International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), Volume 5, Special Issue 2, May 2015)
International Conference on Advances in Computer and Communication Engineering (ACCE-2015)

Privacy Preserving Publishing of Social Network Data Privacy and Big Data Mining

Suma Reddy[1], Shilpa G V[2]
[1]PG Scholar, [2]Asst.Professor, Department of Computer Science and Engineering, Vemana IT, Bengaluru-34
[1]sumarreddy@gmail.com, [2]shilpa.gv@gmail.com

***Abstract*— Privacy is a major concern in big data mining. Advances in development of data mining technologies bring serious risk to the security of individual's sensitive data. An emerging research in data mining, known as privacy-preserving data mining (PPDM), has been broadly studied in recent years. The basic idea of PPDM is to transform or change the data in such a way so as to not to compromise on security of individual's sensitive data and also to perform data mining algorithms effectively. In this paper, I have identified four different category of users who are involved in data mining applications and privacy concerns of each category. Basically, any data mining application will have four kind of users namely, data provider, data collector, data miner and decision maker. I briefly introduce the basics of related research topics and current approaches, and present some basic thoughts on future research. By differentiating the responsibilities of different users with respect to security of sensitive information, I would like to provide some useful insights into the study of PPDM on social networking data. One main characteristic of social networks is that they keep evolving over time. The data collector needs to publish the network data periodically. The privacy issue in sequential publishing of dynamic social network data has recently attracted researchers' attention.**

***Keywords*— Data mining, sensitive data, privacy-preserving data mining, and social network**

I. INTRODUCTION

Data mining has become more and more popular in recent years, probably because of the popularity of the ''big data'' concept. Data mining is the process of discovering interesting patterns and knowledge from large data [2]. As a highly application-driven discipline, data mining has been successfully applied to many domains, such as business intelligence, Web search, scientific discovery, digital libraries, etc.

II. THE KDD PROCESS

''Data mining'' is often treated as ''knowledge discovery from databases'' KDD which highlights the goal of the mining process. KDD has the following steps to perform in an iterative way to get useful patterns and knowledge from (Fig. 1):

III. THE PRIVACY CONCERN AND PPDM

Despite that the information discovered by data mining can be very valuable to many applications, people have shown increasing concern about the other side of the coin, namely the privacy threats posed by data mining [3]. Individual's privacy may be violated due to the unauthorized access to personal data, the undesired discovery of one's embarrassing information, the use of personal data for purposes other than the one for which data has been collected, etc.

To deal with the privacy issues in data mining, privacy preserving data mining (PPDM) has gained a great development in recent years. The objective of PPDM is to safeguard sensitive information from unasked 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 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. [4], [5], numerous studies on PPDM have been conducted [6]–[8].

In this paper, we develop a user-role based methodology to conduct the review of related studies. Based on the stages in KDD process (Fig. 1), we can identify four different types of users, namely four *user roles*, in a typical data mining scenario:

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), Volume 5, Special Issue 2, May 2015)
International Conference on Advances in Computer and Communication Engineering (ACCE-2015)

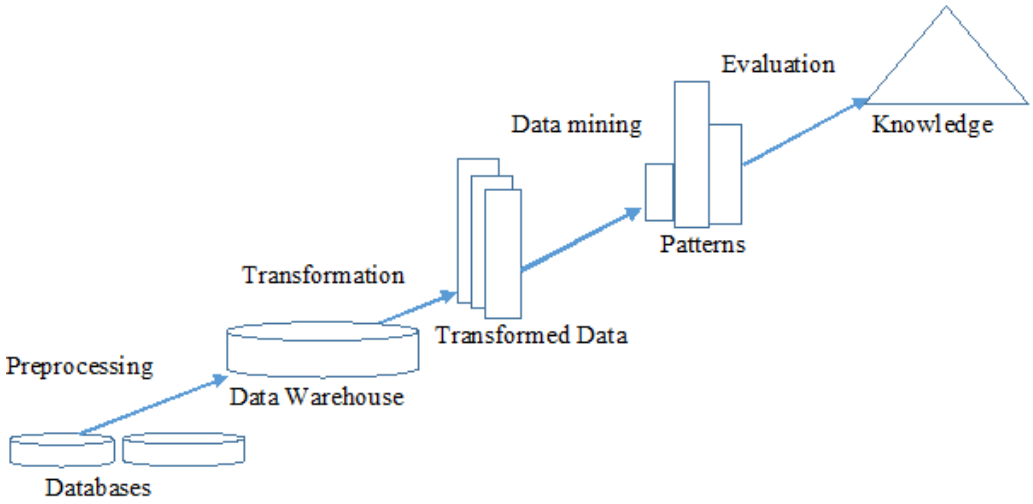

Figure I. Kdd Process

- *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 publish 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.

