Quantitative approaches to information protection

Course in Pisa, April 2014
Lecture 5

Differential Privacy

- Differential privacy [Dwork et al.,2006] is a notion of privacy originated from the area of **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

The problem

- The statistical queries should not reveal private information.

- Example: in a database meant to study a certain disease, we may want to ask queries that reveal the correlation between the disease and the age, but we should not be able to derive from this 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

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

Answer:
40

The problem

- The statistical queries should not unveil private information.

- Example: in a database meant to study a certain disease, we may want to ask queries that reveal the correlation between the disease and the age, but we should not be able to derive from this 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

k-anonymity: the answer always partition the space in groups of at least k elements

Alice	Bob
Carl	Don
Ellie	Frank

The problem

Unfortunately, k-anonymity is very fragile 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

Solution

Introduce some 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 answers: a typical mechanism

- Randomized mechanism for a query $f: \mathcal{X} \to \mathcal{Y}$.
  Instead of the exact answer to the query, the curator gives a randomized answer $\mathcal{K}: \mathcal{X} \to \mathcal{Z}$ ($\mathcal{Z}$ may be different from $\mathcal{Y}$)
- The principle: little noise in global info produces large noise in individual info
- A typical randomized method: **the Laplacian noise.** If the exact answer is y, the reported answer is z, with a probability density function defined as:

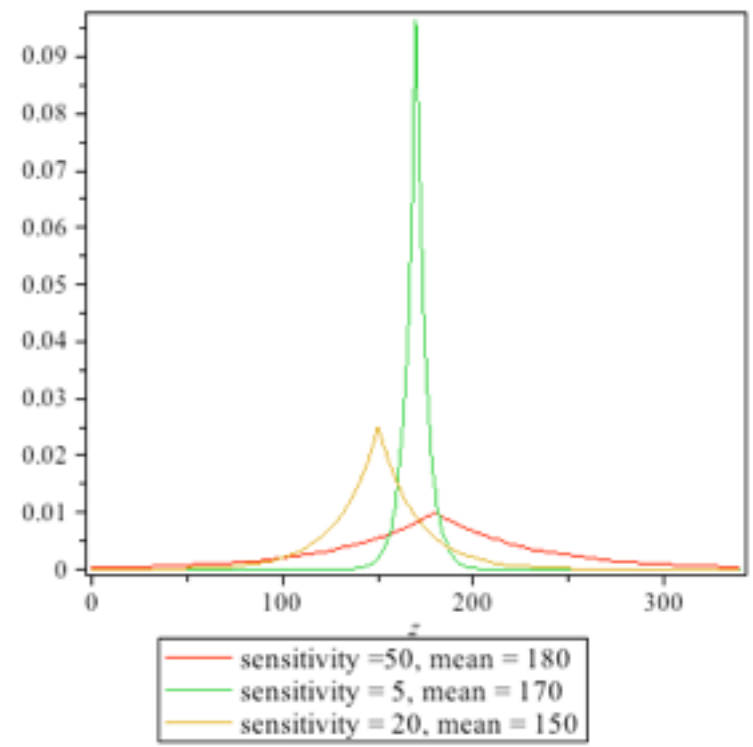
$$dP(z) = c\, e^{-\frac{|z-y|}{\Delta f}}$$

where Δf is the *sensitivity* of f:

$$\Delta f = \max_{x \sim x' \in \mathcal{X}} |f(x) - f(x')|$$

and c is a normalization factor:

$$c = \frac{1}{2\,\Delta f}$$



Privacy and Utility

- The two main criteria by which we judge a randomized mechanism:
  - **Privacy:** how good is the protection against leakage of private information
  - **Utility:** how useful is the reported answer

- Clearly there is a trade-off between privacy and utility, but they are not the exact opposites: privacy is about the individual data, while utility is about the aggregate data.

