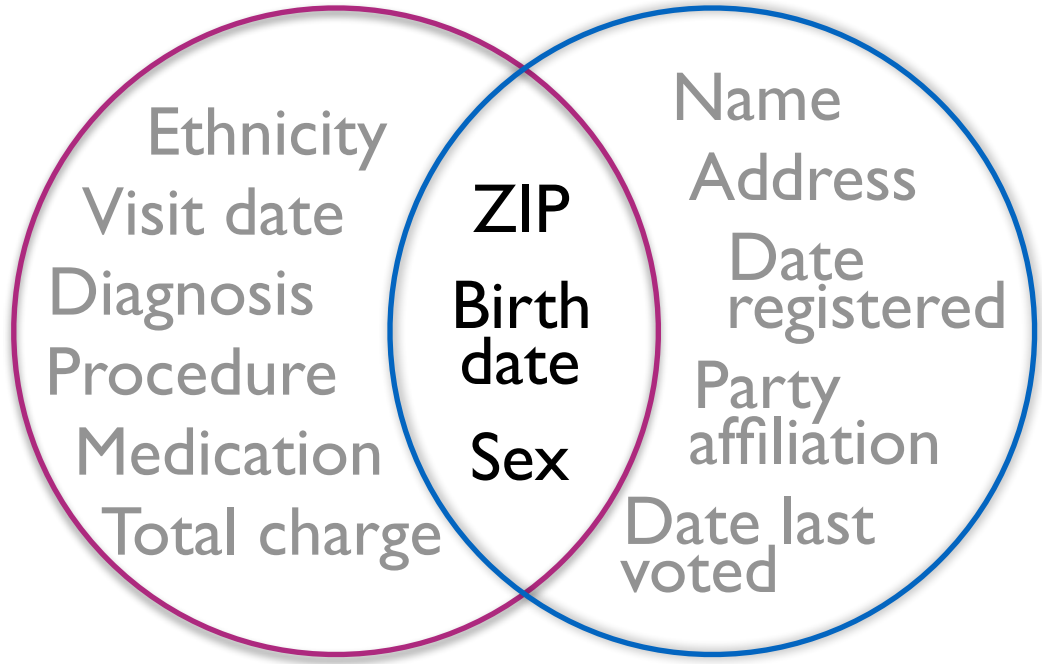
- In 2006, AOL Research released a text file containing twenty million search keywords for over 650,000 users, intended for research purposes.
- The file was anonymized (names where substituted by numbers as pseudonyms), but personally identifiable information was present in many of the queries. The NYT was able to locate an individual from the search records by cross referencing them with phonebook listings
- <<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.", "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake" It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow with three dogs who lives in Lilburn, Ga. >>



Sweeney's de-anonymization attack by linking



Sweeney's de-anonymization attack by linking



87 % of US population is uniquely identifiable by 5-digit ZIP, gender, DOB

This attack has lead to the proposal of k-anonymity (that I will present later)

De-anonymization attacks (II)

Robust De-anonymization of Large Sparse Datasets. Narayanan and Shmatikov, 2008.

Showed the limitations of K-anonymity

De-anonymization of the **Netflix Prize dataset** (500,000 anonymous records of movie ratings), using **IMDB** as the source of background knowledge.

They demonstrated that an adversary who knows just a few preferences about an individual subscriber can identify his record in the dataset.

De-anonymization attacks (III)

De-anonymizing Social Networks.
Narayanan and Shmatikov, 2009.

By using only the network topology, they were able to show that 33% of the users who had accounts on both **Twitter** and **Flickr** could be re-identified in the anonymous Twitter graph with only a 12% error rate.