Here we briefly describe the privacy concerns of each user role. Detailed discussions will be presented in following sections.

IV. Data Provider

The data provider's major concern is whether he can control the sensitivity of the data he provides to data collector. Primarily, the provider should be able to make sure his private data will not be known anyone else. Secondly, if the provider has to provide some data to the data collector, he wants to hide his sensitive information as much as possible and get enough cost for the possible loss in privacy.

A. Concerns For Data Provider

A user (data provider) owns some data from which sensitive information can be extracted. In the data mining scenario, there are actually two types of data providers: one is the data provider who gives data to data collector and data collector in turn acts a data provider to the data miner.

To distinguish the privacy preserving methods adopted by different user roles, here in this section, we restrict ourselves to the ordinary data provider, the one who owns a relatively small amount of data which contain only information about herself. Data reporting information about an individual are often referred to as ``microdata" [10]. If a data provider reveals his microdata to the data collector, his privacy might be comprised due to the exposure of sensitive information. So, the privacy concern of a data provider is can he take command over what kind of and how much information others can obtain from his/her data. To investigate the measures that the data provider can adopt to protect privacy, we consider the following three situations:

- If the data provider considers his/her data may reveal some information that he does not want anyone else to know, the provider can just refuse to provide such data. Effective access control measures are desired by the data provider, so that he can prevent his sensitive data from being stolen by the data collector.
- Realizing 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 benefits, such as better services or monetary rewards. The data provider needs to know how to negotiate with the data collector, so that he will get enough compensation for any possible loss in privacy.
- If the data provider can neither prevent the access to his sensitive data nor make a lucrative deal with the data collector, the data provider can distort his data that will be fetched by the data collector, so that his true information cannot be easily disclosed.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), Volume 5, Special Issue 2, May 2015)
International Conference on Advances in Computer and Communication Engineering (ACCE-2015)

B. Approaches To Privacy Protection

- *LIMIT THE ACCESS:* A data provider provides his data to the collector knowingly or unknowingly. Knowingly means provider actively participates in collector's survey. Unknowingly means data collector might have collected provider's data from collector's routine activities. While the provider may not have any clue of disclosure of data. Current security tools are three types:
    1. Anti-tracking extensions:
    2. Advertisement and script blockers.
    3. Encryption tools.

To make sure a private online communication between two parties cannot be intercepted by third parties, a user can utilize encryption tools.

In addition to the tools mentioned above, an Internet user should always use anti-virus and anti-malware tools to protect his data that are stored in digital equipment such as personal computer, cell phone and tablet. With the help of all these security tools, the data provider can limit other's access to his personal data.

- *BENEFIT OF TRADE PRIVACY:* In some cases, the data provider needs to make a trade-off between the loss of privacy and benefits brought by participating in data-mining. For example, a shopping website may collect our personal information like age, phone number and salary. If the data provider considers his salary as a sensitive information he can give a fuzzy value like 10,000 INR < Salary < 20,000 INR.
- *PROVIDE FALSE DATA:* Some methods to falsify data are "sockpuppets" to hide one's actual activities, fake identity, mask one's identity using security tools.

V. Data Collector

The data collected from data providers may contain individuals' sensitive information. Directly releasing the data to the data miner will violate data providers' privacy, so data modification is required. On the other hand, the data should be useful even after modification, otherwise collecting the data will be of no use. Thus, the major concern of data collector is to make sure that the modified data contain no sensitive information but yet preserve high utility.

A. Concerns Of Data Collector

Data collector collects data from data provider in the data mining process. The original data collected from data provider contains sensitive information about data provider. If the data collector doesn't take enough precautions before delivering data to data miner or public those sensitive information may be disclosed.

B. Approaches To Privacy Protection

1) Basics Of PPDP

PPDP mainly studies anonymization approaches for Publishing useful data while preserving privacy. The original data is assumed to be a private table consisting of multiple records. Each record consists of the following 4 types of attributes:

- Identifier (ID): Attributes that can uniquely identify an individual, such as ID and mobile number.
- Quasi-identifier (QID): Attributes that can be linked with external data to re-identify individual records, such as gender, age and zip code.
- Sensitive Attribute (SA): Attributes that an individual don't want to disclose, such as disease and salary.
- Non-sensitive Attribute (NSA): Attributes other than ID, QID and SA.

