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

Data Abuse is inevitable:
- It compromises individual’s privacy
- Or bridges the security of an institution
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
These privacy concerns need to be mitigated. They have prompted huge research interest to **Protect Data**.

But, there is a trade-off between **Data Utility** and **Privacy**.

- **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

**Hospital Records**

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>Age</th>
<th>Zip Code</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>F</td>
<td>29</td>
<td>52066</td>
<td>Breast Cancer</td>
</tr>
<tr>
<td>Jane</td>
<td>F</td>
<td>27</td>
<td>52064</td>
<td>Breast Cancer</td>
</tr>
<tr>
<td>Jones</td>
<td>M</td>
<td>21</td>
<td>52076</td>
<td>Lung Cancer</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Frank</td>
<td>M</td>
<td>35</td>
<td>52072</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>Ben</td>
<td>M</td>
<td>33</td>
<td>52078</td>
<td>Fever</td>
</tr>
<tr>
<td>Betty</td>
<td>F</td>
<td>37</td>
<td>52080</td>
<td>Nose Pains</td>
</tr>
</tbody>
</table>

**Public Records from Sport Club**

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>Age</th>
<th>Zip Code</th>
<th>Sports</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>F</td>
<td>29</td>
<td>52066</td>
<td>Tennis</td>
</tr>
<tr>
<td>Theo</td>
<td>M</td>
<td>41</td>
<td>52074</td>
<td>Golf</td>
</tr>
<tr>
<td>John</td>
<td>M</td>
<td>24</td>
<td>52062</td>
<td>Soccer</td>
</tr>
<tr>
<td>Betty</td>
<td>F</td>
<td>37</td>
<td>52080</td>
<td>Tennis</td>
</tr>
<tr>
<td>James</td>
<td>M</td>
<td>34</td>
<td>52066</td>
<td>Soccer</td>
</tr>
</tbody>
</table>

Alice has Breast Cancer

Betty had Plastic Surgery
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.
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

### Released Hospital Records

<table>
<thead>
<tr>
<th>Key Attribute</th>
<th>Quasi-Identifier</th>
<th>Sensitive Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Gender</td>
<td>Age</td>
</tr>
<tr>
<td>Alice</td>
<td>F</td>
<td>29</td>
</tr>
<tr>
<td>Jane</td>
<td>F</td>
<td>27</td>
</tr>
<tr>
<td>Jones</td>
<td>M</td>
<td>21</td>
</tr>
<tr>
<td>Frank</td>
<td>M</td>
<td>35</td>
</tr>
<tr>
<td>Ben</td>
<td>M</td>
<td>33</td>
</tr>
<tr>
<td>Betty</td>
<td>F</td>
<td>37</td>
</tr>
</tbody>
</table>

### Public Records from Sport Club

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>Age</th>
<th>Zip Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>F</td>
<td>29</td>
<td>52066</td>
</tr>
<tr>
<td>Theo</td>
<td>M</td>
<td>41</td>
<td>52074</td>
</tr>
<tr>
<td>John</td>
<td>M</td>
<td>24</td>
<td>52062</td>
</tr>
<tr>
<td>Betty</td>
<td>F</td>
<td>37</td>
<td>52080</td>
</tr>
<tr>
<td>James</td>
<td>M</td>
<td>34</td>
<td>52066</td>
</tr>
</tbody>
</table>
**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:**
  - Passed
  - {Excellent}  {Very Good}  {Good, Average}
  - {Sick}  {Poor}  {Very Poor}
  - Not Available
  - Failed
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

<table>
<thead>
<tr>
<th>Key Attribute</th>
<th>Quasi-Identifier</th>
<th>Sensitive Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Gender</td>
<td>Age</td>
</tr>
<tr>
<td>*</td>
<td>2*</td>
<td>520*</td>
</tr>
<tr>
<td>*</td>
<td>2*</td>
<td>520*</td>
</tr>
<tr>
<td>*</td>
<td>2*</td>
<td>520*</td>
</tr>
<tr>
<td>*</td>
<td>3*</td>
<td>520*</td>
</tr>
<tr>
<td>*</td>
<td>3*</td>
<td>520*</td>
</tr>
<tr>
<td>*</td>
<td>3*</td>
<td>520*</td>
</tr>
</tbody>
</table>

### Public Records

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>Age</th>
<th>Zip Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>F</td>
<td>29</td>
<td>52066</td>
</tr>
<tr>
<td>Theo</td>
<td>M</td>
<td>41</td>
<td>52074</td>
</tr>
<tr>
<td>John</td>
<td>M</td>
<td>24</td>
<td>52062</td>
</tr>
<tr>
<td>Betty</td>
<td>F</td>
<td>37</td>
<td>52080</td>
</tr>
<tr>
<td>James</td>
<td>M</td>
<td>34</td>
<td>52066</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age</th>
<th>Zip Code</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>2*</td>
<td>[5206*, 5207*]</td>
<td>Breast Cancer</td>
</tr>
<tr>
<td>*</td>
<td>2*</td>
<td>[5206*, 5207*]</td>
<td>Breast Cancer</td>
</tr>
<tr>
<td>*</td>
<td>2*</td>
<td>[5206*, 5207*]</td>
<td>Lung Cancer</td>
</tr>
<tr>
<td>*</td>
<td>3*</td>
<td>[5207*, 5208*]</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>*</td>
<td>3*</td>
<td>[5207*, 5208*]</td>
<td>Fever</td>
</tr>
<tr>
<td>*</td>
<td>3*</td>
<td>[5207*, 5208*]</td>
<td>Nose Pains</td>
</tr>
</tbody>
</table>

