# Protection of Sensitive Information

Catuscia Palamidessi

catuscia@lix.polytechnique.fr
http://www.lix.polytechnique.fr/~catuscia/

Class 1

# Logistic Information

- The course will be in English
- The grade of the course will be based on:
  - a brief talk (30min) on a paper of your choice among a list of relevant papers that I will put on line shortly
  - a written exam at the end of the course, or a project
- I will propose 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
- I will put the slides on line after every class
- Please feel free to ask questions any time. I love questions, they help to make the class more interactive and lively

# Plan of the lectures

- Motivations and a bit of history
- Differential Privacy and Extensions
- Anonymity
- Location Privacy
- Quantitative Information Flow

# 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

IP address ⇒ location.

History of requests ⇒ interests.

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

Power consumption (smart meters) ⇒ activities at home.

# Motivations

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

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

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).

# Different types of sensitive data

- Sensitive information about an individual :
  - credit card / bank information, valuable belongings, vulnerable behaviors, …
    - sensitive because it can bring to attacks to the person or his properties
  - ethnicity, religious beliefs, political opinions, medical status, …
    - 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

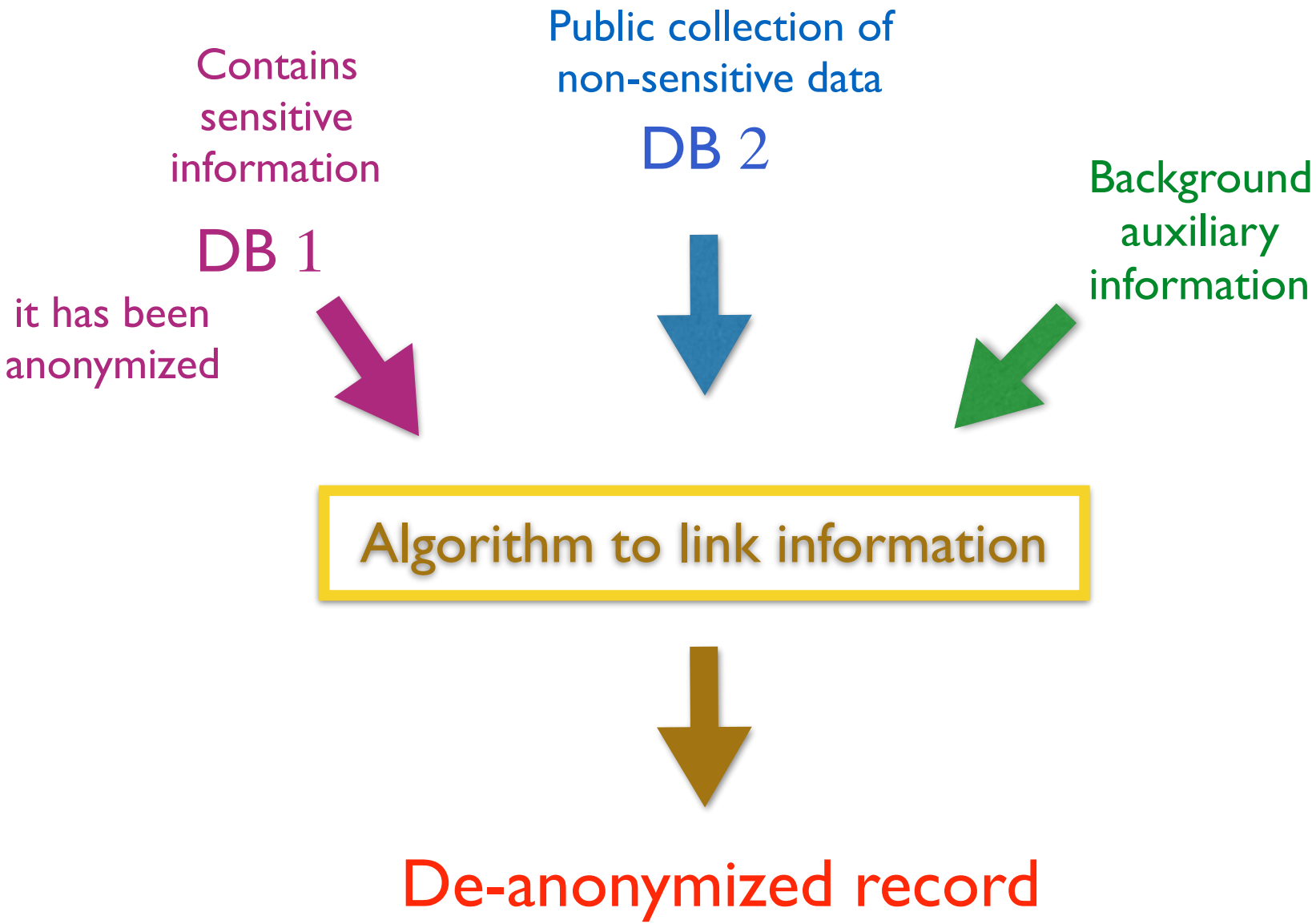
- Traditionally, privacy is protected via:
  - anonymization
  - encryption
  - access control
- However, these methods often fail:
  - encryption and access control cannot protect against the inference of private information from public information
  - anonymization has been proved highly ineffective