Differential Privacy

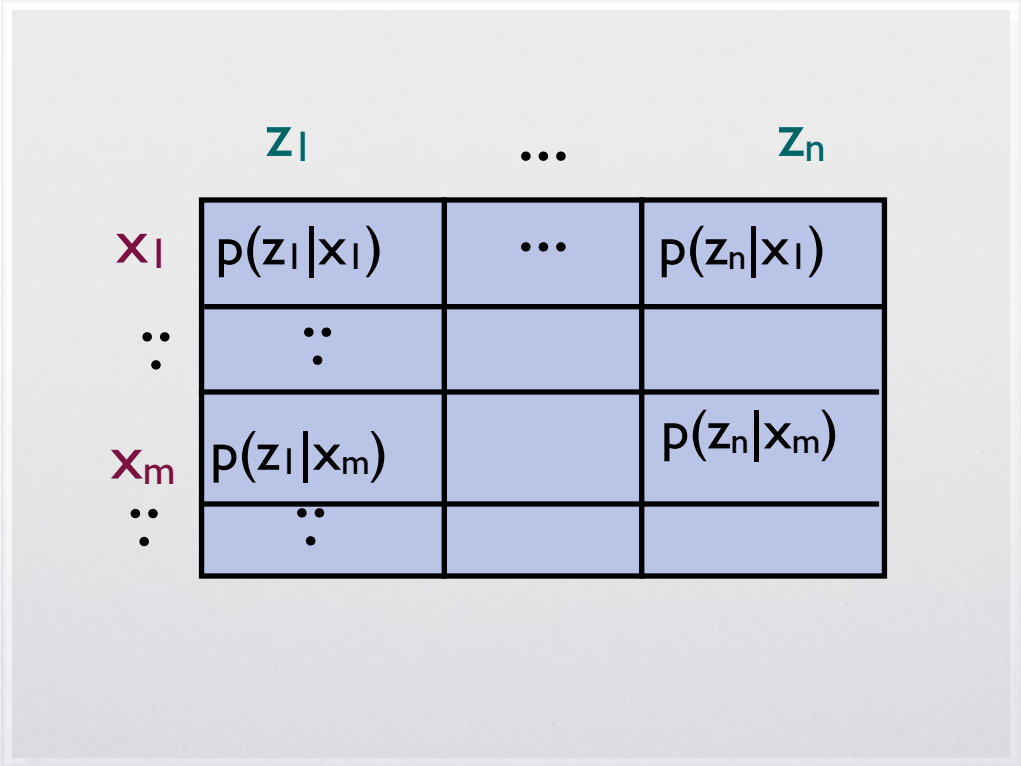
- There have been various attempts to formalize the notion of privacy, but the most successful one is the notion of Differential Privacy, recently introduced by Dwork
- **Differential Privacy** [Dwork 2006]: a randomized function $\mathcal{K}$ provides ε-differential privacy if for all adjacent databases x, x', and for all $z \in \mathcal{Z}$, we have

$$\frac{p(K = z|X = x)}{p(K = z|X = x')} \leq e^{\epsilon}$$

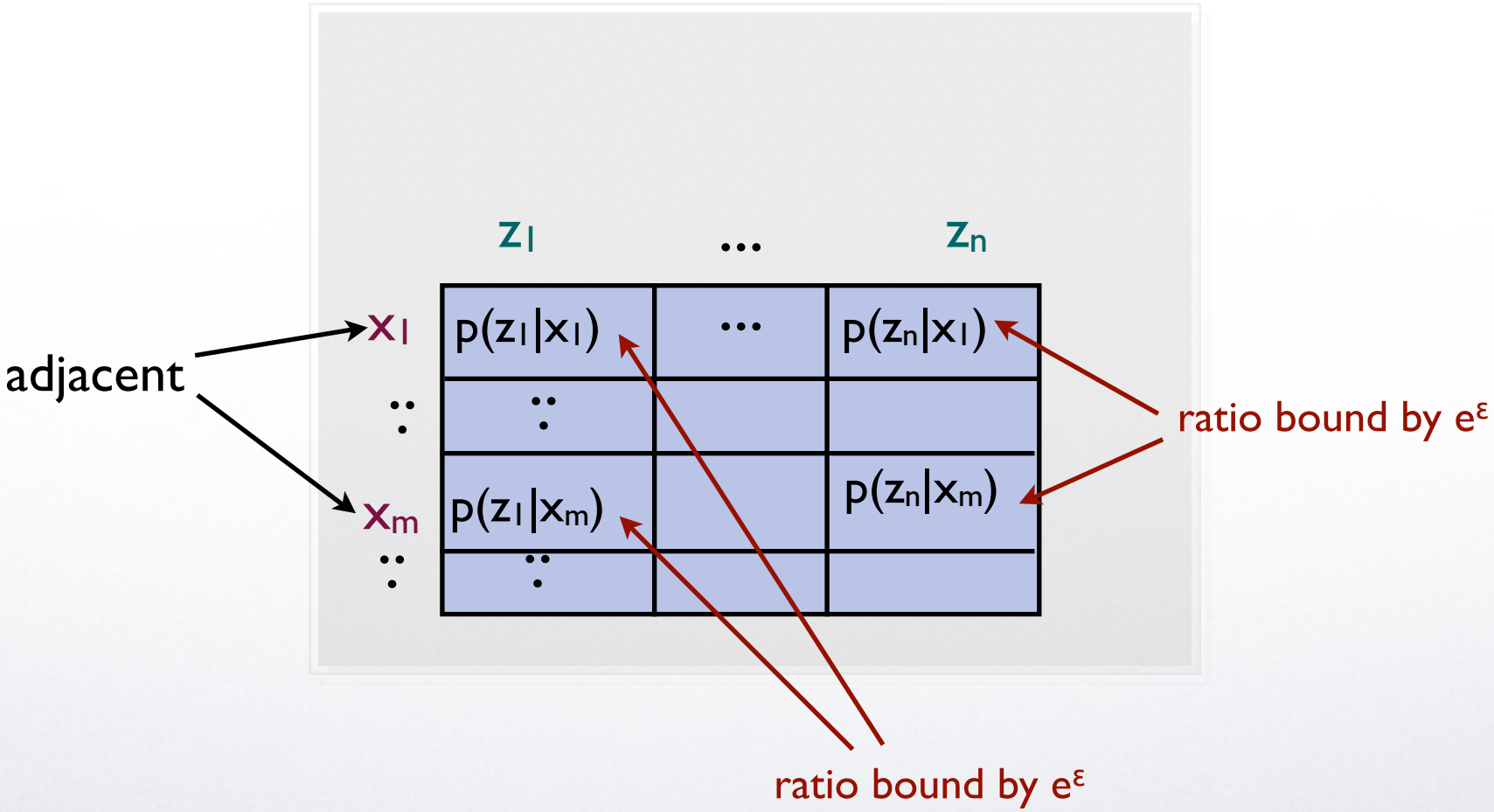
- The idea is that the likelihoods of x and x' are not too far apart, for every S
- Differential privacy is robust with respect to composition of queries.

$\mathcal{K}$ can be seen as a noisy channel, in the information-theoretic sense from the domain $\mathcal{X}$ of databases to the domain $\mathcal{Z}$ of reported answers

Channel matrix



Differential privacy on the channel matrix



Differential Privacy: alternative definition

- Perhaps the notion of differential privacy is easier to understand under the following equivalent characterization.
- In the following, X_i is the random variable representing the value of the individual i, and $X_{\neq i}$ is the random variable representing the value of all the other individuals in the database

- **Differential Privacy, alternative characterization:** a randomized function $\mathcal{K}$ provides ε-differential privacy if:

$$\text{for all } x \in \mathcal{X}, z \in \mathcal{Z}, p_i(\cdot)$$

$$\frac{1}{e^\epsilon} \leq \frac{p(X_i = x_i | X_{\neq i} = x_{\neq i})}{p(X_i = x_i | X_{\neq i} = x_{\neq i} \wedge K = z)} \leq e^\epsilon$$

Utility

The reported answer, i.e. the answer given by the randomized function, should allow to approximate the true (i.e. the exact) answer to some extent

Z = reported answer; Y = exact answer

Utility: $$\mathcal{U}(Y,Z) = \sum_{y,z} p(y,z)\, gain(y, remap(z))$$

The remap allows the user to use side information (i.e. a priori pb) to maximize utility

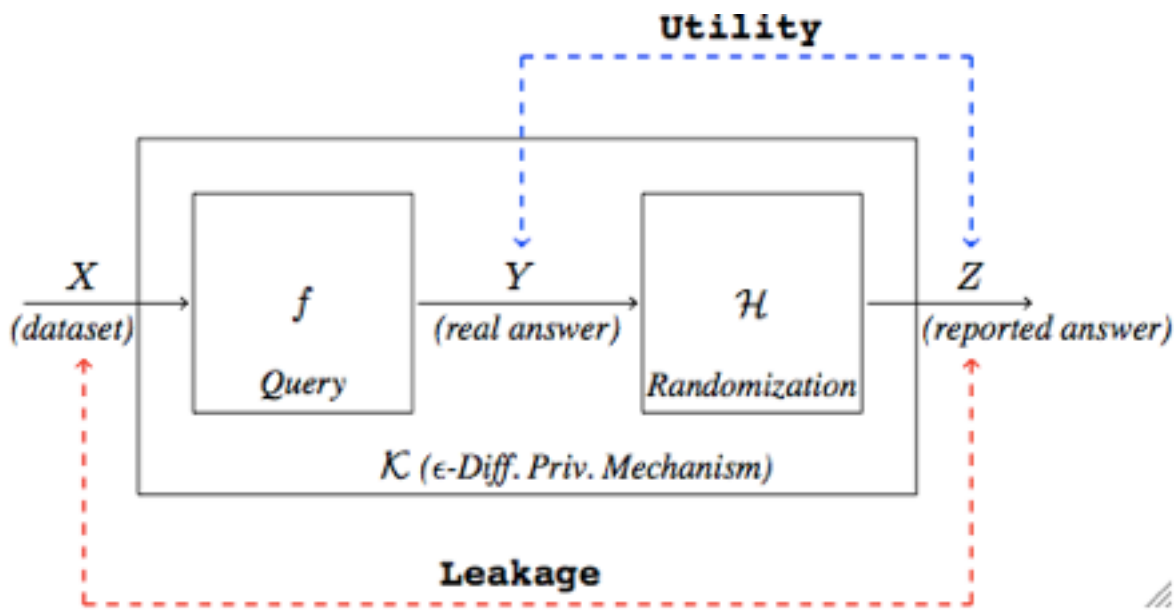
Example: **binary gain function**: $$gain(y_1, y_2) = \begin{cases} 1 & y_1 = y_2 \\ 0 & y_1 \neq y_2 \end{cases}$$

In the binary case the utility is **the expected value of the probability of success** to obtain the true answer (i.e. the Bayes vulnerability)

Oblivious mechanisms

- Given $f: \mathcal{X} \to \mathcal{Y}$ and $\mathcal{K}: \mathcal{X} \to \mathcal{Z}$, we say that $\mathcal{K}$ is oblivious if it depends only on $\mathcal{Y}$ (not on $\mathcal{X}$)

- If $\mathcal{K}$ is oblivious, it can be seen as the composition of f and a randomized mechanism $\mathcal{H}$ defined on the exact answers $\mathcal{K} = f \times \mathcal{H}$



- Privacy concerns the information flow between the databases and the reported answers, while utility concerns the information flow between the correct answer and the reported answer

Differential Privacy and Utility

The fact that privacy and utility are not the exact opposite means that for the same utility we can have mechanisms with different degrees of utility

⇨ **Important research direction: how to increase utility while preserving the intended degree of privacy**

Two fundamental results

1. [Ghosh et al., STOC 2009] The geometric mechanism is **universally optimal** in the case of counting queries, with respect to all (reasonable) notions of utility

   - Counting queries are of the form "how many individuals in the DB satisfy the property P ?"
   - universally optimal means that it provides the best utility, for a fixed ε of differential privacy, for all the a priori distributions (side information)
   - the geometric mechanism is the discrete version of the Laplacian

Two fundamental results

2. [Brenner and Nissim, STOC 2010] The counting queries are practically the only kind of queries for which there exists a universally optimal mechanism

   - This means that for other kind of queries one can only construct optimal mechanisms for specific a priori distributions (side information).
   - The precise characterization is given in terms of the graph structure that the adjacency relation induces on the answer space:
     - line: ok
     - loops: not ok
     - trees: not ok

Some bibliography

M. S. Alvim, M. E. Andrés, K. Chatzikokolakis, and C. Palamidessi. Quantitative Information Flow and Applications to Differential Privacy. In A. Aldini and R. Gorrieri, editors, Foundations of Security Analysis and Design VI – FOSAD Tutorial Lectures, volume 6858 of Lecture Notes in Computer Science, pages 211–230. Springer, 2011.

M. S. Alvim, K. Chatzikokolakis, C. Palamidessi, and G. Smith. Measuring information leakage using generalized gain functions. In Proceedings of the 25th IEEE Computer Security Foundations Symposium, pages 265–279, 2012.

M. Boreale, F. Pampaloni, and M. Paolini. Asymptotic information leakage under one-try attacks. In M. Hofmann, editor, Proceedings of the 14th International Conference on the Foundations of Software Science and Computational Structures, volume 6604 of Lecture Notes in Computer Sci- ence, pages 396–410. Springer, 2011.

C. Dwork. A firm foundation for private data analysis. Communications of the ACM, 54(1):86–96, 2011.

B. Köpf and D. A. Basin. An information-theoretic model for adaptive side-channel attacks. In P. Ning, S. D. C. di Vimercati, and P. F. Syverson, editors, Proceedings of the 2007 ACM Conference on Computer and Communications Security, pages 286–296. ACM, 2007.

Pasquale Malacaria. Algebraic Foundations for Quantitative Information Flow. To appear in MSCS. 2013

Annabelle McIver, Carroll Morgan, Geoffrey Smith, Barbara Espinoza, and Larissa Meinicke, Abstract channels and their robust information-leakage ordering. Proc. of the Int. Conference on Principles of Security and Trust, Grenoble, France, April 2014.

G. Smith. On the foundations of quantitative information flow. In L. de Alfaro, editor, Proceedings of the 12th International Conference on Foundations of Software Science and Computation Structures, volume 5504 of LNCS, pages 288–302, York, UK, 2009. Springer.

Thank you !