Before being published to others, the table is anonymized, that is, identifiers are removed and quasi-identifiers are modified. As a result, individual's identity and sensitive attribute values can be hidden from adversaries.

How the data table should be anonymized mainly depends on how much privacy we want to preserve in the anonymized data. Different privacy models have been proposed to quantify the preservation of privacy. Based on the attack model which describes the ability of the adversary in terms of identifying a target individual, privacy models can be roughly classified into two categories. The first category considers that the adversary is able to identify the record of a target individual by linking the record to data from other sources, such as liking the record to a record in a published data table (called *record linkage*), to a sensitive attribute in a published data table (called *attribute linkage*), or to the published data table itself (called *table linkage*). The second category considers that the adversary has enough background knowledge to carry out a *probabilistic attack*, that is, the adversary is able to make a confident inference about whether the target's record exist in the table or which value the target's sensitive attribute would take. Typical privacy models [18] includes *k*-anonymity (for preventing record linkage), *l*-diversity (for preventing record linkage and attribute linkage), *t*-closeness (for preventing attribute linkage and probabilistic attack), *epsilon*-differential privacy (for preventing table linkage and probabilistic attack), etc.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), Volume 5, Special Issue 2, May 2015)
International Conference on Advances in Computer and Communication Engineering (ACCE-2015)

TABLE 1

Age	Sex	Zip code	Disease
5	Male	560001	HIV
14	Female	560017	Cancer
23	Male	560034	Gastritis
20	Female	566004	HIV
9	Male	560009	Cancer
7	Female	560009	flu
15	Female	561007	pulmonary
50	Male	560009	Parkinson's
25	Male	560012	Flu

TABLE 2

Age	Sex	Zip code	Disease
[1, 10]	People	5*****	HIV
[1, 10]	People	5*****	Cancer
[1, 10]	People	5*****	flu
[11,20]	People	5*****	HIV
[11,20]	People	5*****	pulmonary
[11,20]	People	5*****	Gastritis
[21,50]	People	5*****	pulmonary
[21,50]	People	5*****	Parkinson's
[21,50]	People	5*****	Gastritis

FIGURE 2. An example of 2-anonymity, where QID = {Age; Sex; Zipcode}. (a) Original table. (b) 2-anonymous table.

Among the many privacy models, *k*-anonymity and its variants are most widely used. The idea of *k*-anonymity is to modify the values of quasi-identifiers in original data table, so that every tuple in the anonymized table is indistinguishable from at least *k*-1 other tuples along the quasi-identifiers. The anonymized table is called a *k*-anonymous table. Fig. 2 shows an example of *2*-anonymity. Intuitionally, if a table satisfies *k*-anonymity and the adversary only knows the quasi-identifier values of the target individual, then the probability that the target's record being identified by the adversary will not exceed 1/*k*. To make the data table satisfy the requirement of a specified privacy model, one can apply the following anonymization operations [18]:

- Generalization.
- Suppression.
- Anatomization.
- Permutation.
- Perturbation.

The anonymization operations will reduce the utility of data. The reduction of data utility is usually represented by *information loss*: higher information loss means lower utility of the anonymized data. Various metrics for measuring information loss have been proposed, such as minimal distortion [20], discernibility metric [21], the normalized average equivalence class size metric [22], weighted certainty penalty [23], information-theoretic metrics [24], etc. A fundamental problem of PPDP is how to make a tradeoff between privacy and utility. Given the metrics of privacy preservation and information loss, current PPDP algorithms usually take a greedy approach to achieve a proper tradeoff: multiple tables, all of which satisfy the requirement of the specified privacy model, are generated during the anonymization process, and the algorithm outputs the one that minimizes the information loss.

2) Privacy-Preserving Publishing Of Social Network Data

Social networks have gained great development in recent years. Aiming at discovering interesting social patterns, social network analysis becomes more and more important. To support the analysis, the company who runs a social network application sometimes needs to publish its data to a third party. However, even if the truthful identifiers of individuals are removed from the published data, which is referred to as naïve anonymized, publication of the network data may lead to exposures of sensitive information about individuals, such as one's intimate relationships with others. Therefore, the network data need to be properly anonymized before they are published.

