Schutz der Privatsphäre (WS15/16)
Introduction to Privacy (Part 1)

"You have zero privacy. Get over it."

Scott McNealy, 1999

Privacy, k-anonymity, and differential privacy

Johann Christoph Freytag
Humboldt-Universität zu Berlin

Is it always obvious?

- Is it always obvious that privacy is violated or breached?
- Latanya Sweeney's Finding
  - 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**, diagnosis, procedure, ...)

date **o**f **b**irth

[Sween'02]

http://dataprivacylab.org/people/sweeney/

Schutz der Privatsphäre (WS15/16)
Introduction to Privacy (Part 1)

Latanya Sweeney's Finding (1)

- Sweeney paid $20 and bought the voter registration list for Cambridge, MA:

Voter						GIC		
Name	Adress	…	**ZIP**	**DOB**	**Sex**			
			ZIP	**DOB**	**Sex**	Diagnostic	Medication	…

- William Weld (former governor) 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!
- 87 % of population in U. S. can be identified by ZIP, dob, sex

What is Privacy?

- **Definition 1:** [Sweeney, 2002]
"**Privacy** reflects the ability of a person, organization, government, or entity to control its own space, where the concept of space (or "privacy space") takes on different contexts."
  - Physical space, against invasion
  - Bodily space, medical consent
  - Computer space, spam
  - Web browsing space, Internet privacy
- **Definition 2:** [Agrawal et al., 2002]
"**Privacy** is the right of individuals to determine for themselves when, how, and to what extent information about them is communicated to others."
(We shall call this data/information privacy)

Schutz der Privatsphäre (WS15/16)
Introduction to Privacy (Part 1)

Challenge

- **Given:** person-specific data
  - microdata table *T*
- **Goal:** privacy preserving public release table *T**
  - Information should remain practically useful

SSN	Name	Zipcode	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

→ attributes A_j

tuples *t*

Microdata *T*

Quasi-identifier

- **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	Zipcode	Age	Sex
Chris	12211	18	M
Jack	19221	20	M

Public database

Zipcode	Age	Sex	Disease
12211	18	M	Arthritis
12244	19	M	Cold
…	…	…	…

T

Linking attack

Name	Zipcode	Age	Sex	Disease
Chris	12211	18	M	Arthritis

QI_T = {Zipcode, Age, Sex}

Ω = {Chris, David, Jack, …}

Schutz der Privatsphäre (WS15/16)
Introduction to Privacy (Part 1)

Microdata

Identifier | Quasi-identifier | Sensitive attributes

SSN	Name	Zipcode	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

→ attributes A_j

tuples t

Microdata T

Introduced by Latanya Sweeney, 2002

K-Anonymity

Schutz der Privatsphäre (WS15/16)
Introduction to Privacy (Part 1)

k-anonymity
Definition

- **Definition (*k*-anonymity)**
  A table T satisfies k-anonymity if for every tuple $t \in T$ there exist $k-1$ other tuples $t_1, t_2, \ldots, t_{k-1} \in T$ such that $t[QI_T] = t_1[QI_T] = t_2[QI_T] = \cdots = t_{k-1}[QI_T]$ for all quasi-identifier QI_T.

Zipcode	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

Microdata table T

Zipcode	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-anomynous table T^*

k-anonymity

QI-group/ equivalence class

Zipcode	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

T^*

Public database

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

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

Disease of Chris?
Arthritis or Cold?

Schutz der Privatsphäre (WS15/16)
Introduction to Privacy (Part 1)

Privacy protection vs. information

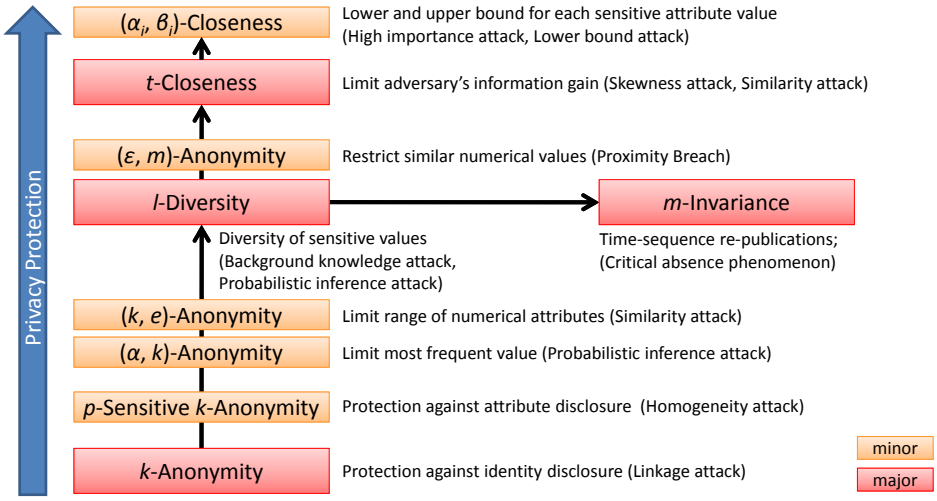
Zipcode	Age	Sex	Disease
122**	18–19	M	Arthritis
122**	18–19	M	Cold
*	27	*	Heart problem
*	27	*	Flu
*	27	*	Arthritis
12391	≥ 30	F	Diabetes
12391	≥ 30	F	Flu

2-anonymous table
high information content

Zipcode	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

Anonymization Methods
Overview

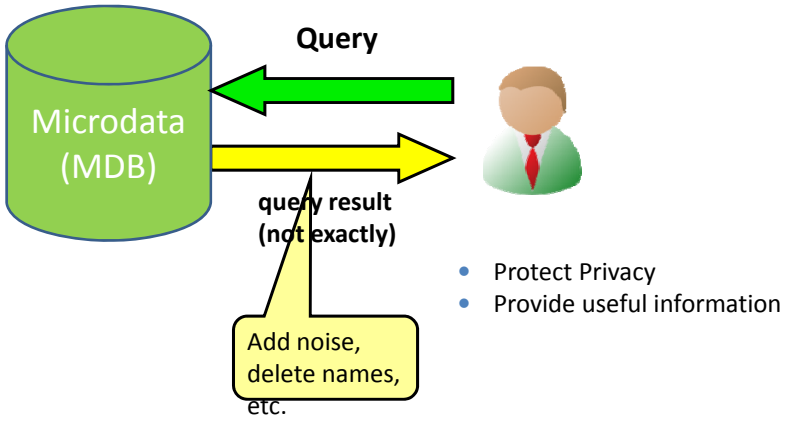


Schutz der Privatsphäre (WS15/16)
Introduction to Privacy (Part 1)

Introduced by Cynthia Dwork (2006)

Differential Privacy

Model



Schutz der Privatsphäre (WS15/16)
Introduction to Privacy (Part 1)

Differential Privacy (informal)

- Output of a query is similar whether any single individual's record is included in the database or not

Query: # of persons with a cold?

Database D

Name	Disease
Chris	Arthritis
David	Cold
Ethan	Heart problem

Query → R1 ≈ R2 ← Query

Database D'

Name	Disease
Chris	Arthritis
Ethan	Heart problem

- David is no worse off because his record is/is not included in the output of a query

Definitions

Definition 1 (neighboring databases):

Two databases D, D' are **neighbors** if they differ by at most one tuple

Definition 2 (ε-differential privacy):

A randomized algorithm G provides **ε-differential privacy** if:

- for all neighboring databases D and D', andprivacy
- for any outputs *O:*

$$\Pr[G(D) = O] \leq e^{\varepsilon} * \Pr[G(D') = O]$$

Schutz der Privatsphäre (WS15/16)
Introduction to Privacy (Part 1)

Differential Privacy – additional remarks

- $\Pr[G(D) = O] \leq e^{\varepsilon} * \Pr[G(D') = O]$

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

ε is a privacy parameter

- Epsilon is usually small: e.g. if $\varepsilon = 0.1$ then $e^{\varepsilon} \approx 1.10$

⇓ epsilon = ⇑ stronger privacy

Query sensitivity

Definition 3: The **sensitivity** of a query Q is

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

where D, D' are any two neighboring databases

Query Q	Sensitivity Δq
Q1: Count tuples	1
Q2: Count (patients with "Cold")	1
Q3: Count (patients with property X)	1
Q4: Max (age of patients)	max age

Schutz der Privatsphäre (WS15/16)
Introduction to Privacy (Part 1)

Differential privacy

[Dwork, ICALP06]

- How to add noise: ***Laplace distribution***

$$\Pr[\eta = x] = \frac{1}{2\lambda} e^{-|x-\mu|/\lambda}$$

- with
  - μ is the mean of the distribution (usually $\mu = 0$)
  - λ (referred to as the noise scale) is a parameter that controls the degree of privacy protection
  - $\lambda = \Delta q / \varepsilon$,
    i.e. sensitivity (of query) / strength of protection

Calibrate Noise & Sensitivity (1)

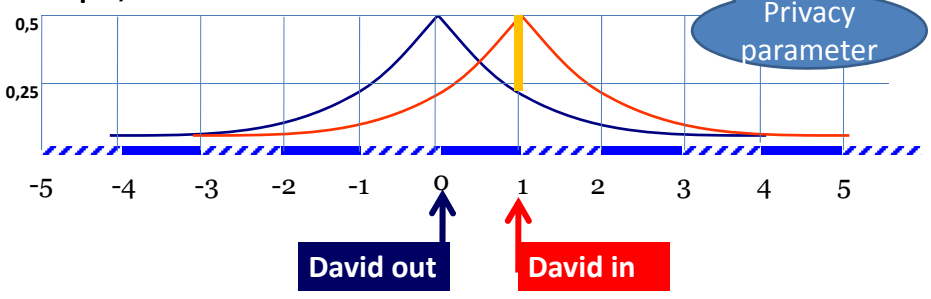
- Example 1:

Sensitivity

Q(D) + Laplace(Δq / ε)

Privacy parameter

Δq=1, ε=1.0



Schutz der Privatsphäre (WS15/16)
Introduction to Privacy (Part 1)

Challenges

- Semantic knowledge
  - Add chances for attacker (background knowledge)
    - Problem for k-anonymity, not for differential privacy
  - New protection necessary?
- … more??

Questions ??