# Security Issues in Data Mitigation and Mining: An Overview

Prof. Saleha S. Saudagar
Information Technology Department
Prof. Ram Meghe Institute of Technology
and Research Badnera, India
salehasaudagar@gmail.com

Prof. Priyanka A. Chorey
Information Technology Department
Prof. Ram Meghe Institute of Technology
and Research Badnera, India
priyankachorey@rediffmail.com

Prof. Smeet D. Thakur
Information Technology Department
Prof. Ram Meghe Institute of Technology
and Research Badnera, India
sdthakur@mitra.ac.in

***Abstract*— The improvising trends in information technology has enabled collection and processing of vast amounts of personal information, such as shopping habits, criminal records, credit and medical history, and driving records. This data is undoubtedly very useful in many areas, including medical research, national security and law enforcement. Privacy is commonly seen as the right of each user to control information about themselves. However, there is an increasing public concern about the individuals ' privacy. In this paper, we view the privacy issues related to data mining from a wider perspective and investigate various remedies that can help to protect sensitive data. In particular, we identify four different types of user roles involved in data mining applications, namely, data provider, data collector, data miner, and decision maker. For each kind of user, we focus on his privacy issues and techniques to protect sensitive Knowledge.**

***Keywords—Sensitive Data, Data Mining, Data Provider, Data Collector, Data Miner, Decision Maker.***

## I. Introduction

Data mining is catching tremendous attention in recent years, probably because of the popularity of the ``big data''. Data mining is the technology of extracting interesting patterns and knowledge from large amounts of data [3]. As a purely application-driven discipline, data mining is being applying many domains successfully, such as digital libraries, Web search, business intelligence, scientific discovery, etc.

### *A. The Knowledge Discovery and Data Mining*

Knowledge Discovery from Data (KDD) is often treated as a synonym for another ``Data Mining" which highlights the purpose of the mining process. To obtain useful knowledge from data, the following steps can be performed (see Figure 1):

Step 1: Data preprocessing. Its operations include retrieve data relevant to Knowledge Discovery from the database (data selection), remove noise (data cleaning), and inconsistent data, to handle the missing data fields, etc.) and combine data from multiple Sources(data integration)

Step 2: Data transformation. The aim is to transform data into forms appropriate for the mining, that is, to find useful features to represent the data. Property selection and property transformation are basic operations.

Step 3: Data mining. This is an analytical process where intelligent methods are employed to explore data patterns

Step 4: Pattern extraction and presentation. These basic operations comprise of identifying the interesting patterns which show knowledge, and exploring the mined knowledge in an easy-to-understand way.

### *B. Integration Of Privacy into Data Mining Operations*

Information discovered by data mining can be very valuable to many applications, despite of that demand for security violation remedies are increasing tremendously [2]. Individual's privacy might be disturb due to the unauthorized disclosure to personal information, the undesired disclosure of one's embarrassing information, the use of personal information for reason other than the one for which data has been collected, etc. For instance, the Indian retailer Target once received complaints from a user who was angry that Target sent coupons for newly born baby' product to her teenager sister. .However, it is being consider that the sister was pregnant at that time, and Target correctly guesses the fact

by mining its customer data. From this example, we can observe that the conflict between data mining and privacy security does exist. To deal with the privacy concerns in data mining, a sub- field of data mining, referred to as *privacy preserving data mining* (PPDM) nowadays has gained a tremendous . The objective of PPDM is to safeguard sensitive information from unsolicited or unsanctioned disclosure, and meanwhile, preserve the utility of the data. The consideration of PPDM is two-fold. First, sensitive raw data, such as individual's ID card number and cell phone number, should not be directly used for mining. Second, sensitive mining results, whose disclosure will result in privacy violation, should be excluded. After the pioneering work of Agrawal et al. [3],[4], numerous studies on PPDM have been conducted [5][7].