A social network is usually modelled as a graph, where the vertex represents an entity and the edge represents the relationship between two entities. Thus, PPDP in the context of social networks mainly deals with anonymizing graph data, which is much more challenging than anonymizing relational table data. Zhou et al. [25] have identified the following three challenges in social network data anonymization:

First, modelling adversary's background knowledge about the network is much difficult. For relational data tables, a small set of quasi-identifiers are used to define the attack models. While given the network data, various information, such as attributes of an entity and relationships between different entities, may be utilized by the adversary.

Second, measuring the information loss in anonymizing social network data is harder than that in anonymizing relational data.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), Volume 5, Special Issue 2, May 2015)
International Conference on Advances in Computer and Communication Engineering (ACCE-2015)

It is difficult to determine whether the original network and the anonymized network are different in certain properties of the network.

Third, devising anonymization methods for social network data is much harder than that for relational data. Anonymizing a group of tuples in a relational table does not affect other tuples. However, when modifying a network, changing one vertex or edge may affect the rest of the network. Therefore, ``divide-and-conquer'' methods, which are widely applied to relational data, cannot be applied to network data. To deal with above challenges, many approaches have been proposed. According to [26], anonymization methods on simple graphs, where vertices are not associated with attributes and edges have no labels, can be classified into three categories, namely edge modification, edge randomization, and clustering-based generalization. Comprehensive surveys of approaches to on social network data anonymization can be found in [19], [26], and [27]. In this paper, we briefly review some of the very recent studies, with focus on the following three aspects: attack model, privacy model, and data utility.

3) Attack Model

Given the anonymized network data, adversaries usually rely on background knowledge to de-anonymize individuals and learn relationships between de-anonymized individuals. Zhou et al. [25] identify six types of the background knowledge, i.e. attributes of vertices, vertex degrees, link relationship, neighbourhoods, embedded subgraphs and graph metrics. Peng et al. [28] propose an algorithm called *Seed-and-Grow* to identify users from an anonymized social network graph, solely based on graph structure. The algorithm identifies a seed sub-graph which is either planted by an attacker or divulged by collusion of a small user group, and then grows the seed larger based on the adversary's existing knowledge of users' social relations. Zhu et al. [29] design a *structural attack* to de-anonymize data of social graph.

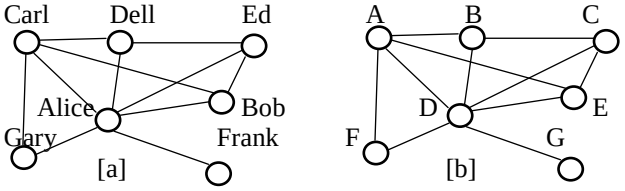

FIGURE 3. Instance of mutual friend attack: (a) original network; (b) naïve anonymized network.

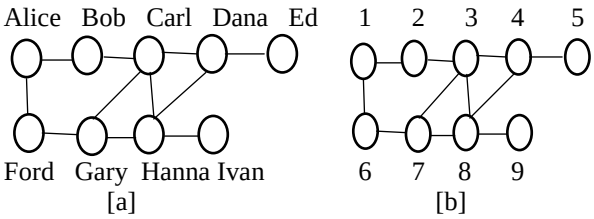

FIGURE 4. Instance of friend attack: (a) original network; (b) naïve anonymized network.

The attack makes use of the cumulative degree of *n*-hop neighbours of a vertex as the regional feature, and combines it with the simulated annealing-based graph matching method to re-identify vertices in anonymous social graphs. Sun et al. [30] introduce a relationship attack model called *mutual friend attack*, which is based on the number of mutual friends of two connected individuals. Fig. 3 shows an instance of the mutual friend attack. The original social network *G* with vertex identities is shown in Fig. 3[a], and Fig. 3[b] shows the corresponding anonymized network where all individuals' names are removed. In this network, only Alice and Bob have 4 mutual friends. If an adversary knows this information, then he can uniquely re-identify the edge *D*; *E* in Fig. 3[b] is *Alice and Bob*. In [31], Tai et al. investigate the *friendship attack* where an adversary utilizes the degrees of two vertices connected by an edge to re-identify related victims in a published social network data set. Fig. 4 shows an example of friendship attack. Suppose that each user's friend count (i.e. the degree of the vertex) is publicly available. If the adversary knows that Bob has 2 friends and Carl has 4 friends, and he also knows that Bob and Carl are friends, then he can uniquely identify that the edge 2 and 3 in Fig. 4[b] corresponds to B*ob* and *Carl*. In [32], another type of attack, namely *degree attack*, is explored. The motivation is that each individual in a social network is inclined to associate with not only a vertex identity but also a community identity, and the community identity rejects some sensitive information about the individual. It has been shown that, based on some background knowledge about vertex degree, even if the adversary cannot precisely identify the vertex corresponding to an individual, community information and neighbourhood information can still be inferred. For instance, the network shown in Fig. 5 have two communities, and the community identity reveals sensitive information (i.e. status of disease) about its members. In case if an adversary knows Jhon has 5 friends, then he can infer that Jhon has AIDS, even though if he is not sure which of the two vertices vertex 2 and vertex 3 in the anonymized network (Fig. 5[b]) corresponds to Jhon.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), Volume 5, Special Issue 2, May 2015)
International Conference on Advances in Computer and Communication Engineering (ACCE-2015)