# Privacy via anonymity

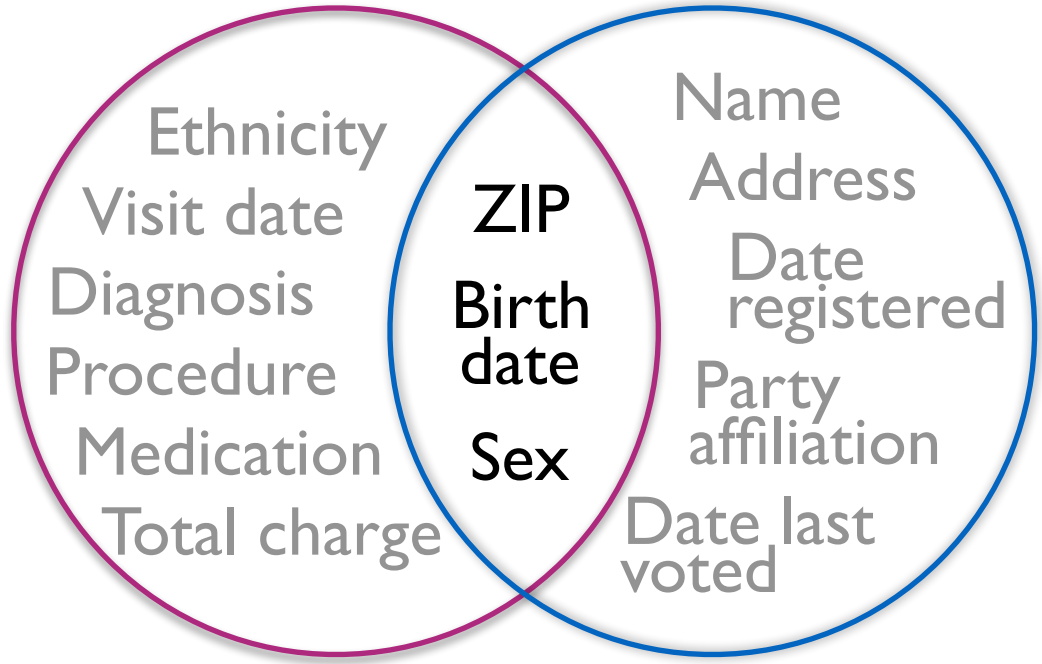
- Organizations that collect / release large collections sanitize the data by removing all personal identifiers: name, address, SSN, … Thus the data are anonymous and they claim that there is no risk for privacy.

- This method has been proved highly ineffective.
  The quasi-identifiers allow to retrieve the identity in a large number of cases.
  - Several famous de-anonymization attacks have been carried out in the last decade.

# 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

Another famous attack is that by Narayanan et Smatikov. They showed that by combining the information of two popular social network (Twitter and Flickr) they were able to de-anonymize a large percentage of the users (about 80%) and retrieve their private information with only a small probability of error (12%).

De-anonymizing Social Networks, Arvind Narayanan and Vitaly Shmatikov. Security & Privacy '09.