Statistical Databases

- **The problem:** we want to use databases to get statistical information (aka aggregated information), but without violating the privacy of the people in the database
  - We assume that the database itself is hidden. The only way to access information is by querying it
- For instance, medical databases are often used for research purposes. Typically we are interested in studying the correlation between certain diseases, and certain other attributes: age, sex, weight, etc.
  - A typical query would be: "*Among the people affected by the disease, what percentage is over 60 ?* "
  - Personal queries are forbidden. An example of forbidden query would be: " *Does Don have the disease ?* "

The problem

- Statistical queries should not reveal private information, but it is not so easy to prevent such privacy breaches.

- Example: in a medical database, we may want to ask queries that help to figure the correlation between a disease and the age, but we want to keep private the info whether a certain person has the disease.

name	age	disease
Alice	30	no
Bob	30	no
Don	40	yes
Ellie	50	no
Frank	50	yes

Query:
What is the youngest age of a person with the disease?

Answer:
40

Problem:
The adversary may know that Don is the only person in the database with age 40

The problem

- Statistical queries should not reveal private information, but it is not so easy to prevent such privacy breach.

- Example: in a medical database, we may want to ask queries that help to figure the correlation between a disease and the age, but we want to keep private the info whether a certain person has the disease.

name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

A famous approach to solve this problem: **k-anonymity**. The idea is that the answer should always partitions the space in groups of at least k elements

Alice	Bob
Carl	Don
Ellie	Frank

K-anonymity

- Quasi-identifier: Set of attributes that can be linked with external data to uniquely identify individuals
- Make every record in the table indistinguishable from a least $k-1$ other records with respect to quasi-identifiers. This can be done by:
  - suppression of attributes, and/or
  - generalization of attributes, and/or
  - addition of dummy records
- Linking on quasi-identifiers yields at least k records for each possible value of the quasi-identifier

K-anonymity

Example: 4-anonymity w.r.t. the quasi-identifiers (nationality, ZIP, age)

- achieved by suppressing the nationality and generalizing ZIP and age

	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

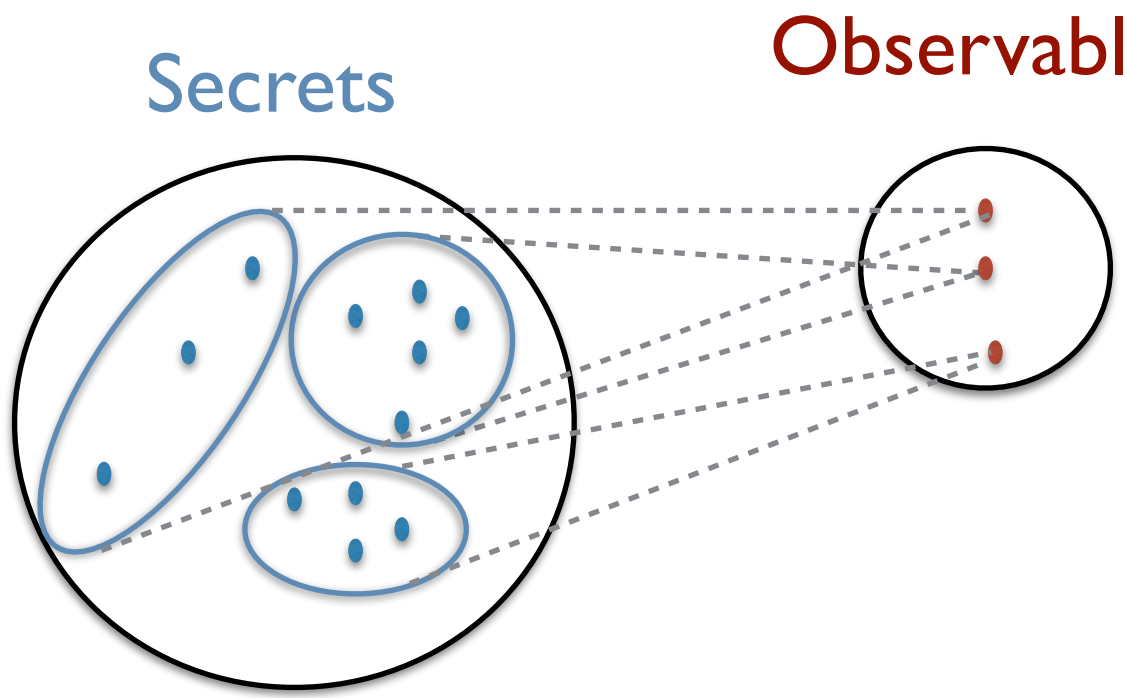
Figure 1. Inpatient Microdata

	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

Correlation: Many-to-one

- Principle: Ensure that there are **many** secret values that correspond to **one** (publicly available) result
- This is the general principle of most deterministic approaches to protection of confidential information (group anonymity, k-anonymity, ℓ-diversity, cloacking, etc.)



The problem

Unfortunately, the many-to-one approach is not robust under **composition**:

name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

Alice	Bob
Carl	Don
Ellie	Frank

The problem of composition

Consider the query:

What is the minimal weight of a person with the disease?

Answer: 100

name	weight	disease
Alice	60	no
Bob	90	no
Carl	90	no
Don	100	yes
Ellie	60	no
Frank	100	yes

Alice	Bob
Carl	Don
Ellie	Frank

The problem of composition

Combine with the two queries:

minimal weight and the minimal age of a person with the disease

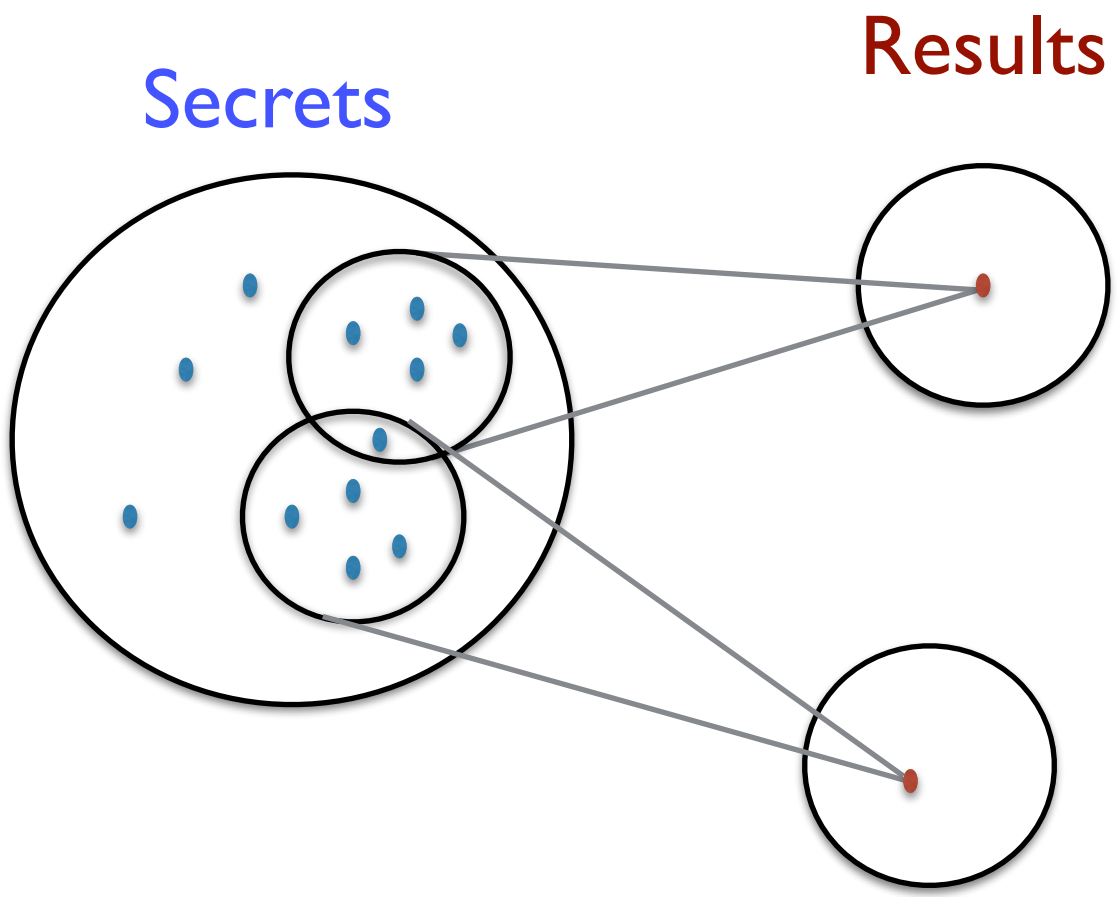
Answers: 40, 100

name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

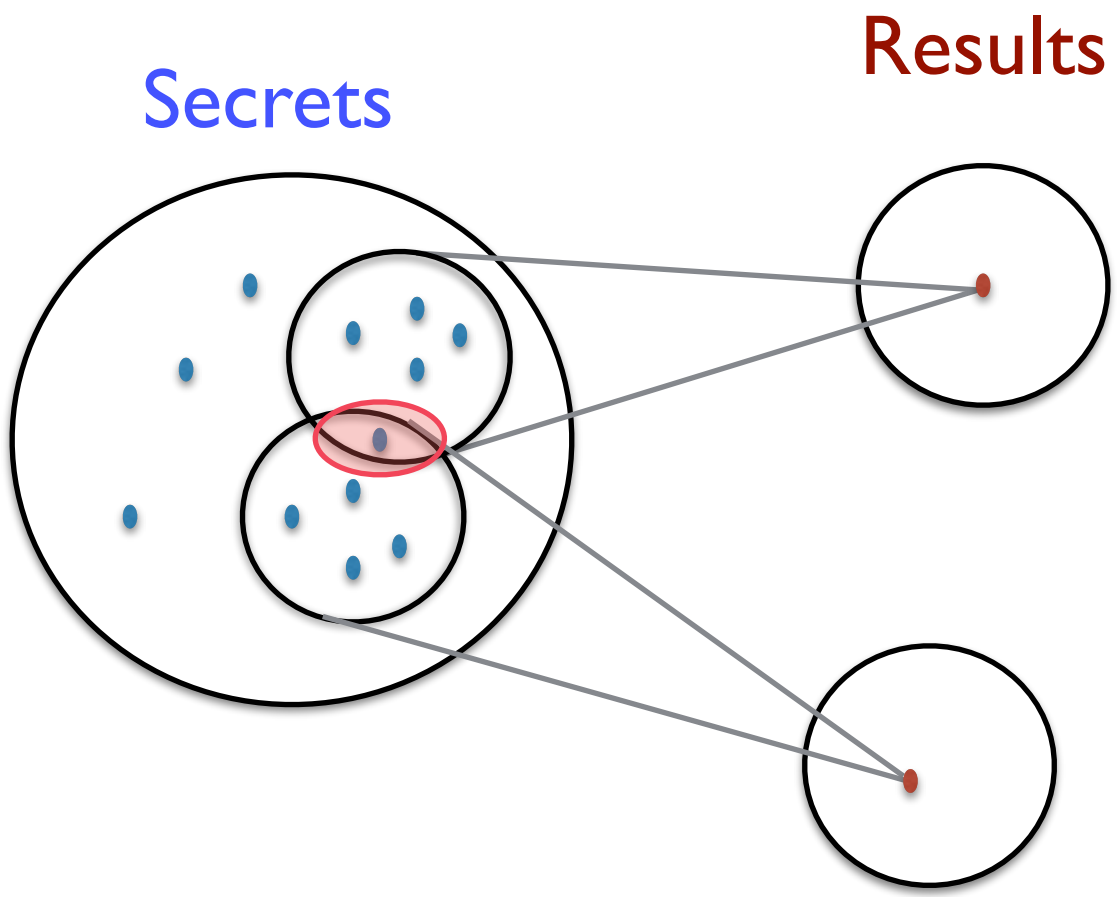
name	weight	disease
Alice	60	no
Bob	90	no
Carl	90	no
Don	100	yes
Ellie	60	no
Frank	100	yes