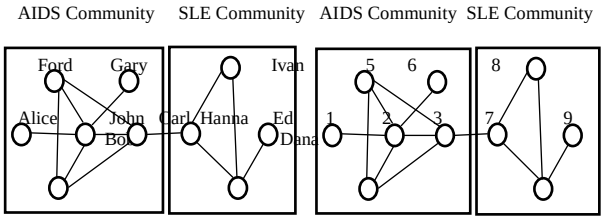

FIGURE 5. Instance of degree attack: (a) original network; (b) naïve anonymized network.

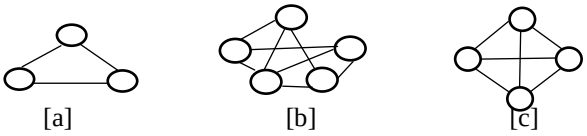

FIGURE 6. Instances of k-NMF anonymity: (a) 3-NMF; (b) 4-NMF; (c) 6-NMF.

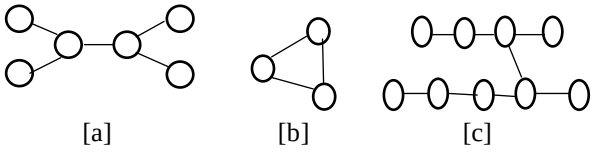

FIGURE 7. Instances of k2-degree anonymous graphs: [a] 22-degree; [b] 32-degree; [c] 22-degree.

From above discussion we can see that, the graph data contain rich information that can be explored by the adversary to initiate an attack. Modelling the background knowledge of the adversary is difficult yet very important for deriving the privacy models.

A. Privacy Model

A *n*-anonymity model, which is a base for number of privacy models have been proposed for graph data. Some of the models have been summarized in the survey [33], such as *k*-degree, *k*-neighbourhood, *k*-automorphism, *k*-isomorphism, and *k*-symmetry. In order to protect the privacy of relationship from the mutual friend attack, Sun et al. [30] introduce a variant of *k*-anonymity, called *k-NMF* anonymity. *NMF* is a property defined for the edge in an undirected simple graph, representing the number of mutual friends between the two individuals linked by the edge. If a network satisfies *k-NMF* anonymity (see Fig. 6), then for each edge *e*, there will be at least *k* -1 other edges with the same number of mutual friends as *e*. It can be guaranteed that the probability of an edge being identified is not greater than 1/*k*. Tai et al. [31] introduce the concept of *k*2-degree anonymity to prevent friendship attacks. A graph *G* is *k*2-degree anonymous if, for every vertex with an incident edge of degree pair (*d*1; *d*2) in *G*, there exist at least *k* - 1 other vertices, such that each of the *k* -1 vertices also has an incident edge of the same degree pair (Fig. 7).

Intuitively, if a graph is *k*2-degree anonymous, then the probability of a vertex being re-identified is not greater than 1/*k*, even if an adversary knows a certain degree pair (*dA*, *dB*) where community ID is indicated beside each vertex.

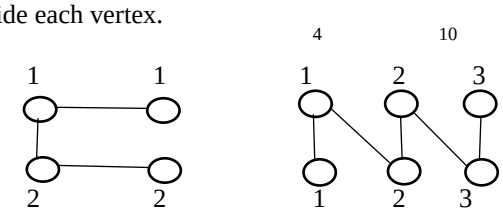

FIGURE 8. Examples of 2-structurally diverse graphs, where the community ID is indicated beside each vertex.

