CHAPTER 4

PRIVACY PRESERVING DATA MINING

University of Kentucky
February 2011

Based partly on "Privacy Preserving Data Mining: Challenges & Opportunities" by Ramakrishnan Srikant from Google, Inc,
OVERVIEW

- Basic Concepts
  - What is privacy-preserving data mining
  - Difference between privacy-preserving data mining and data mining
  - Why do we need privacy-preserving data mining

- Technologies for Privacy-Preserving Data Mining
  - Statistical disclosure control
  - Randomization
  - Cryptography

- Privacy Attacks

- Challenges
WHAT IS PRIVACY-PRESERVING DATA MINING

- Privacy-preserving data mining (PPDM) is to conduct data mining operations under the condition of preserving data privacy.
- PPDM can be considered in two aspects:
  - Protecting sensitive data values, e.g., names, social security numbers, etc., of some people.
  - Protecting confidential knowledge in data, e.g., hiding confidential knowledge and not affecting the non-confidential knowledge and data utilities.

Ideal model for privacy-preserving data publishing:

- Data cleaning
- D(R+R') → D' (R')

D: original data
D': published data
R': non-sensitive rules
R: sensitive rules
DIFFERENCE BETWEEN PPDM AND DM

- During data collection, process data by removing private information or adding noise, this is privacy protection for individual data entries.

- In pre-processing, process data for data mining purposes, e.g., reconstructing original data distributions or statistical properties.

- Modifying data mining algorithms so that data mining can be performed without disclosing private information or reducing the information disclosure, e.g., secure multiparty computation (SMC). In most cases, data mining algorithms have to be modified.
WHY DO WE NEED PPDM

- Various data are collected at an increasing rate
- Data mining is threatening the security of sensitive data

- For example, disease control centers need to collect patient information from various hospitals and clinics, for disease prevention and control. During this process, sensitive information such as patient’s diseases may be disclosed. But the data owner does not want this information to be disclosed to other people or organization
WHY DO WE NEED PPDM

- Several commercial partners need cooperation in data analysis, but they do not want to share customer information, and have to prevent others from knowing their business secrets.
- Cooperation and competition are business strategies.

- After privacy-preserving data processing, data can be published.
- Some data knowledge can be hidden.

- Obtaining benefits and protecting itself, is one goal of PPDM.
- Obtaining benefits and protecting customers, is another goal of PPDM.
PRESERVING PRIVACY AND DATA MINING ARE CONTRADICTION

- If we emphasize data privacy, we may compromise the benefits of data mining.
- If we focus on knowledge discovery, we may not guarantee the protection of confidential data. However, data with privacy are vitally important in many data analysis applications.
OVERVIEW

- Basic Concepts
  - What is privacy-preserving data mining
  - Difference between PPDM and data mining
  - Why do we need privacy-preserving data mining

- Technologies for Privacy-preserving Data Mining
  - Statistical disclosure control
  - Randomization
  - Cryptography

- Privacy Attacks

- Challenges
TECHNIQUES FOR PPDM

- According to data distribution, technologies for PPDM can be in two categories:
  - Distributed, data to be mined are in different locations
  - Centralized, all data needed are in one location

- According to data processing, technologies for PPDM can be in three categories:
  - Statistical disclosure control
  - Randomization
  - Cryptography
DISTRIBUTED DATA

- Data are owned by two or more owners
- None of them trust the others or a third party

First Class of Methods:
- Every data owners process their own data with PPDM
- They may use different PPDM methods
- Are the combined data still useful?

Second Class of Methods:
- No need for data combination
- Directly use distributed data structure (e.g., SMC)
FLOW CHART FOR DISTRIBUTED PPDM

Dataset 1

Data records without sensitive data values

Single data entry privacy protection

Dataset 2

Data records without sensitive data values

Privacy-preserving data mining on distributed dataset

Knowledge
CENTRALIZED DATA

- Data is owned by one or more owners
- Data mining is performed by a third party
- Perturb, distort, or modify data values
- Hide certain data patterns

- If parts of the data are from different owners, do they have to use the same data perturbation methods?
- Can the third party keep the data after data mining and analysis?
FLOW CHART FOR CENTRALIZED PPDM

- Dataset
- Data records without sensitive data values
- Perturbed dataset
- Perturbation
- Reconstruction
- Data mining
- Knowledge
- Single data entry privacy protection
- Data patterns
STATISTICAL DISCLOSURE CONTROL

- Statistical Disclosure Control (SDC) is techniques to protect statistical data. It allows the data to be published and analyzed by the public, but protects private information of certain individuals or groups.

- SDC uses special methods to modify data. The aim is to protect the data privacy and minimize the information loss.

- Some SDC Methods: Tabular data protection, Dynamic database, Microdata protection.
**RANDOMIZATION**

- Randomization is an important method to provide privacy protection for centralized data. In some sense, it is a statistical disclosure control method.