### Public Records

<table>
<thead>
<tr>
<th>Name</th>
<th>Gender</th>
<th>Age</th>
<th>Zip Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>F</td>
<td>29</td>
<td>52066</td>
</tr>
<tr>
<td>Theo</td>
<td>M</td>
<td>41</td>
<td>52074</td>
</tr>
<tr>
<td>John</td>
<td>M</td>
<td>24</td>
<td>52062</td>
</tr>
<tr>
<td>Betty</td>
<td>F</td>
<td>37</td>
<td>52080</td>
</tr>
<tr>
<td>James</td>
<td>M</td>
<td>34</td>
<td>52066</td>
</tr>
</tbody>
</table>

Adversary cannot identify Alice or her disease from the released record

However, $k$-Anonymity still has several shortcomings
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.

<table>
<thead>
<tr>
<th>Released Records 1</th>
<th>Released Records 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quasi-Identifier</strong></td>
<td><strong>Sensitive Attribute</strong></td>
</tr>
<tr>
<td>Gender</td>
<td>Age</td>
</tr>
<tr>
<td>*</td>
<td>2*</td>
</tr>
<tr>
<td>*</td>
<td>2*</td>
</tr>
<tr>
<td>*</td>
<td>2*</td>
</tr>
<tr>
<td>*</td>
<td>3*</td>
</tr>
<tr>
<td>*</td>
<td>3*</td>
</tr>
<tr>
<td>*</td>
<td>3*</td>
</tr>
</tbody>
</table>

Jones is at Row three. Jones has Lung Cancer!

Unsorted attack can be solved by Randomizing the order of the rows.
Background Knowledge attack
Lack of diversity of the sensitive attribute values (homogeneity)

1. Background Knowledge

   - 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!

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!

This led to the creation of a new privacy model called \( l \)-diversity
I-Diversity

Addresses the homogeneity and background knowledge attacks
Accomplishes this by providing “well represented” sensitive attributes for each sequence of quasi-identifiers (Distinct I-Diversity)

<table>
<thead>
<tr>
<th>Micro Data</th>
<th>Anonymized 1</th>
<th>Anonymized 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quasi-Identifier</strong></td>
<td><strong>Sensitive Attribute</strong></td>
<td><strong>Quasi-Identifier</strong></td>
</tr>
<tr>
<td>...</td>
<td>Headache</td>
<td>QI 1</td>
</tr>
<tr>
<td>...</td>
<td>Headache</td>
<td>QI 1</td>
</tr>
<tr>
<td>...</td>
<td>Headache</td>
<td>QI 1</td>
</tr>
<tr>
<td>...</td>
<td>Headache</td>
<td>QI 2</td>
</tr>
<tr>
<td>...</td>
<td>Cancer</td>
<td>QI 2</td>
</tr>
<tr>
<td>...</td>
<td>Cancer</td>
<td>QI 2</td>
</tr>
<tr>
<td>...</td>
<td>Cancer</td>
<td>QI 4</td>
</tr>
</tbody>
</table>

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
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

Database $D_1$

$x_1$
$x_2$
$x_3$
.. 
$x_n$

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

Database $D_2$

$x_1$
$x_3$
.. 
$x_n$

$x_2$ Missing

Randomization Mechanism $A(x)$

Query Outputs $S_1$

Queries

Answers

Query Outputs $S_2$

Queries

Answers

Ratio of probabilities of $s_1$ and $s_2$ is at most $\varepsilon$
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 $A(x)$ provides ε-differential privacy if for any two databases $D_1$ and $D_2$ that differ on at most one element, and all output $S \text{ Range}(A)$,

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

ε is the privacy parameter called privacy budget or privacy level
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 $D_2$.

\[ \| f(D_2) - f(D_1) \| \leq S(f) \]

Types of Sensitivities

- i) Global Sensitivity
- ii) Local Sensitivity

(LOCAL SENSITIVITY): Local Sensitivity of a function $f : D^n \rightarrow \mathbb{R}^d$ 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 $f : D^n \rightarrow \mathbb{R}^d$ 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

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

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

Output query is \(\varepsilon\)-indistinguishable when sensitivity \(\frac{GSf}{\varepsilon}\) and noise of \(\text{Lap} \left( \frac{GSf}{\varepsilon} \right)\) stronger is used for perturbation

**Theorem**  For a given function \(f : D^n \to \mathbb{R}^d\), which has sensitivity \(S(f)\), a mechanism \(A(x) = f(x) + \text{Lap} \left( \frac{S(f)}{\varepsilon} \right)^d\) provides \(\varepsilon\)-differential privacy.
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 $u : (\mathcal{X} \times y) \rightarrow \mathbb{R}$, an algorithm that chooses an output $y$ with a probability $\propto \exp(-\frac{\epsilon u(x, y)}{2\Delta u})$ is $\epsilon$-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 \( \epsilon \)-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