A and *B* are friends. To prevent degree attacks, Tai et al. [32] introduce the concept of *structural diversity*. A graph satisfies *k-structural diversity anonymization* (*k-SDA*), if for every vertex *v* in the graph, there are at least *k* communities, such that each of the communities contains at least one vertex with the same degree as *v* (Fig. 8). In other words, for each vertex *v*, there are at least *k* -1 other vertices located in at least *k* - 1 other communities.

B. Data Utility

In the network data anonymization, the implication of data utility is: whether and to what extent properties of the graph are preserved. Wu et al. [26] summarize three types of properties considered in current studies. The first type is graph topological properties, which are defined for applications aiming at analyzing graph properties. Various measures have been proposed to denote the structure characteristics of the network. The second type is graph spectral properties. The spectrum of a graph is usually defined as the set of eigenvalues of the graph's adjacency matrix or other derived matrices, which has close relations with many graph characteristics. The third type is aggregate network queries. An aggregate network query calculates the aggregate on some paths or subgraphs satisfying some query conditions. The accuracy of answering aggregate network queries can be considered as the measure of utility preservation. Most existing *k*-anonymization algorithms for network data publishing perform edge insertion and/or deletion operations, and they try to reduce the utility loss by minimizing the changes on the graph degree sequence. Wang et al. [34] consider that the degree sequence only captures limited structural properties of the graph and the derived anonymization methods may cause large utility loss.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), Volume 5, Special Issue 2, May 2015)
International Conference on Advances in Computer and Communication Engineering (ACCE-2015)

They propose utility loss measurements built on the community-based graph models, including both the community model and the hierarchical community model, to better capture the impact of anonymization on network topology. One important characteristic of social networks is that they keep evolving over time. Sometimes the data collector needs to publish the network data periodically. The privacy issue in sequential publishing of dynamic social network data has recently attracted researchers' attention. Medforth and Wang [35] identify a new class of privacy attack, named *degree-trail attack*, arising from publishing a sequence of graph data. They demonstrate that even if each published graph is anonymized by strong privacy preserving techniques, an adversary with little background knowledge can re-identify the vertex belonging to a known target individual by comparing the degrees of vertices in the published graphs with the degree evolution of a target. In [36], Tai et al. adopt the same attack model used in [35], and propose a privacy model called dynamic *kw-structural diversity anonymity* (*kw-SDA*), for protecting the vertex and multi-community identities in sequential releases of a dynamic network. The parameter *k* has a similar implication as in the original *k*-anonymity model, and *w* denotes a time period that an adversary can monitor a target to collect the attack knowledge. They develop a heuristic algorithm for generating releases satisfying this privacy requirement.

VI. Data Miner

The data miner applies mining algorithms to the data provided by data collector, and he wishes to extract useful information from data in a privacy-preserving manner. PPDM covers two types of protections, namely the protection of the sensitive data themselves and the protection of sensitive mining results. With the user role-based methodology proposed in this paper, we consider the data collector should take the major responsibility of protecting sensitive data, while data miner can focus on how to hide the sensitive mining results from untrusted parties.

A. Concerns For Data Provider

In order to discover useful pattern desired by the decision maker, the data miner applies data mining algorithms to the data received from data collector. The privacy issues coming with the data mining operations are two types. On one hand, if personal information can be directly observed in the data and data breach happens, privacy of the 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 owners. Different from existing surveys on privacy-preserving data mining (PPDM), in this paper, we consider it is the data collector's responsibility to ensure that sensitive raw data are modified or trimmed out from the published data. The primary concern of data miner is how to prevent sensitive information from appearing in the mining results. To perform a privacy-preserving data mining, the data miner usually needs to modify the data he got from the data collector. As a result, the decline of data utility is inevitable.

B. Approaches To Privacy Protection

Extensive PPDM approaches have been proposed (see [6]_[8] for detailed surveys). These approaches can be classified by different criteria [54], such as data distribution, data modification method, data mining algorithm, etc. Based on the distribution of data, PPDM approaches can be categorized into two categories, namely approaches for centralized data mining and approaches for distributed data mining. Distributed data mining can be further categorized into data mining over horizontally partitioned data and data mining over vertically partitioned data. Based on the technique adopted for data modification, PPDM can be classified into perturbation-based, blocking-based, swapping based, etc. Since we define the privacy-preserving goal of data miner as preventing sensitive information from being revealed by the data mining results, in this section, we classify PPDM approaches according to the type of data mining tasks. Specially, we review recent studies on privacy-preserving association rule mining, privacy-preserving classification, and privacy-preserving clustering, respectively.