Alice	Bob
Carl	Don
Ellie	Frank

This is a general problem of the deterministic approaches (based on the principle of many-to-one): the combination of observations determines smaller and smaller intersections on the domain of the secrets, and eventually result in singletones



This is a general problem of the deterministic approaches (based on the principle of many-to-one): the combination of observations determines smaller and smaller intersections on the domain of the secrets, and eventually result in singletones



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

A better solution

Introduce some probabilistic noise on the answer, so that the answers of minimal age and minimal weight can be given also by other people with different age and weight

name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

name	weight	disease
Alice	60	no
Bob	90	no
Carl	90	no
Don	100	yes
Ellie	60	no
Frank	100	yes

Alice	Bob
Carl	Don
Ellie	Frank

Noisy answers

minimal age:

40 with probability 1/2

30 with probability 1/4

50 with probability 1/4

name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

Alice	Bob
Carl	Don
Ellie	Frank

Noisy answers

minimal weight:

100 with prob. 4/7

90 with prob. 2/7

60 with prob. 1/7

name	weight	disease
Alice	60	no
Bob	90	no
Carl	90	no
Don	100	yes
Ellie	60	no
Frank	100	yes

Alice	Bob
Carl	Don
Ellie	Frank

Noisy answers

Combination of the answers

The adversary cannot tell for sure whether a certain person has the disease

name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

name	weight	disease
Alice	60	no
Bob	90	no
Carl	90	no
Don	100	yes
Ellie	60	no
Frank	100	yes

Alice	Bob
Carl	Don
Ellie	Frank

Noisy mechanisms

- The mechanisms reports an approximate answer, typically generated randomly on the basis of the true answer and of some probability distribution
- The probability distribution must be chosen carefully, in order to not destroy the utility of the answer
- A good mechanism should provide a good trade-off between privacy and utility. Note that, for the same level of privacy, different mechanisms may provide different levels of utility.

Randomization

- In this course, we will consider the general case of probabilistic systems (note that the deterministic oness can be seen as a special case), and develop quantitative (probabilistic) foundations
  - Randomization is often used in protection mechanisms, as it is quite effective in obfuscating the link between public and private information (aka observables and secret information)
  - We need to reason about the knowledge of the adversary, which can often be represented in terms of a probability distribution on the set of the possible values of the secret (probabilistic knowledge)

Exercise. Bob wants to find out whether Don is affected by a certain disease d. He knows Don's age and weight, and that Don is going to check in a hospital that maintains an anonymized database of all patients, and that can be queried with queries of the form:

- How many patients are affected by the disease d ?

- What is the average age and weight of the patients affected by the disease d?

Discuss whether Bob can determine, with high probability, whether Don has the disease. What kind of background information Don needs? What kind of queries should he ask?

Research internships

We have various internship (stage) subjects, ranging from rather theoretical to rather practical

- Privacy-friendly Machine Learning
  - Focus on the Bayesian methods for ML (also DP is based on Bayesian principles)
- Bisimulation metrics for analysis of leakage in concurrent systems
- Location privacy: various research directions
  - Collaboration with Renault R&D

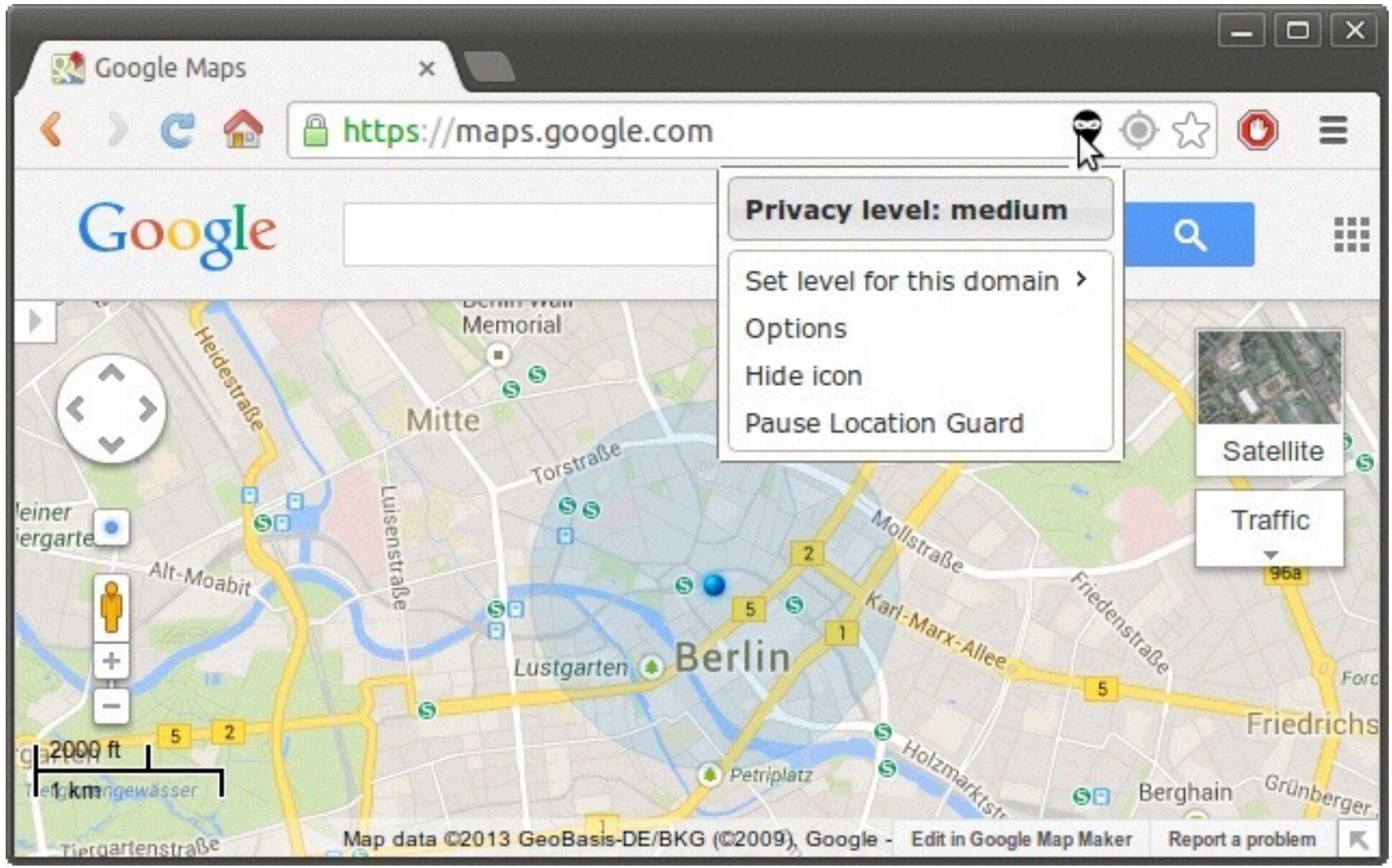
Research internships

- Location of the internship : LIX, Ecole Polytechnique, within an Equipe INRIA
- The internships will be “remunerés”
- It will be possible to continue the research as a PhD student

Tool: "Location Guard"

http://www.lix.polytechnique.fr/~kostas/software.html

About 50,000 active users to date



Location guard for Chrome