Knowledge Discovery in Databases
SS 2016

Chapter 8: Privacy Preserving Data Mining

Lecture: Prof. Dr. Thomas Seidl

Tutorials: Julian Busch, Evgeniy Faerman,
Florian Richter, Klaus Schmid

- Introduction
  - Data Privacy
  - Privacy Preserving Data Mining
- k-Anonymity Privacy Paradigm
  - k-Anonymity
  - l-Diversity
  - t-Closeness
- Differential Privacy
  - Sensitivity, Noise Perturbation, Composition

Data Privacy

Huge volume of data is collected
from a variety of devices and platforms

Such as Smart Phones, Wearables,
Social Networks, Medical systems

Such data captures human behaviors,
routines, activities and affiliations

While this overwhelming data collection
provides an opportunity to perform data analytics

ACTIVITY

Data Abuse

Data Abuse is inevitable:
- It compromises individual’s privacy
- Or bridges the security of an institution

DATABASE SYSTEMS GROUP

Data Privacy: Attacks

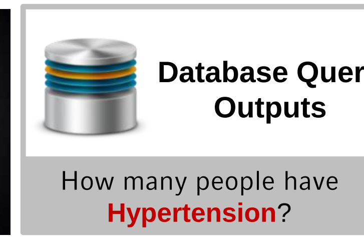
An attacker queries a database for sensitive records

Targeting of vulnerable or strategic nodes of large networks to
- Bridge an individual's privacy
- Spread virus

Adversary can track
- Sensitive locations and affiliations
- Private customer habits

These attacks pose a threat to privacy

Database Privacy



Network Privacy

Location & Customer Privacy

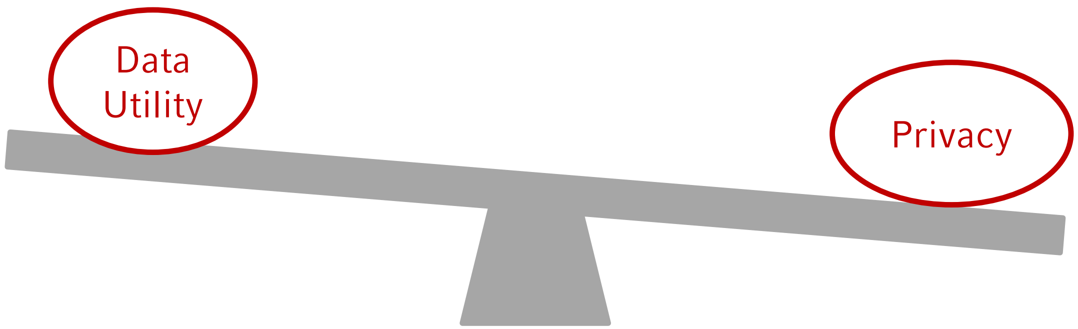
These privacy concerns need to be mitigated

They have prompted huge research interest to **Protect Data**

But,

- Strong Privacy Protection → Poor Data Utility
- Good Data Utility → Weak Privacy Protection

The challenge is to find a good trade-off between **Data Utility** and **Privacy**



Objectives of Privacy Preserving Data Mining in Database/Data Mining:

- Provide new plausible approaches to ensure data privacy when executing database and data mining operations
- Maintain a good trade-off between data utility and privacy

Privacy Breach

Linkage Attack: different public records can be linked to it to breach privacy

Alice has **Breast Cancer**

Hospital Records

Name	Gender	Age	Zip Code	Disease
Alice	F	29	52066	Breast Cancer
Jane	F	27	52064	Breast Cancer
Jones	M	21	52076	Lung Cancer
…				
….				
…				
Frank	M	35	52072	Heart Disease
Ben	M	33	52078	Fever
Betty	F	37	52080	Nose Pains

Public Records from Sport Club

Name	Gender	Age	Zip Code	Sports
Alice	F	29	52066	Tennis
Theo	M	41	52074	Golf
John	M	24	52062	Soccer
Betty	F	37	52080	Tennis
James	M	34	52066	Soccer

Betty had **Plastic Surgery**

k-Anonymity

A privacy paradigm for protecting database records before *Data Publication*

Three kinds of attributes:

- i) Key Attribute ii) Quasi-identifier ii) Sensitive Attribute

Key Attribute:

- Uniquely identifiable attributes (E.g., Name, Social Security Number, Telephone Number)

Quasi-identifier:

- Groups of attributes that can be combined with external data to uniquely re-identify an individual
- For Example: Date of Birth, Zip Code, Gender

Sensitive Attribute:

- Disease, Salary, Habit, Location etc.

k-Anonymity

- Example of partitioning a table into *Key, Quasi-Identifier* and *Sensitive* Attributes
- Hiding of **Key Attributes** does not guarantee privacy
- Quasi-Identifiers have to be altered to enforce privacy

Alice has **Breast Cancer**

Betty had **Plastic Surgery**

Released Hospital Records

Key Attribute	Quasi-Identifier			Sensitive Attribute
Name	**Gender**	**Age**	**Zip Code**	**Disease**
Alice	F	29	52066	Breast Cancer
Jane	F	27	52064	Breast Cancer
Jones	M	21	52076	Lung Cancer
Frank	M	35	52072	Heart Disease
Ben	M	33	52078	Fever
Betty	F	37	52080	Nose Pains

Public Records from Sport Club

Name	Gender	Age	Zip Code
Alice	F	29	52066
Theo	M	41	52074
John	M	24	52062
Betty	F	37	52080
James	M	34	52066

k-Anonymity

k-Anonymity ensures privacy by Suppression or Generalization of quasi-identifiers.

(k-ANONYMITY): Given a set of quasi-identifiers in a database table, the database table is said to be k-Anonymous, if the sequence of records in each quasi-identifier exists at least (k-1) times.

Suppression:

- Accomplished by replacing a part or the entire attribute value by "*"
- Suppress **Postal Code** : 52057 → 52***
- Suppress **Gender** : i) Male → * ii) Female → *

Generalization:

- **Exam**:

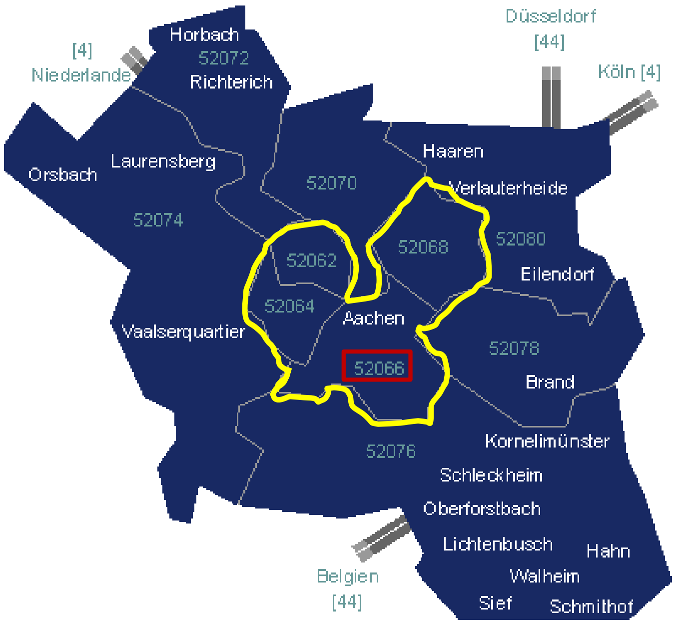
Not Available

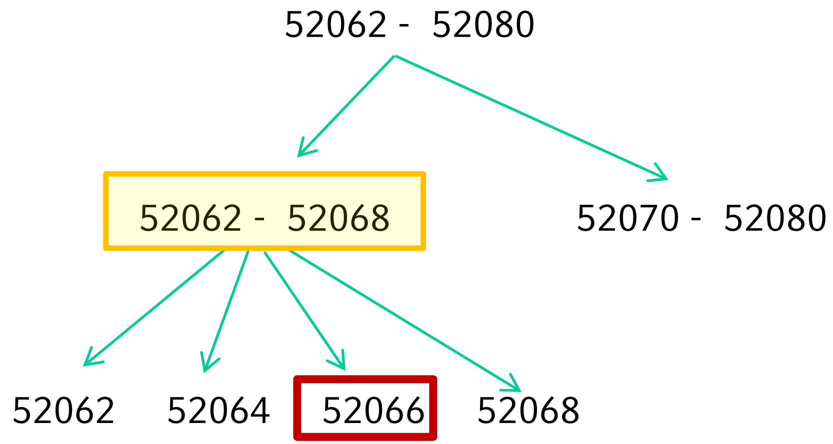
Passed Failed

{Excellent} {Very Good} {Good, Average} {Sick} {Poor } {Very Poor}

Generalization

Generalization of Postal Code:





Generalization can be achieved by (Spatial) Clustering

Example of *k*-Anonymity

Remove **Key Attributes**

Suppress or Generalize Quasi-Identifiers

Released Hospital Records

Key Attribute	Quasi-Identifier			Sensitive Attribute
Name	**Gender**	**Age**	**Zip Code**	**Disease**
	*	2*	520*	Breast Cancer
	*	2*	520*	Breast Cancer
Remove	*	2*	520*	Lung Cancer
	*	3*	520*	Heart Disease
	*	3*	520*	Fever
	*	3*	520*	Nose Pains

?

?

Public Records

Name	Gender	Age	Zip Code
Alice	F	29	52066
Theo	M	41	52074
John	M	24	52062
Betty	F	37	52080
James	M	34	52066

This database table is **3-Anonymous**

Oversuppression leads to stronger privacy but poorer Data Utility

Example of *k*-Anonymity

Generalize postal code to [5206*,5207*] and [5207*,5208*]

K-Anonymity is still satisfied with better Data Utility

Released Hospital Records

Quasi-Identifier			Sensitive Attribute
Gender	**Age**	**Zip Code**	**Disease**
*	2*	[5206*, 5207*]	Breast Cancer
*	2*	[5206*, 5207*]	Breast Cancer
*	2*	[5206*, 5207*]	Lung Cancer
*	3*	[5207*, 5208*]	Heart Disease
*	3*	[5207*, 5208*]	Fever
*	3*	[5207*, 5208*]	Nose Pains

?

?

Public Records

Name	Gender	Age	Zip Code
Alice	F	29	52066
Theo	M	41	52074
John	M	24	52062
Betty	F	37	52080
James	M	34	52066

Adversary cannot identify Alice or her disease from the released record

However, *k*-Anonymity still has several shortcomings

Shortcomings of *k*-Anonymity

Unsorted Attack: Different subsets of the record are released unsorted

Linkage Attack: Different versions of the released table can be linked to compromise *k*-Anonymity results.

Released Records 1

Quasi-Identifier			Sensitive Attribute
Gender	**Age**	**Zip Code**	**Disease**
*	2*	[5206*, 5207*]	Breast Cancer
*	2*	[5206*, 5207*]	Breast Cancer
*	2*	[5206*, 5207*]	Lung Cancer
*	3*	[5207*, 5208*]	Heart Disease
*	3*	[5207*, 5208*]	Fever
*	3*	[5207*, 5208*]	Nose Pains

Released Records 2

Quasi-Identifier			Sensitive Attribute
Gender	**Age**	**Zip Code**	**Disease**
F	2*	520*	Breast Cancer
F	2*	520*	Breast Cancer
M	2*	520*	Lung Cancer
M	3*	520*	Heart Disease
M	3*	520*	Fever
F	3*	520*	Nose Pains

Jones is at Row three. Jones has Lung Cancer!

Unsorted attack can be solved by *Randomizing* the order of the rows.

DATABASE SYSTEMS GROUP

Attack on *k*-Anonymity

LMU

Background Knowledge attack

Lack of diversity of the sensitive attribute values (homogeneity)

1. Background Knowledge

   Attacker's Knowledge: Alice is
   i) 29 years old ii) Female

   Attacker's Knowledge: Jones is
   i) 21 years old ii) Male

2. Homogeneity
   - All Females within 20 years have Breast Cancer. No diversity!!!
     → Alice has Breast Cancer!
   - All 2*-aged males have lung cancer
     → Jones has Lung Cancer!