# 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 |

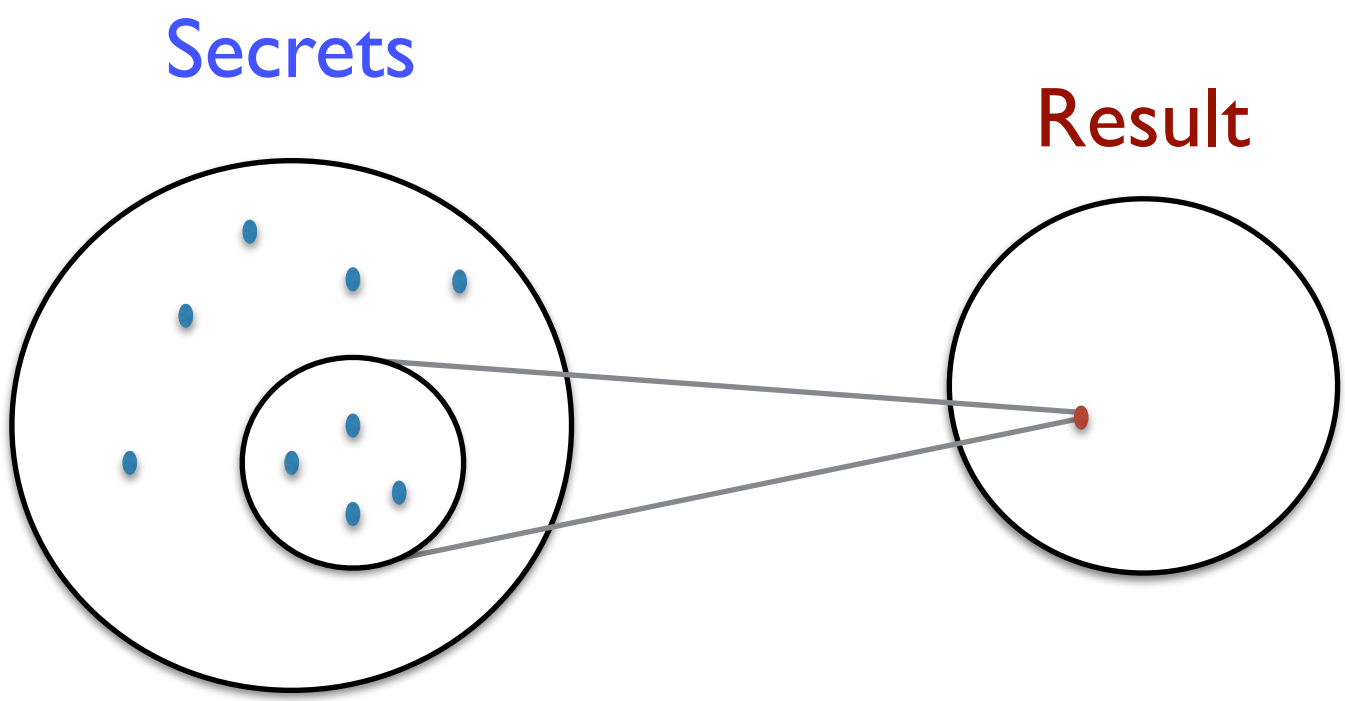
Figure 1. Inpatient Microdata

| | 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, $\ell$-anonymity, 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 |