### C. User Role-Based Techniques

Current models and algorithms proposed for PPDM mainly focus on how to hide those sensitive information from certain mining operations. However, as depicted in Fig. 1, the whole KDD process involve multi-phase operations. Besides the

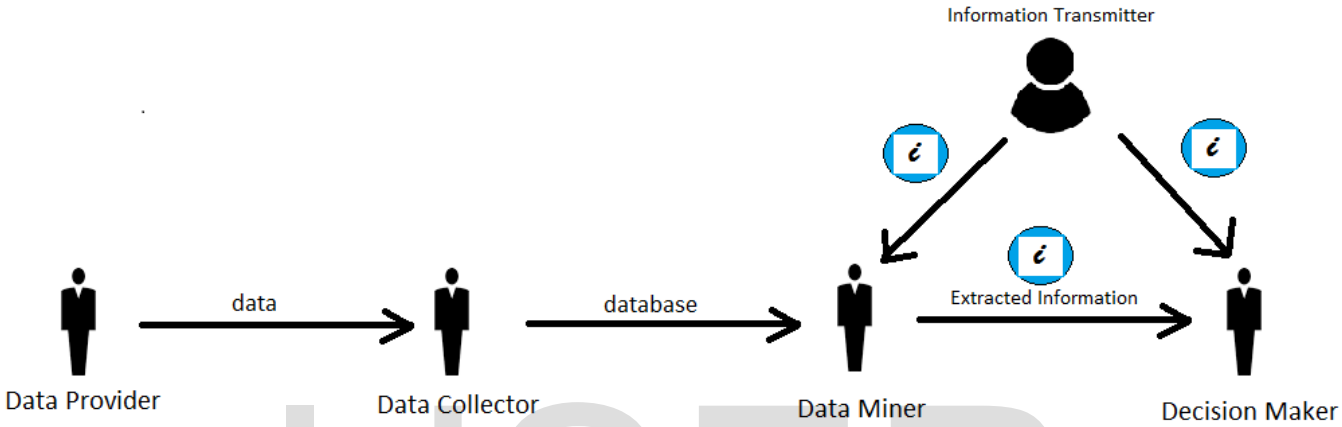


Figure 1. Data Mining Application Scenario



mining phase, privacy issues may also arise in the phase of data collecting or data preprocessing, even in the delivery process of the mining results. In this paper, we investigate the privacy aspects of data mining by considering the whole knowledge-discovery process. We present an overview of the many approaches which can help to make proper use of sensitive data and protect the security of sensitive information discovered by data mining. We use the term ``sensitive information'' to refer to privileged or proprietary information that only certain people are allowed to see and that is therefore not accessible to everyone. If sensitive information is lost or used in any way other than intended, the result can be severe damage to the person or organization to which that information belongs. The term ``sensitive data'' refers to data from which sensitive information can be extracted. Throughout the paper, we consider the two terms ``privacy'' and ``sensitive information'' are interchangeable. In this paper, we develop a user-role based methodology to conduct the review of related studies. Based on the stage division in KDD process (see Fig. 1), we can identify four different types of users, namely four *user roles*, in a typical data mining scenario (see Fig. 2):

- **Data Provider**: the user who owns some data that are desired by the data mining task.
- **Data Collector**: the user who collects data from data providers and then publishes the data to the data miner.
- **Data Miner**: the user who performs data mining tasks on the data.
- **Decision Maker**: the user who makes decisions based on the data mining results in order to achieve certain goals. In the data mining scenario depicted in Fig. 2, a user represents either a person or an organization. Also, one user can play multiple roles at once. For example, in the Target story we mentioned above, the customer plays the role of data provider,

and the retailer plays the roles of data collector, data miner and decision maker. By differentiating the four different user roles, we can explore the privacy issues in data mining in a principled way. All users care about the security of sensitive information, but each user role views the security issue from its own perspective. What we need to do is to identify the privacy problems that each user role is concerned about, and to find appropriate solutions the problems.

### D. Paper Organization

The remainder of this paper is organized as follows: Section 2, 3, 4, 5 discusses the privacy issues and remedies to these

problems for data provider, data collector, data miner and decision maker, respectively. The paper is concluded in Section 6 with some future research direction.