- The basic ideas are:
  - Add noise to the data, but maintain its original probability distribution.
  - Guarantee that individual records are difficult to recover, which gives the meaning of privacy-preserving.
RANDOMIZATION METHOD – AN EXAMPLE

- Volvo S40 website targets people in their 20s
  - Are visitors in their 20s or 40s?
  - Which demographic groups like/dislike the website?
RANDOMIZATION APPROACH OVERVIEW

30 | 70K |

Randomizer

65 | 20K |

Randomizer

50 | 40K |

Randomizer

25 | 60K |

Randomizer

Reconstruct distribution of Age

Reconstruct distribution of Salary

Data Mining Algorithms

Model

...
RECONSTRUCTION PROBLEM

- Original values $x_1, x_2, ..., x_n$
  - from probability distribution $X$ (unknown)
- To hide these values, we use $y_1, y_2, ..., y_n$
  - from probability distribution $Y$
- Given
  - $x_1 + y_1, x_2 + y_2, ..., x_n + y_n$
  - the probability distribution of $Y$
- Estimate the probability distribution of $X$
RECONSTRUCT SINGLE POINT

- Use Bayes' rule for density functions

---

**Age**

- Original distribution for age
- Probability estimate of the original value of V
RECONSTRUCT DISTRIBUTION

- Combine estimates of where point came from for all the points:
  Give estimate of the original distribution

![Graph showing combined and single point probability estimates](image)
**RECONSTRUCTION (BOOTSTRAPPING)**

- $f_x^0 = \text{uniform distribution}$
- $j = 0$ //iteration
- repeat
  $$f_{x}^{j+1}(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{f_Y((x_i + y_i) - a) f_{X}^{j}(a)}{\int_{-\infty}^{\infty} f_Y((x_i + y_i) - a) f_{X}^{j}(a)}$$ (Bayes’s rules)
  
  $$j = j + 1$$
  until (stopping criterion met)

- Converges to maximum likelihood estimate.
RESULTS

The graph shows the number of visitors by age. The highest number of visitors is around the ages of 20 and 60, with a peak number of 1200. The number of visitors decreases as the age increases and decreases further as the age increases beyond 60. The graph also shows a trend where the number of visitors is lower in the middle age range (30-50) compared to the young and old age ranges.
RANDOMIZATION – OTHER METHODS

- In addition to adding noise (addition methods) to the data, multiplication methods can also be used.
- There are three classes of multiplication methods: rotation, projection, and geometric perturbation.

Summary of randomization methods:
- Easy to implement.
- Independent of datasets, allowing randomization at the data collection phase.
- Do not consider the distribution of original data, cannot guarantee the randomized data not to be reconstructed.
- For densely populated data segments, randomized data may be easier to be attacked than the original one.
CRYPTOGRAPHY

- SDC and randomization are for centralized data, cryptography is to provide privacy protection for distributed data.

- The most important issue in privacy protection in a distributed data computation is communications, encryption meets this requirements.

- In privacy-preserving data mining on distributed data, there are two categories of data models: horizontally partitioned datasets, vertically partitioned datasets.
DATA MINING IN MULTI-NATIONAL CORPORATION

Problem: Two departments of the corporation have their own confidential data, they want to combine the data to construct a decision classifier, but they also do not want the other department to know unnecessary information.

- Horizontally partitioned data
  - Partition is based on different customers

- Vertically partitioned data
  - Partition is based on different attributes
**Privacy Attacks**

- **Malicious attacks**: Attackers can do anything to obtain private information of the other parties, e.g., do not honor agreements, send false information, involve in collaborations with other attackers.

- **Semi-honest attacks**: Attackers follow appropriate computation protocols, but want to gain others’ private information.
SECURE MULTIPARTY COMPUTATION

- In a distributed environment, PPDM needs the participations of several parties, they do not want to disclose anything but the data mining results.

- For this purpose, we need Secure Multiparty Computation, or SMC.

- What is SMC:
  - A group of participants want to compute the value of a given function. Every participant provides an input. For security purpose, the input cannot be disclosed to other participants. It also requires that correct results be computed even if there are some participants who are semi-honest.
SECURE MULTIPARTY COMPUTATION AND TRUSTED THIRD PARTY

- In general, every data owners give their original data to a trusted third party for computation.

- The trusted third party provides results. After the data analysis, the trusted third party destroys the data.

- In SMC, privacy protection is modeled as no trusted third party, i.e., every node only knows its own inputs and the final computed results of all data.
SMC – MAIN ALGORITHMS

- Based on cryptography theory, assume semi-honest attacker model
  - Secure Sum
  - Secure Dot Product
  - Secure Polynomial Evaluation
  - Secure Logarithm
  - Secure Set Union
  - Secure Set Intersection
SUMMARY

- SMC solves different problems (vs. randomization methods)
  - It is efficient for semi-honest attackers and not too many such attackers
  - It gives the same results as would be obtained using non-secure methods
  - It cannot be generalized to data of single users
OVERVIEW

- **Basic Concepts**
  - What is privacy-preserving data mining
  - Difference between PPDM and data mining
  - Why do we need privacy-preserving data mining

- **PPDM Methods**
  - Statistical disclosure control
  - Randomization
  - Cryptography

- **Privacy Attacks**

- **Challenges**
PRIVACY ATTACKS

- Use the perturbed data and any prior knowledge to estimate the original data
- Attacks with respect to noise addition methods
- Attacks with respect to matrix multiplication perturbation methods
ATTACK TECHNIQUES WITH RESPECT TO NOISE ADDITION METHODS

- In noise addition methods, data owner adds a noise matrix to the original data matrix to obtain a perturbed matrix Y and publishes the data matrix Y

\[ Y = X + R \]

- If the probability density function of \( R \) is known, and the attacker knows the perturbed data records and these records comes from independent samples of the stochastic vectors of \( Y = X + R \)
ATTACK TECHNIQUES FOR NOISE ADDITION METHODS (CONT.)