Released Records

Quasi-Identifier			Sensitive Attribute
Gender	**Age**	**Zip Code**	**Disease**
F	2*	520*	Breast Cancer
F	2*	520*	Breast Cancer
M	2*	520*	Lung Cancer
M	2*	520*	Lung Cancer
M	3*	520*	Heart Disease
M	3*	520*	Fever
F	3*	520*	Nose Pains

This led to the creation of a new privacy model called ***l*-diversity**

l-Diversity

Addresses the homogeneity and background knowledge attacks

Accomplishes this by providing "well represented" sensitive attributes for each sequence of quasi-identifiers (Distinct *l*-Diversity)

Micro Data

Quasi-Identifier	Sensitive Attribute
…	Headache
…	Headache
…	Headache
…	Headache
…	Cancer

Anonymized 1

Quasi-Identifier	Sensitive Attribute
QI 1	Headache
QI 1	Headache
QI 1	Headache
QI 2	Cancer
QI 2	Cancer

Anonymized 2

Quasi-Identifier	Sensitive Attribute
QI 1	Headache
QI 3	Cancer
QI 2	Headache
QI 2	Headache
QI 4	Cancer

Diversity of Equivalent class

not "diverse"

QI 1: 50% "diverse"

Other variants of *l*-Diversity

- **Entropy l-Diversity**: For each equivalent class, the entropy of the distribution of its sensitive values must be at least $\log(l)$
- **Probabilistic l-Diversity**: The most frequent sensitive value of an equivalent class must be at most $1/l$

Limitations of *l*-Diversity

- Is not necessary at times
- Is difficult to achieve: For large record size, many equivalent classes will be needed to satisfy *l*-Diversity
- Does not consider the distribution of sensitive attributes

t-Closeness

The *l*-diversity approach is insufficient to prevent sensitive attribute disclosure

This led to the proposal of another privacy definition called ***t*-Closeness**

t-Closeness achieves privacy by keeping the distribution of each quasi-identifier's sensitive attribute "close" to their distribution in the database

For Example: Let P be the distribution of a sensitive attribute and Q denotes the distribution of all attributes in the database table

Given a threshold t:

an equivalent class satisfies *t*-closeness if the distance between P and Q is less than or equal to t

A table satisfies *t*-closeness if all its equivalent classes have *t*-closeness

Background Attack Assumptions

k-Anonymity, *l*-Diversity, *t*-Closeness make assumptions about the adversary

They at times fall short of their goal to prevent data disclosure

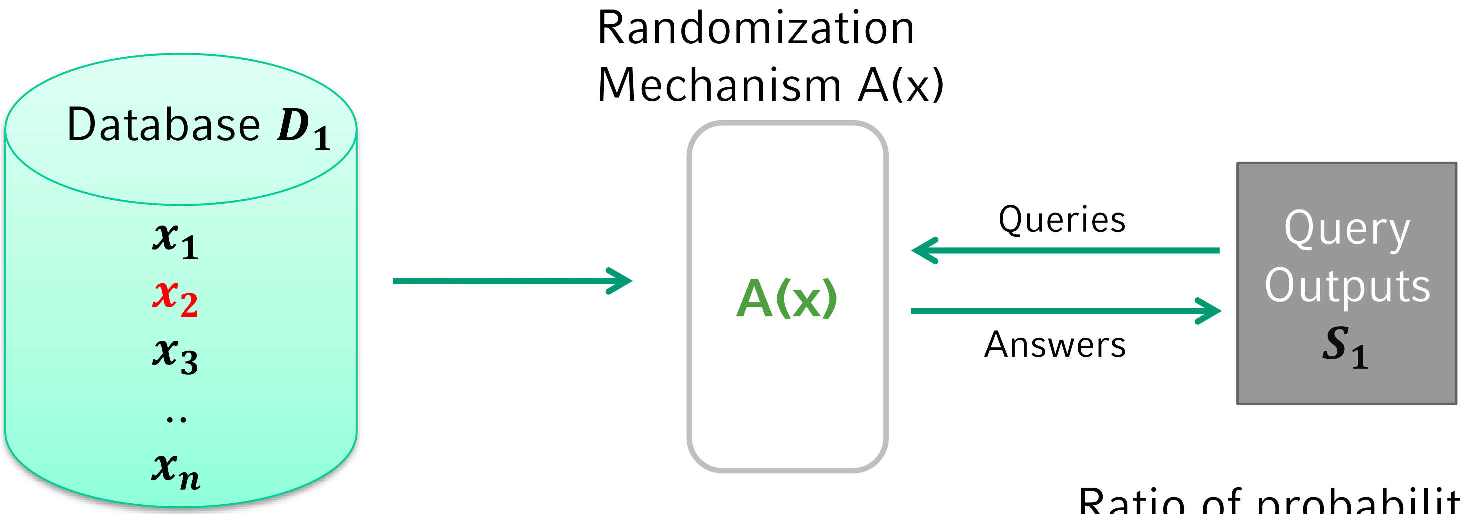
There is another privacy paradigm which does not rely on background knowledge