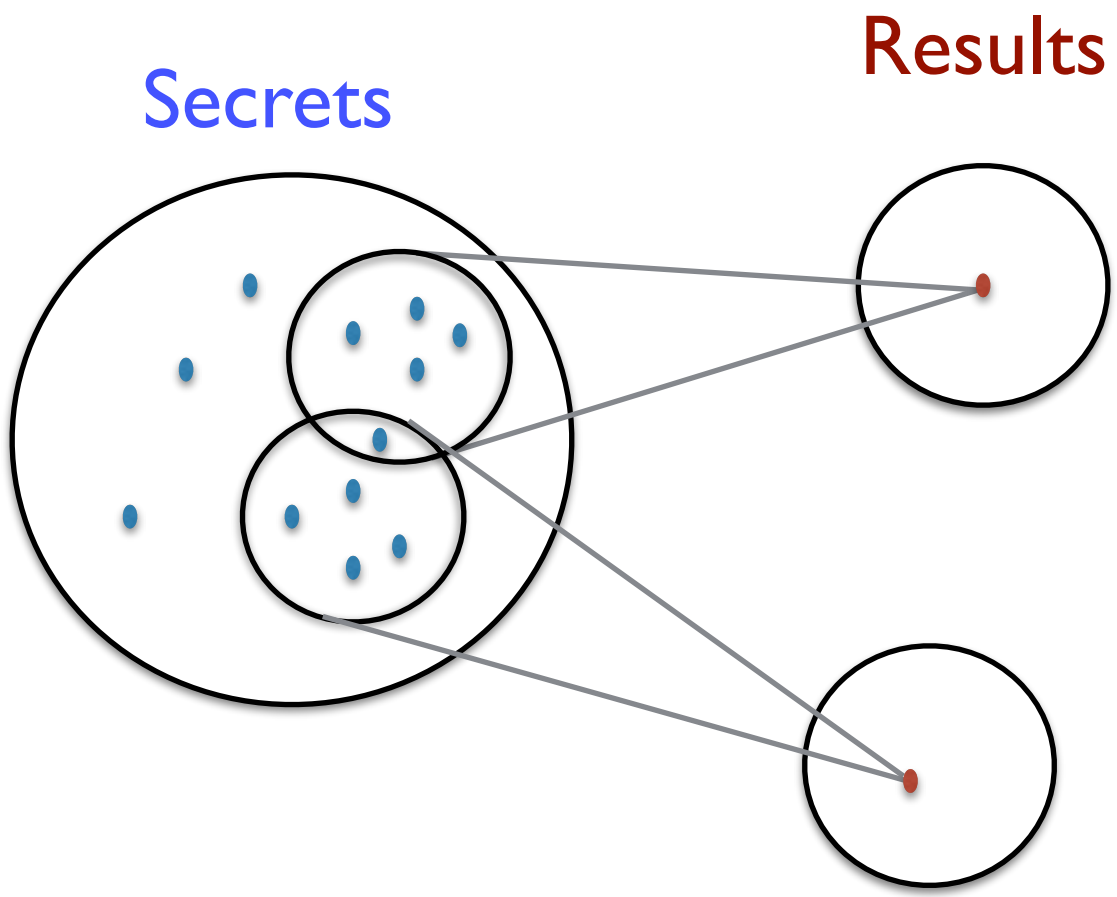
# Composition attacks

Attacks of this kind are called

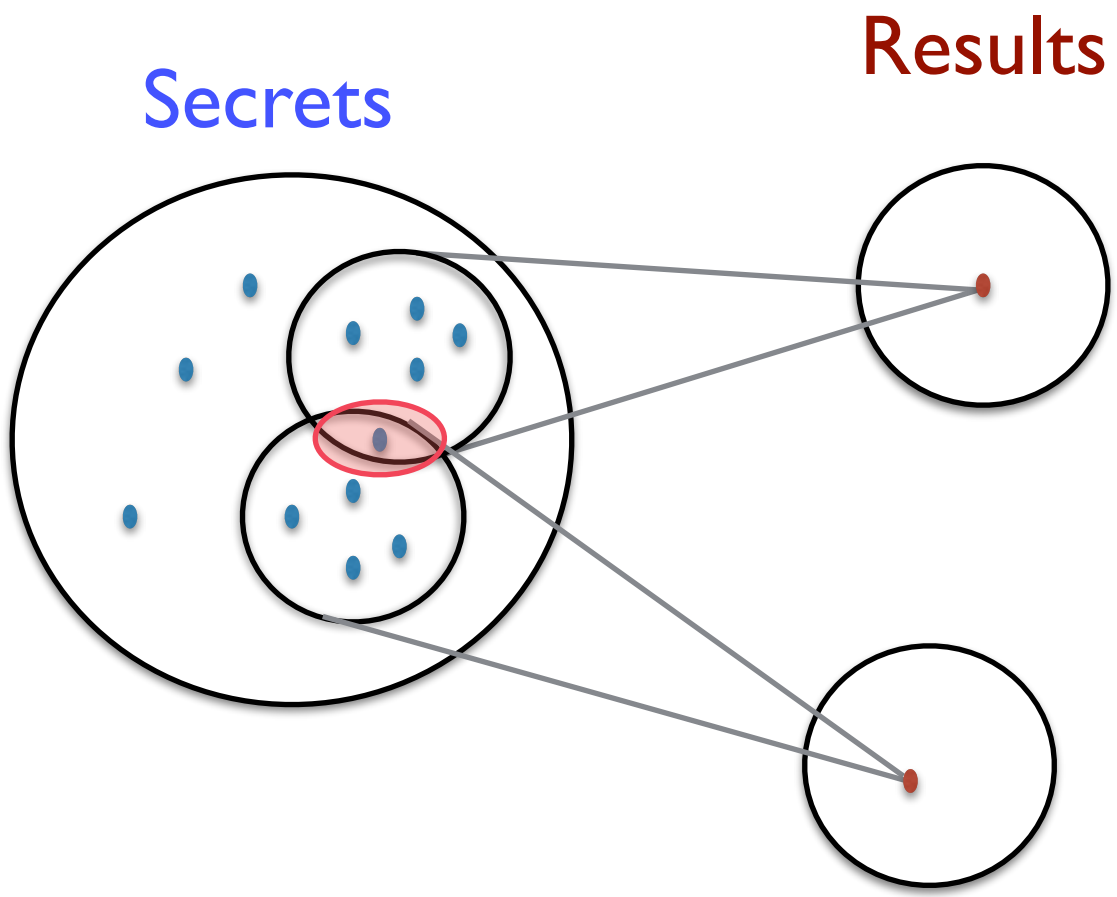
**composition attacks**

the adversary combine information deriving from different sources

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 ones can be seen as a special case), and study the quantitative (probabilistic) foundations of privacy and information leakage. The reasons are:
  - 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 and of the user, as these may influence privacy and utility, respectively. This is usually 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 that Don is going to check in a hospital that maintains a 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 of the patients affected by the disease d?

Is it possible for Bob to determine, with high probability, whether Don has the disease? If you answer yes, what is the strategy? If you answer no, what other kind of queries or knowledge should Bob have at his disposal?