# II. Data Provider

## A. *Privacy Issues Related to Data Provider*

A data provider owns some data from which valuable information can be extracted. In the data mining scenario depicted in Figure 1, there are actually two types of data providers: one refers to the data provider who provides data to data collector, and the other refers to the data collector who provides data to data miner. To differentiate the privacy protecting methods adopted by different user roles, here in this section, we discuss about the ordinary data provider who owns a relatively small amount of data which contain only information about himself. Data reporting information about an individual are often referred to as ``microdata'' [4]. If a data provider reveals his microdata to the data collector, his privacy might be comprised due to the unexpected data breach or exposure of sensitive information. Hence, the privacy concern of a data provider is whether he can take control over what kind of and how much information other people can obtain from his data. To evaluate the ways that the data provider can adopt to protect privacy, we consider the following three situations:

1. If the data provider seems his data to be very sensitive, that is, the data may expose some information that he does not want to reveal it, the provider can just disagree to provide such data. Effective access control measures are desired by the data provider, so that he can stop his sensitive information from being stolen by the data collector.
2. Data Provider knows that his data are valuable to the data collector (as well as the data miner), the data provider may be willing to hand over some of his private data in exchange for certain profit, such as good services or monetary rewards. The data provider has to know how to negotiate with the data collector, so that he will get enough recompense for any possible loss in privacy.
3. If the data provider can neither prevent the access to his sensitive information nor have a profitable deal with the data collector, the data provider can disturb his data that will be fetched by the data collector, so that his true information cannot be easily reveal.

## B. *Remedies for Privacy Protection*

A data provider provides his data to the collector in an active way or a passive way. By ``active'' we mean that the data provider voluntarily opts in a survey initiated by the data collector, or fill in some registration forms to create an account in a website. By ``passive'' we mean that the data, which are generated by the provider's routine activities, are recorded by the data collector, while the data provider may even have no awareness of the disclosure of his data. When the data provider provides his data actively, he can simply ignore the collector's demand for the information that he deems very sensitive. If his data are passively provided to the data collector, the data provider can take some measures to limit the collector's access to his sensitive data.

In some cases, the data provider needs to make a tradeoff between the loss of privacy and the benefits brought by participating in data mining. In such cases the preference of privacy or benefits can affects benefits of data collector.

Once the data have been handed over to others, there is no guarantee that the provider's sensitive information will be safe. So it is important for data provider to make sure his sensitive data are out of reach for anyone untrustworthy..

# III. Data Collector

## A. *Privacy Issues Related to Data Collector*

As shown in Figure 1, data collector takes data from data providers in order to take part in the subsequent data mining operations. The original information collected from data providers usually contain sensitive information about individuals. If the data collector doesn't take sufficient precautions before releasing the data to public or data miners, those sensitive information may be disclosed, even though this is not the collector's original intention.

Modification of the original data is necessary before releasing it to others, so that sensitive information about data providers can neither be found in the modified data nor be inferred by anyone with malicious intent. Generally, the modification will cause a loss in data utility. The data collector should also make sure that sufficient utility of the data can be retained after the modification; otherwise collecting the data will be a wasted effort. The data modification process adopted by data collector, with the goal of preserving privacy and utility simultaneously, is usually called *privacy preserving data publishing* (PPDP).

## B. *Remedies for Privacy Protection*

Privacy-preserving data publishing provides methods to hide identity or sensitive attributes of original data owner. Despite the many advances in the study of data anonymization, there remain some research topics awaiting to be explored. Here we highlight two topics that are important for developing a practically effective anonymization method, namely personalized privacy preservation and modeling the background knowledge of adversaries. Current studies on PPDP mainly manage to achieve privacy preserving in a statistical sense, that is, they focus on a universal approach that exerts the same amount of preservation for all individuals. While in practice, the implication of privacy varies from person to person. For example, someone considers salary to be sensitive information while someone doesn't; someone cares much about privacy while someone cares less. Therefore, the ``personality'' of privacy must be taken into account when anonymizing the data.

# IV. Data Miner

To withdraw useful knowledge from the data received from data collector the data miner applies data mining algorithms which is required by the decision maker. The privacy related issues coming with the data mining operations are twofold. On one hand, if personal information can be directly seen and analyze in the data and data breach happens, privacy of the original data owner (i.e. the data provider) will be compromised.

On the other hand, equipping with the many powerful data mining techniques, the data miner is able to find out various kinds of information underlying the data Sometimes the data mining results may reveal sensitive information about the data owners To encourage data providers to participate in the data mining activity and provide more sensitive data, the data miner confirm that the above two privacy threats should be eliminated.

# V. Decision Maker

The ultimate goal of data mining is to provide useful information to the decision maker. so that the decision maker can choose a better way to achieve his objective, such as increasing sales of products or making correct diagnoses of diseases. As wider perspective, it seems that the decision maker has nothing to do with for protecting privacy, since we usually consider privacy as sensitive information about the original data owners (i.e. data providers)

Provenance, which describes what is the actual source of information and how it evolved over time, can help people estimate the credibility of data. For a decision maker, if he can acquire complete provenance of the data mining results, then he can easily determine whether the mining results are reliable. However, in most cases, provenance of the data mining results is not available. If the mining results are not directly delivered to the decision maker, it is very likely that they are propagated in a less controlled environment. As we introduced earlier, a major approach to represent the provenance information is adding annotations to data.

## *Conclusion*

How to protect sensitive data from the security violator brought by data mining has become a hot topic in recent years. In this paper we have reviewed the privacy issues related to data mining by using user-role based techniques. We differentiate four different user roles that are equally involved in data mining application process, i.e. data provider, data collector, data miner and decision maker. Each user involved has its own privacy issues; hence the privacy-preserving remedies adopted by one user role are generally different from those adopted by others

## *References*

[1] R. Agrawal and R. Srikant, ``Privacy-preserving data mining,'' *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 439_450, 2000.

[2] L. Brankovic and V. Estivill-Castro, ``Privacy issues in knowledge discovery and data mining,'' in *Proc. Austral. Inst. Comput. Ethics Conf.*, 1999, pp. 89_99.

[3] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2006.

[4] Y. Lindell and B. Pinkas, ``Privacy preserving data mining,'' in *Advances in Cryptology*. Berlin, Germany: Springer-Verlag, 2000, pp. 36_54.

[5] C. C. Aggarwal and S. Y. Philip, *A General Survey of Privacy-Preserving Data Mining Models and Algorithms*. New York, NY, USA: Springer-Verlag, 2008.

[6] M. B. Malik, M. A. Ghazi, and R. Ali, ``Privacy preserving data mining techniques: Current scenario and future prospects,'' in *Proc. 3rd Int. Conf. Comput. Commun. Technol. (ICCCT)*, Nov. 2012, pp. 26_32.

[7] M. B. Malik, M. A. Ghazi, and R. Ali, ``Privacy preserving data mining techniques: Current scenario and future prospects,'' in *Proc. 3rd Int. Conf. Comput. Commun. Technol. (ICCCT)*, Nov. 2012, pp. 26_32.

[8] S. Matwin, ``Privacy-preserving data mining techniques: Survey and challenges,'' in *Discrimination and Privacy in the Information Society*. Berlin ,Germany: Springer-Verlag, 2013, pp. 209_221.

[9] A. Narayanan and V. Shmatikov, ``Robust de-anonymization of large sparse datasets,'' in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2008, pp. 111_125.

[10] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, ``Privacy-preserving data publishing: A survey of recent developments,'' *ACM Comput. Surv.*, vol. 42, no. 4, Jun. 2010, Art. ID 14.

[11] R. C.-W. Wong and A. W.-C. Fu, ``Privacy-preserving data publishing: An overview,'' *Synthesis Lectures Data Manage.*, vol. 2, no. 1, pp. 1_138, 2010.

[12] L. Sweeney, ``*k*-anonymity: A model for protecting privacy,'' *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557_570, 2002.

[13] R. J. Bayardo and R. Agrawal, ``Data privacy through optimal k-anonymization,'' in *Proc. 21st Int. Conf. Data Eng. (ICDE)*, Apr. 2005, pp. 217_228.

[14] J. Xu,W.Wang, J. Pei, X.Wang, B. Shi, and A.W.-C. Fu, ``Utility-based anonymization for privacy preservation with less information loss,'' *ACM SIGKDD Explorations Newslett.*, vol. 8, no. 2, pp. 21_30, 2006.

[15] B. Zhou, J. Pei, andW. Luk, ``A brief survey on anonymization techniques for privacy preserving publishing of social network data,'' *ACM SIGKDD Explorations Newslett.*, vol. 10, no. 2, pp. 12_22, 2008.

[16] X. Wu, X. Ying, K. Liu, and L. Chen, ``A survey of privacy-preservation of graphs and social networks,'' in *Managing and Mining Graph Data*. New York, NY, USA: Springer-Verlag, 2010, pp. 421_453.