It is called Differential Privacy

Differential Privacy

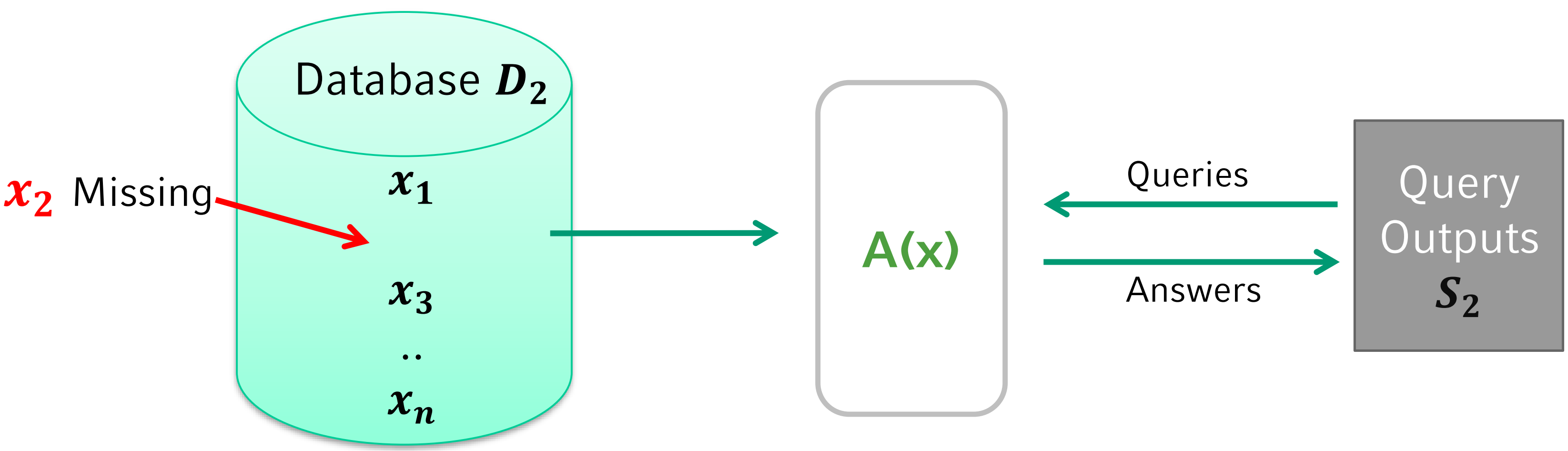
- Privacy through data perturbation
- Addition of a small amount of noise to the true data
- True value of a data can be masked from adversaries
- Used for the perturbation of query results of count, sum, mean functions, as well as other statistical query functions.

Differential Privacy



Row x_2 is removed. Meaning databases D_1 and D_2 differ by only 1 entry

Ratio of probabilities of s_1 and s_2 is at most ε



Core Idea:

- The addition or removal of one record from a database does not reveal any information to an adversary
- This means your presence or absence in the database does not reveal or leak any information from the database
- This achieves a strong sense of privacy

ε-DIFFERENTIAL PRIVACY:

A randomized mechanism $\boldsymbol{A}(x)$ provides ε-differential privacy if for any two databases D_1 and D_2 that differ on at most one element, and all output S Range($\boldsymbol{A}$),

$$\frac{\Pr[\,\boldsymbol{A}(D_1)\ \in S]}{\Pr[\,\boldsymbol{A}(D_2)\ \in\ S]} \leq \exp(\epsilon)$$

ε is the privacy parameter called privacy budget or privacy level

Sensitivity of a Function

Sensitivity is important for noise derivation

The sensitivity of a function is defined as the maximum change that occurs if one record is added or removed from a database D_1 to form another database D_2.

$$\| f(D_2) - f(D_1) \| \quad \leq \quad S(f)$$

Types of Sensitivities

- i) Global Sensitivity ii) Local Sensitivity

(LOCAL SENSITIVITY): *Local Sensitivity of a function* $f : D^n \to \mathbb{R}^d$ *for all* x *and* x' *which differ in one entry is* $LS_f(x) = \max_{d(x,x')=1} \|f(x) - f(x')\|_1$.

(GLOBAL SENSITIVITY): *Global Sensitivity of a function* $f : D^n \to \mathbb{R}^d$ *is given by* $GS_f = \max_x LS_f(x)$.

Data Perturbation

Data Perturbation in Differential Privacy is achieved by noise addition

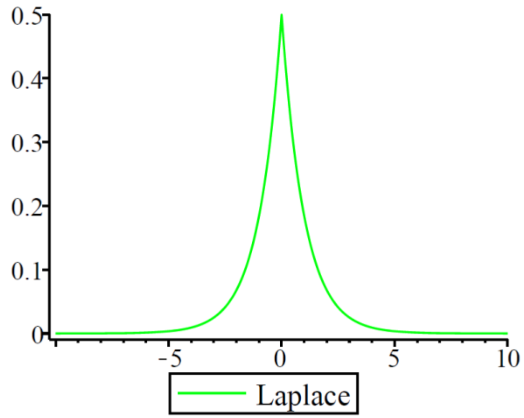
Different kinds of noise

- Laplace noise
- Gaussian noise
- Exponential Mechanism

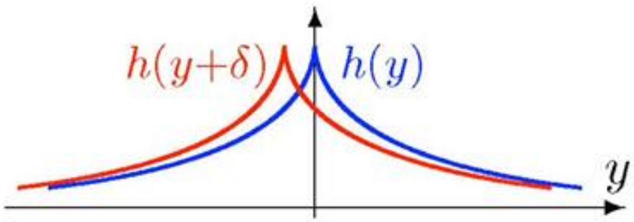
Laplace Noise

Stems from the Laplace Distribution

$$Lap(x) = \frac{1}{2b}\exp\left(\frac{-(x-\mu)}{b}\right)$$



$Lap(\lambda)$ consists of a density $Lap(\lambda) \propto \exp\left(\frac{\|y\|_1}{\lambda}\right)$



Output query is *ε-indistinguishable* when sensitivity $\frac{GS_f}{\epsilon}$ and noise of $Lap\left(\frac{GS_f}{\epsilon}\right)$ stronger is used for perturbation

Theorem *For a given function $f : D^n \to \mathbb{R}^d$, which has sensitivity S(f) , a mechanism $A(x) = f(x) + Lap(\frac{S(f)}{\epsilon})^d$ provides ϵ-differential privacy.*

Exponential Mechanism

- Extension the notion of differential privacy to incorporate non-real value functions
  - Example: Color of a car, category of a car
- Guarantees privacy by approximating the true value of a data using quality function or utility function.
- Exponential Mechanism requires: 1) Input dataset 2) Output range 3) Utility function
- It maps several input data to some outputs
- The output whose mapping has the best score is chosen and sampled with a given probability such that differential privacy is guaranteed.

Theorem *For a given input $\mathcal{X}$ and a function $u : (\mathcal{X} \times y) \to \mathbb{R}$, an algorithm that chooses an output y with a probability $\propto \exp(-\epsilon\frac{u(\mathcal{X},y)}{2\Delta u})$ is ϵ-differential private.*

Composition

There are two types of composition

- Sequential Composition
- Parallel Composition

Sequential Composition:

- Exhibited when a sequence of computation provides differential privacy in isolation.
- The final privacy guarantee is said to be the sum of each ε-differential privacy.

Parallel Composition:

- Occurs when the input data is partitioned in disjoint sets, independent of the original data
- The final privacy from such a sequence of computation depends on the worst computation guarantee of the sequence

Summary

- Privacy Preserving Data Mining
- k-Anonymity Privacy Paradigm
  - k-Anonymity
  - l-diversity
  - t-Closeness
- Differential Privacy
  - Sensitivity
  - Noise Perturbation
  - Composition