- Three attack techniques for noise addition methods
  - Analyzing eigenvalues of the data matrix to filter out the noise, typical methods include Spectral Filtering, Singular Value Decomposition Filtering, and Principal Component Analysis Filtering
  - Using Bayes’ method, typical method is to Maximize a Posteriori Probability Estimation
  - Based on the assumption: If the probability density function of $X$ can be reconstructed, in some cases, this can cause the disclosure of the original data information. Such attacks are called Distribution Analysis
ATTACK TECHNIQUES WITH RESPECT TO MATRIX MULTIPLICATION

- Data owner uses $Y = MX$ to replace the original data $X$, where $M$ is a special $n' \times n$ matrix with some properties.

- If $M$ is orthogonal, the perturbation will maintain Euclidean distance. If $x_1$ and $x_2$ are two columns in $X$, and the corresponding columns in $Y$ are $y_1$ and $y_2$, then we have $\|x_1 - x_2\| = \|y_1 - y_2\|$

- Since matrix multiplication can maintain Euclidean distance, many data mining algorithms can be applied on the perturbed data and result in the same results as on the original data. A typical algorithm is $K$-Means clustering.
Privacy attacks with respect to matrix multiplication (cont.)

- If $M$ is not known to the attackers, without any prior knowledge, it will be difficult for the attackers to reconstruct the original data $X$

- However, in many situations, attackers may have prior knowledge or background information, which facilitates the attacks

- Attacking methods are mainly based on two classes of prior knowledge
  - Known input-output: Attackers know a few original records and know the corresponding perturbed records in $Y$
  - Known samples: Attackers collect a few independent samples in $X$
SUMMARY

- There are a few situations that may leak private information from the perturbed data:
  - Re-Identification: In real world, many data have strong associations, which could be used to filter additive noise
  - Known Samples: Attackers have some background information, such as the probability density function, or some partially overlapped or unoverlapped independent samples
  - Known Input-out: Sometimes, attackers know some private data records and their corresponding perturbed values. These data records can be used to estimate the original values of other data records
SUMMARY (CONT.)

- Data Mining Results: Data patterns obtained from data mining can also be used by the attackers to accurately guess the original data records.

- Sample Dependence: For some types of data, such as time series data, there exist automatic correlations/dependence between samples. These dependence relationships can be used by attackers to estimate the original data.
CHALLENGES IN PPDM

- Privacy-sensitive security profiling
- Potential privacy breaches
- Or the basics: What is privacy? How to measure privacy?
PRIVACY-SENSITIVE SECURITY PROFILING

- Heterogeneous, distributed data.
- New domains: text, graph

"Frequent Traveler" Rating Model

- Email
- Phone
- Demographic
- Criminal Records
- Credit Agencies
- Birth
- Marriage
- State
- Local
POTENTIAL PRIVACY BREACHES

○ Distribution is a spike.
  • Example: Everyone is of age 40

○ Some randomized values are only possible from a given range.
  • Example: Add U[-50,+50] to age and get 125, if the true age is 75. This can be easily figured out as untrue.
  • Not an issue with Gaussian.
POTENTIAL PRIVACY BREACHES (CONT.)

- Most randomized values in a given interval come from a given interval.
  - **Example:** 60% of the people whose randomized value is in [120,130] have their true age in [70,80].
  - Implication: Higher levels of randomization will be required.

- Correlations can make previous effect worse.
  - **Example:** 80% of the people whose randomized value of age is in [120,130] and whose randomized value of income is [...] have their true age in [70,80].

- Challenge: How do you limit privacy breaches?
CHALLENGES: HOW TO PREVENT PRIVACY BREACHES?

- Enhance privacy awareness of citizens and governments
  - Do not publish your or other’s private information, such as ID numbers, SSC
  - Be aware of private information in online communications and transactions
  - Establish laws for privacy protection
  - Add privacy-preserving properties to data that need to be published

- Develop privacy-preserving technologies
  - Exercising privacy protection at data collection phase
  - Using different privacy-preserving techniques in collaborative analysis
  - Inter-department, inter-business, different types of data need to have different privacy-preserving techniques
DEFINITION AND MEASURE OF PRIVACY

- What is private data
  - National secrets are the secrete data of a nation
  - Personal privacy has different definitions
  - Data value protection and data pattern protection
  - Time interval for private data

- How to measure privacy
  - Can we quantify privacy?
  - How do we consider that privacy is preserved?
  - Possibility of private data being breached?
  - (How to define data utilities?)