VII. Privacy-Preserving Association Rule Mining

Various kinds of approaches have been proposed to perform association rule hiding [57], [58]. These approaches can roughly be categorized as below:

- Heuristic distortion approaches: resolve how to select the appropriate data sets for data modification.
- Heuristic blocking approaches: reduce the degree of support and confidence of the sensitive association rules by replacing certain attributes of some data items with a specific symbol (e.g. `?').

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), Volume 5, Special Issue 2, May 2015)
International Conference on Advances in Computer and Communication Engineering (ACCE-2015)

- Probabilistic distortion approaches: distort the data through random numbers generated from a predefined probability distribution function.
- Exact database distortion approaches: formulate the solution of the hiding problem as a constraint satisfaction problem (CSP), and apply linear programming approaches to its solution.
- Reconstruction-based approaches: generate a database from the scratch that is compatible with a given set of non-sensitive association rules.

VIII. Decision Maker

A decision maker can get the data mining results directly from the data miner, or from some *Information Transmitter*. The information transmitter is likely to change the mining results intentionally or unintentionally, which may cause serious loss to the decision maker. Therefore the decision maker concerns is whether the mining results are credible. In addition to investigate the privacy-protection approaches adopted by each user role.

A. Concerns Of Decision Maker

The obvious objective 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 goal, such as increasing sales of products or making proper diagnoses of diseases. At first glance, it seems to be decision maker has no responsibility for protecting privacy, since we usually misinterpret privacy as sensitive information about the data providers or owners. Usually, the data provider, the data collector and the data miner himself are considered to be responsible for the safety of privacy. However, if we look at the privacy issue from a wider perspective, it seems that the decision maker also has his own privacy concerns. The data mining results provided by the data miner are of high importance to the decision maker. If the results are disclosed to someone else, e.g. a competing company, the decision maker may undergo a loss. That is to say, from the perspective of decision maker, the data mining results are sensitive information. On the other hand, if the decision maker does not get the data mining results directly from the data miner, but from someone else which we called *information transmitter*, the decision maker should be accountable for the credibility of the results, in case that the results have been distorted. So, the privacy concerns of the decision maker are two type: how to prevent unwanted disclosure of sensitive mining results, and how to evaluate the credibility of the received mining results.

B. Approaches To Privacy Protection

To deal with the first privacy issue proposed above, i.e. to prevent unwanted disclosure of sensitive mining results, usually the decision maker has to resort to legal measures. For example, making a contract with the data miner to forbid the miner from disclosing the mining results to a third party. To handle the second issue, i.e. to determine whether the received information can be trusted, the decision maker can utilize methodologies from data provenance, credibility analysis of web information, or other related research fields. In the rest part of this section, we will first briefly review the studies on data provenance and web information credibility, and then present a preliminary discussion about how these studies can help to analyze the credibility of data mining results.

Below are some of the approaches to privacy protection:

- Data Provenance
- Web information Credibility

IX. Future Research Directions

PPDP and PPDM provide methods to explore the utility of data while preserving privacy. However, most current studies only manage to achieve privacy preserving in a statistical sense. Considering that the definition of privacy is essentially personalized, developing methods that can support personalized privacy preserving is an important direction for the study of PPDP and PPDM. As mentioned in Section II, some researchers have already investigated the issue of personalized anonymization, but most current studies are still in the theoretical stage. Developing practical personalized anonymization methods is in urgent need. Besides, introducing personalized privacy into other types of PPDP/PPDM algorithms is also required. In addition, since complex socioeconomic and psychological factors are involved, quantifying individual's privacy preference is still an open question which expects more exploration.

X. Conclusion

How to protect sensitive information from the security threats brought by data mining has become a hot topic in recent years. In this paper we review the privacy issues related to data mining by using a user-role based methodology. We distinguish four different user roles that are commonly involved in data mining applications, i.e. data provider, data collector, data miner and decision maker.

Each user role has its own privacy concerns, hence the privacy-preserving approaches adopted by one user role are generally different from those adopted by others:

For data provider, his privacy-preserving objective is to effectively control the amount of sensitive data revealed to others. To achieve this goal, he can utilize security tools to limit other's access to his data, sell his data at auction to get enough compensations for privacy loss, or falsify his data to hide his true identity.

For data collector, his privacy-preserving objective is to release useful data to data miners without disclosing data providers' identities and sensitive information about them. To achieve this goal, he needs to develop proper privacy models to quantify the possible loss of privacy under different attacks, and apply anonymization techniques to the data.

For data miner, his privacy-preserving objective is to get correct data mining results while keep sensitive information undisclosed either in the process of data mining or in the mining results. To achieve this goal, he can choose a proper method to modify the data before certain mining algorithms are applied to, or utilize secure computation protocols to ensure the safety of private data and sensitive information contained in the learned model.

For decision maker, his privacy-preserving objective is to make a correct judgement about the credibility of the data mining results he's got. To achieve this goal, he can utilize provenance techniques to trace back the history of the received information, or build classifier to discriminate true information from false information. To achieve the privacy-preserving goals of different user roles, various methods from different research fields are required. We have reviewed recent progress in related studies, and discussed problems awaiting to be further investigated.

REFERENCES

[1] LEI XU, CHUNXIAO JIANG AND JIAN WANG, Information Security in Big Data: Privacy and Data Mining Beijing 100084, China
[2] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. San Mateo, CA, USA: Morgan Kaufmann, 2006.
[3] 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.
[4] R. Agrawal and R. Srikant, ``Privacy-preserving data mining," ACM SIGMOD Rec., vol. 29, no. 2, pp. 439_450, 2000.
[5] Y. Lindell and B. Pinkas, ``Privacy preserving data mining," in Advances in Cryptology. Berlin, Germany: Springer-Verlag, 2000, pp. 36_54.
[6] 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.
[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] E. Rasmusen, Games and Information: An Introduction to Game Theory, vol. 2. Cambridge, MA, USA: Blackwell, 1994.
[10] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, ``Microdata protection," in Secure Data Management in Decentralized Systems. New York, NY, USA: Springer-Verlag, 2007, pp. 291_321.
[11] O. Tene and J. Polenetsky, ``To track or `do not track': Advancing transparency and individual control in online behavioral advertising," Minnesota J. Law, Sci. Technol., no. 1, pp. 281_357, 2012.
[12] R. T. Fielding and D. Singer. (2014). Tracking Preference Expression (DNT). W3C Working Draft. [Online]. Available: http://www.w3.org/ TR/2014/WD-tracking-dnt-20140128/
[13] R. Gibbons, A Primer in Game Theory. Hertfordshire, U.K.: Harvester Wheatsheaf, 1992.
[14] D. C. Parkes, ``Iterative combinatorial auctions: Achieving economic and computational ef_ciency," Ph.D. dissertation, Univ. Pennsylvania, Philadelphia, PA, USA, 2001.
[15] S. Carter, ``Techniques to pollute electronic pro_ling," U.S. Patent 11/257 614, Apr. 26, 2007. [Online]. Available: https://www.google.com/patents/US20070094738
[16] Verizon Communications Inc. (2013). 2013 Data Breach Investigations Report. [Online]. Available: http://www.verizonenterprise.com/ resources/reports/rp_data-breach-investigations-report-2013_en_xg.pdf
[17] A. Narayanan and V. Shmatikov, ``Robust de-anonymization of large sparse datasets," in Proc. IEEE Symp. Secur. Privacy (SP), May 2008, pp. 111_125.
[18] 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.
[19] 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.
[20] L. Sweeney, ``k-anonymity: A model for protecting privacy," Int. J. Uncertainty, Fuzziness Knowl.-Based Syst., vol. 10, no. 5, pp. 557_570, 2002.
[21] 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.
[22] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, ``Mondrian multidimensional k-anonymity," in Proc. 22nd Int. Conf. Data Eng. (ICDE), Apr. 2006, p. 25.
[23] 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.
[24] A. Gionis and T. Tassa, ``k-anonymization with minimal loss of information," IEEE Trans. Knowl. Data Eng., vol. 21, no. 2, pp. 206_219, Feb. 2009.
[25] 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.
[26] [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.
[27] S. Sharma, P. Gupta, and V. Bhatnagar, ``Anonymisation in social network: A literature survey and classi_cation," Int. J. Soc. Netw. Mining, vol. 1, no. 1, pp. 51_66, 2012.

[28] W. Peng, F. Li, X. Zou, and J. Wu, ``A two-stage deanonymization attack against anonymized social networks," IEEE Trans. Comput., vol, 63, no. 2, pp. 290_303, 2014.
[29] T. Zhu, S. Wang, X. Li, Z. Zhou, and R. Zhang, ``Structural attack to anonymous graph of social networks," Math. Problems Eng., vol. 2013, Oct. 2013, Art. ID 237024.
[30] C. Sun, P. S. