TDDD17
Information Security

Topic: Database Privacy

Olaf Hartig

olaf.hartig@liu.se

Acknowledgement: Many of the slides in this slide set are adaptations of lecture slides of Prof. Johann-Christoph Freytag (Humboldt Universität zu Berlin).

LINKÖPING UNIVERSITY

What is Privacy?

Definitions of Privacy

- Alan Westin, Privacy and Freedom, 1967

  "Privacy is the claim of individuals, groups or institutions to determine for themselves when, how, and to what extent information about them is communicated to others."

- Control over information
- Relevant when you give personal information on a Web site (agree to privacy policy of the Web site)
- You may not always have control
  - e.g., personal health information

Definitions of Privacy (cont'd)

- Latanya Sweeney, in Int. Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002

  "Privacy reflects the ability of a person, organization, government, or entity to control its own space, where the concept of space takes on different contexts."

- Examples of privacy spaces:
  - Physical space (e.g., against invasion)
  - Bodily space (e.g., medical consent)
  - Computer space (e.g., spam)
  - Web browsing space (Internet privacy)

Dimensions of Privacy

- Personal privacy
  - Protecting a person against undue interference (e.g., physical searches) and information that violates his/her moral sense
- Territorial privacy
  - Protecting a physical area surrounding a person that may not be violated without the acquiescence of the person
- Informational privacy
  - Deals with the gathering, compilation, and selective dissemination of information

Privacy and Utility

- Ruth Gavison, Privacy and the Limits of Law, 1980

  "We start from the obvious fact that both perfect privacy and total loss of privacy are undesirable. Individuals must be in some intermediate state – a balance between privacy and interaction […] Privacy thus cannot be said to be a value in the sense that the more people have of it, the better."

- Balance between privacy and utility
  - e.g., health data could be shared with medical researchers

Picture source: https://www.flickr.com/photos/61056899@N06/5751301741

Example

The Massachusetts Governor Privacy Breach

Latanya Sweeney: *Achieving k-Anonymity Privacy Protection Using Generalization and Suppression*. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(5), 2002.

Massachusetts Governor Privacy Breach

- In Massachusetts, USA, the Group Insurance Commission (GIC) is responsible for purchasing health insurance for state employees
- GIC has to publish the data:

GIC	ZIP	DOB	Sex	Diagnostic	Medication	...

Sweeney's Experiment

- Is it always obvious that privacy is violated/breached?
- Sweeney paid $20 to buy the voter registration list for Cambridge, MA

VOTER	Name	Address	...	ZIP	DOB	Sex

Sweeney's Findings

- William Weld (former governor of MA) lives in Cambridge, hence is in VOTER
- 6 people in VOTER share his date of birth (dob)
- only 3 of them were man (same sex)
- Weld was the only one in that zip
- Sweeney learned Weld's medical records!

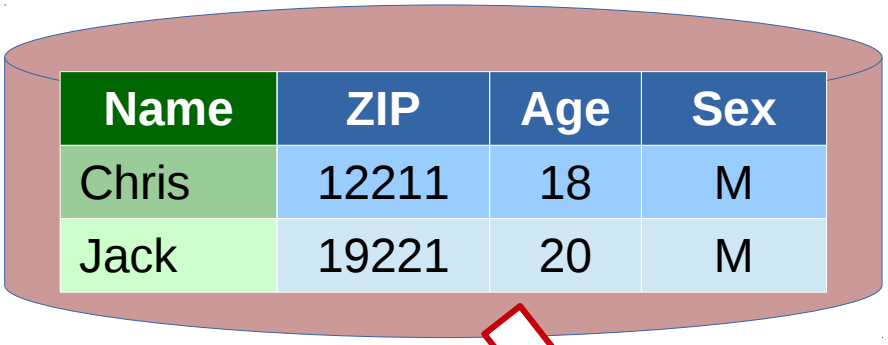
				GIC	ZIP	DOB	Sex	Diagnostic	Medication	...
VOTER	Name	Address	...	ZIP	DOB	Sex			

- 87 % of US population can be identified by the combination of ZIP, DOB, and sex

Basic Terminology and Goals of Database Privacy

Definition: Quasi-Identifier

A set of non-sensitive attributes $QI_T = \{A_i, \ldots, A_j\}$ of a table T is called a quasi-identifier if these attributes can be linked with external data to uniquely identify at least one individual in the general population Ω.



Name	ZIP	Age	Sex
Chris	12211	18	M
Jack	19221	20	M

public database

ZIP	Age	Sex	Disease
12211	18	M	Arthritis
12244	19	M	Cold
...	...	...	...

T

Name	ZIP	Age	Sex	Disease
Chris	12211	18	M	Arthritis

Ω = {Chris, David, Jack, … }

QI_T = {ZIP, Age, Sex}

Challenge

- **Given:** person-specific data *T*

SSN	Name	ZIP	Age	Sex	Disease
003	Chris	12211	18	M	Arthritis
004	David	12244	19	M	Cold
010	Ethan	12245	27	M	Heart problem
029	Frank	12377	27	M	Flu
034	Gillian	12377	27	F	Arthritis
059	Helen	12391	34	F	Diabetes
077	Ireen	12391	45	F	Flu

identifier | quasi-identifier | sensitive attributes

- **Goal:** privacy-preserving public release table *T**
  - Information should remain practically useful

Picture source: https://www.flickr.com/photos/61056899@N06/5751301741

k-Anonymity

Latanya Sweeney: *Achieving k-Anonymity Privacy Protection Using Generalization and Suppression*. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(5), 2002.

Definition

A table T satisfies k-anonymity if for every tuple t in T there exist (at least) k–1 other tuples t_1, t_2, …, t_{k-1} in T such that we have $t[\mathrm{QI}_T] = t_1[\mathrm{QI}_T] = t_2[\mathrm{QI}_T] = t_{k-1}[\mathrm{QI}_T]$ for each quasi-identifier QI_T.

ZIP	Age	Sex	Disease
12211	18	M	Arthritis
12244	19	M	Cold
12245	27	M	Heart problem
12377	27	M	Flu
12377	27	F	Arthritis
12391	34	F	Diabetes
12391	45	F	Flu

T

ZIP	Age	Sex	Disease
122**	18-19	M	Arthritis
122**	18-19	M	Cold
*	27	*	Heart problem
*	27	*	Flu
*	27	*	Arthritis
12391	≥ 30	F	Diabetes
12391	≥ 30	F	Flu

2-anonymous table T^*

Example

QI group / equivalence class

ZIP	Age	Sex	Disease
122**	18-19	M	Arthritis
122**	18-19	M	Cold
*	27	*	Heart problem
*	27	*	Flu
*	27	*	Arthritis
12391	≥ 30	F	Diabetes
12391	≥ 30	F	Flu

2-anonymous table *T**

Name	ZIP	Age	Sex
Chris	12211	18	M
Jack	19221	20	M

public database

Name	ZIP	Age	Sex	Disease
Chris	12211	18	M	Arthritis
Chris	12211	18	M	Cold

Disease of Chris?
Arthritis or cold?

Privacy vs. Utility

ZIP	Age	Sex	Disease
122**	18-19	M	Arthritis
122**	18-19	M	Cold
*	27	*	Heart problem
*	27	*	Flu
*	27	*	Arthritis
12391	≥ 30	F	Diabetes
12391	≥ 30	F	Flu

2-anonymous table
high information content

ZIP	Age	Sex	Disease
*	≤ 19	M	Arthritis
*	≤ 19	M	Cold
*	18-65	*	Heart problem
*	18-65	*	Flu
*	18-65	*	Arthritis
12***	≥ 20	*	Diabetes
12***	≥ 20	*	Flu

2-anonymous table
low information content

- Optimization problem: achieving *k*-anonymity by hiding the minimum amount of information
  - L. Sweeney: *Achieving k-Anonymity Privacy Protection Using Generalization and Suppression*. Int. Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002
  - G. Aggarwal et al.: *Approximation Algorithms for k-Anonymity*. Journal of Privacy Technology, 2005

Two Types of Information Disclosure

- Identity disclosure: individual can be linked to a particular record in the released table
  - Achieved by *k*-anonymity
- Attribute disclosure: learning something new about an individual or a group of individuals
  - i.e., the released data makes it possible to infer the characteristics of an individual more accurately than it would be possible without the data release

Example: Attribute Disclosure

ZIP	Age	Sex	Disease
12211	18	M	Heart disease
12244	19	M	Heart disease
12245	19	M	Heart disease
12245	27	M	Cancer
12377	27	F	Arthritis
12377	27	F	Diabetes
12391	34	F	Breast cancer
12391	45	F	Flu
12391	47	M	Flu

T

ZIP	Age	Sex	Disease
122**	18-19	M	Heart disease
122**	18-19	M	Heart disease
122**	18-19	M	Heart disease
12***	27	*	Cancer
12***	27	*	Arthritis
12***	27	*	Diabetes
12391	≥ 30	*	Breast cancer
12391	≥ 30	*	Flu
12391	≥ 30	*	Flu

3-anonymous table *T**

Example: Attribute Disclosure (cont'd)

ZIP	Age	Sex	Disease
122**	18-19	M	Heart disease
122**	18-19	M	Heart disease
122**	18-19	M	Heart disease
12***	27	*	Cancer
12***	27	*	Arthritis
12***	27	*	Diabetes
12391	≥ 30	*	Breast cancer
12391	≥ 30	*	Flu
12391	≥ 30	*	Flu

Name	ZIP	Age	Sex
Chris	12211	18	M
Jack	19221	20	M

public database

3-anonymous table *T**

Chris = ?

ZIP	Age	Sex	Disease
12211	18	M	Heart disease
12211	18	M	Heart disease
12211	18	M	Heart disease

no identity disclosure

Disease of Chris?
→ heart disease

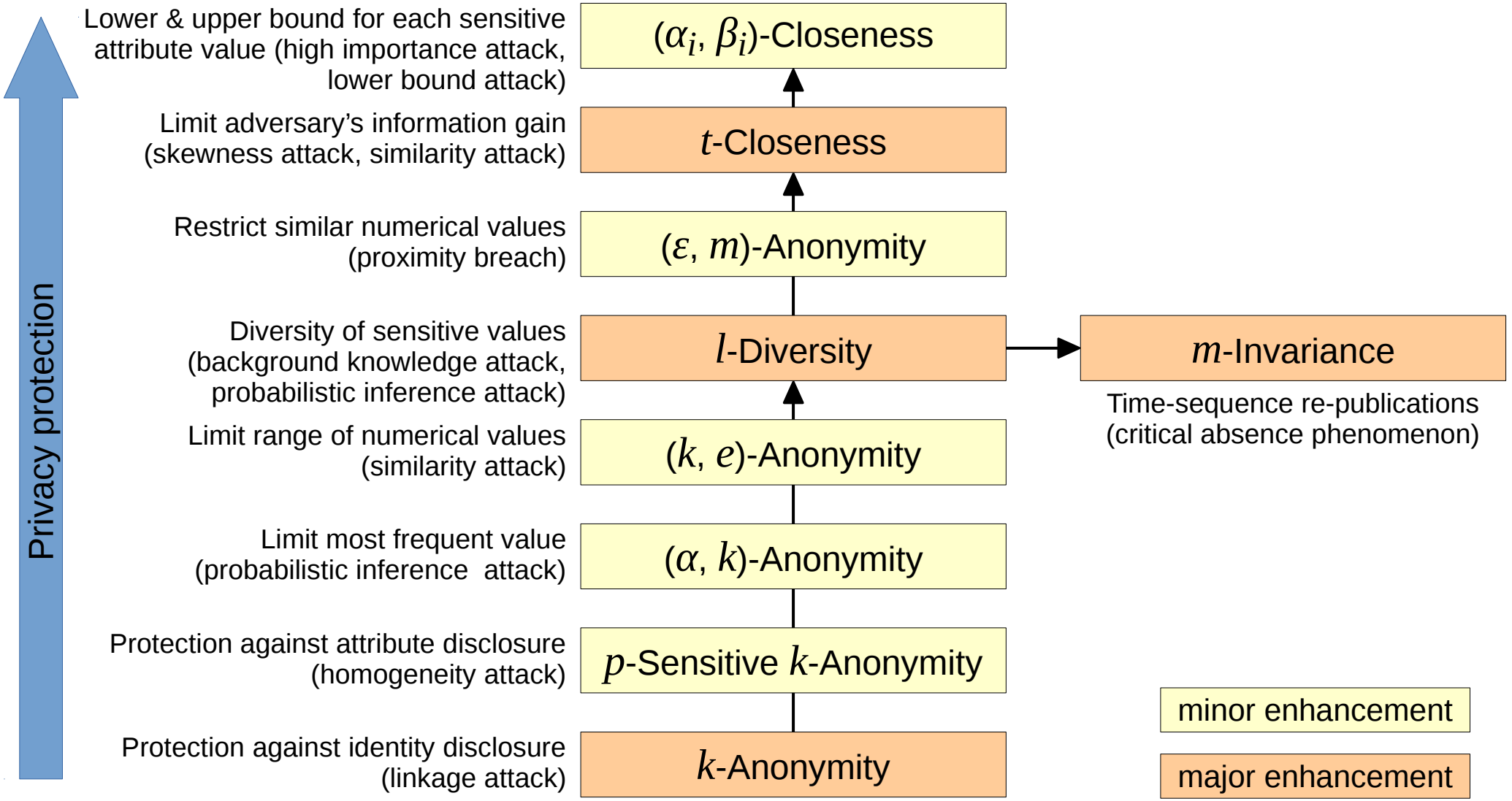
no protection against attribute disclosure

Two Types of Information Disclosure (cont'd)

- Identity disclosure: individual can be linked to a particular record in the released table
  - Achieved by *k*-anonymity
- Attribute disclosure: learning something new about an individual or a group of individuals
  - i.e., the released data makes it possible to infer the characteristics of an individual more accurately than it would be possible without the data release
  - Can not be guaranteed by *k*-anonymity
- Identity disclosure leads to attribute disclosure
- On the other hand, attribute disclosure may occur with or without identity disclosure

Enhancements to k-Anonymity

Overview of Anonymization Methods



Differential Privacy

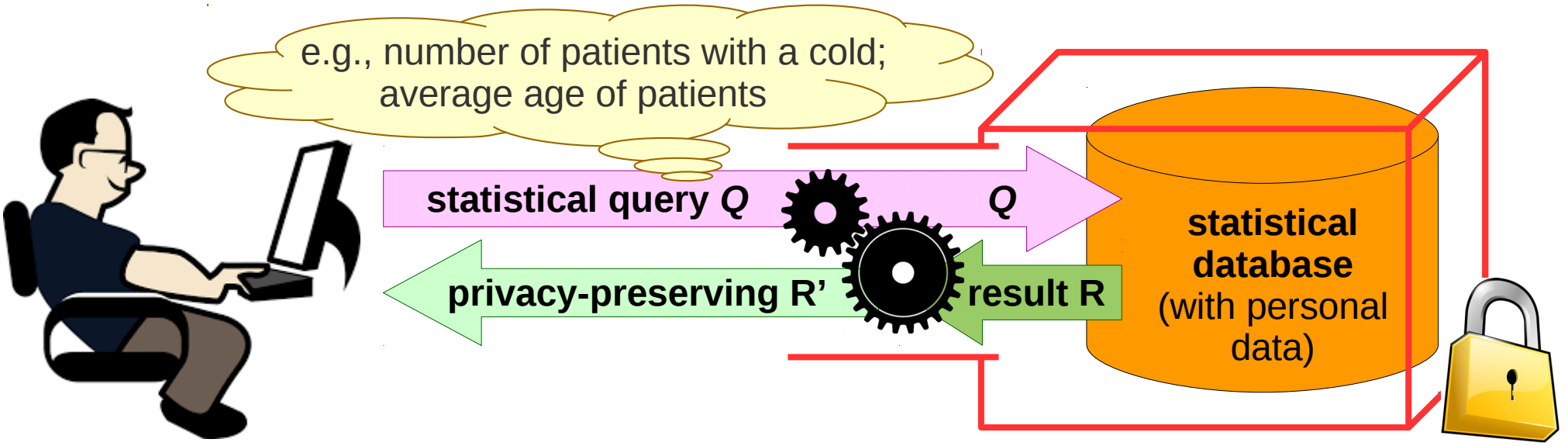
Cynthia Dwork: *Differential Privacy*. ICALP (2), 2006.

Cynthia Dwork: *A Firm Foundation for Private Data Analysis*. Communications of the ACM 54 (1), 2011.

General Idea

- Instead of releasing an anonymized database, offer a query mechanism that protects privacy (even if an adversary employs other databases)
  - Query mechanism may delete names, add noise, randomize the result, etc.

Cynthia Dwork



Picture sources: https://pixabay.com/en/coding-computer-computer-user-pc-1294361/
https://commons.wikimedia.org/wiki/File:Gears.png
http://cliparts101.com/free_clipart/17475/cadenas
https://www.seas.harvard.edu/directory/dwork

General Idea (cont'd)

- Returned result of a query is similar whether any single individual's record is included in the database or not
- A Example: Number of persons with a cold?

Name	Disease
Chris	Arthritis
David	Cold
Ethan	Heart problem
...	...

Name	Disease
Chris	Arthritis
Ethan	Heart problem
...	...

  - David is no worse off because his record is included in the returned query results
- Any two databases that differ by at most one tuple are called ***neighbors***

Definition

- A privacy-preserving query mechanism M_Q for query Q provides *ε-differential privacy* if
  - for every pair of neighboring databases D and D', and
  - for every possible output O of M,

  we have that:

$$\underbrace{\Pr[\, M_Q(D) = O \,]}_{\text{probability that the output of } M_Q \text{ over } D \text{ is } O} \leq e^{\varepsilon} \cdot \underbrace{\Pr[\, M_Q(D') = O \,]}_{\text{probability that the output of } M_Q \text{ over } D' \text{ is } O}$$

Remarks

- A privacy-preserving query mechanism M_Q for query Q provides *ε-differential privacy* if
  - for every pair of neighboring databases D and D', and
  - for every possible output O of M,

  we have that:

$$\Pr[\, M_Q(D) = O \,] \leq e^{\varepsilon} \cdot \Pr[\, M_Q(D') = O \,]$$

which is equivalent to

$$\frac{\Pr[\, M_Q(D) = O \,]}{\Pr[\, M_Q(D') = O \,]} \leq e^{\varepsilon} \approx 1 \pm \varepsilon$$

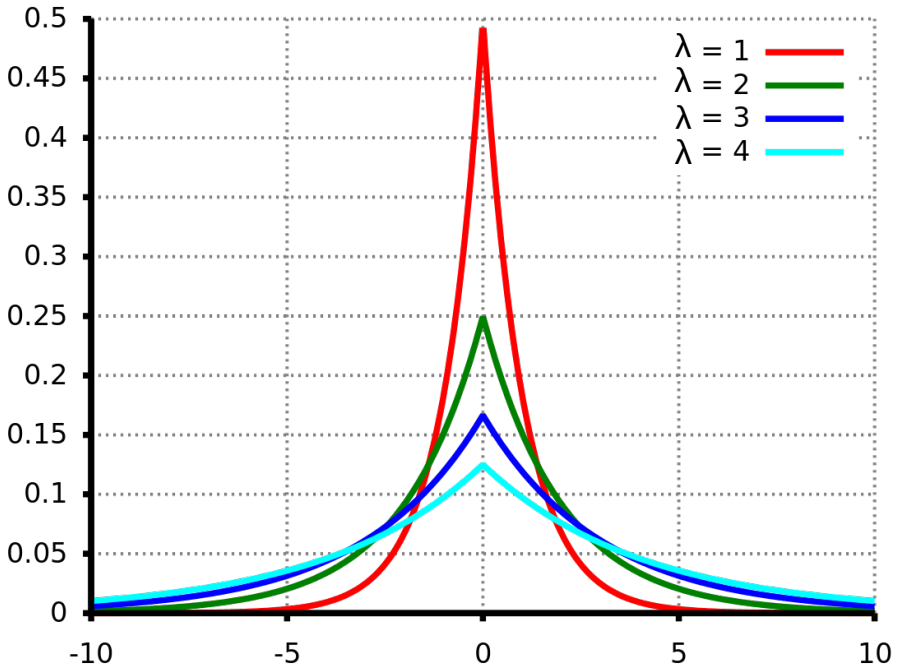
- The privacy parameter ε is usually small
  - e.g., if $\varepsilon = 0.1$, then $e^{\varepsilon} \approx 1.10$

How to add noise? Laplace-based Approach

- For queries Q whose results are real numbers
- $M_Q(D) = Q(D) + \eta$
  - where noise η is determined using the Laplace distribution with $\lambda = \Delta q / \varepsilon$

Laplace Distribution

- For queries Q whose results are real numbers
- $M_Q(D) = Q(D) + \eta$
  - where noise η is determined using the Laplace distribution with $\lambda = \Delta q / \varepsilon$



Picture sources: https://commons.wikimedia.org/wiki/File:Laplace-verteilung.svg

Query Sensitivity

- For queries Q whose results are real numbers
- $M_Q(D) = Q(D) + \eta$
  - where noise η is determined using the Laplace distribution with $\lambda = \Delta q / \varepsilon$

Definition: The *sensitivity* of a query Q is

$$\Delta q = \max | Q(D) - Q(D') |$$

for any two neighboring databases D and D'.

Examples:

- Δq for "count all tuples" is: 1
- Δq for "count all patients with a cold" is: 1
- Δq for "maximum age of all patients" is: max age

www.liu.se

LINKÖPING UNIVERSITY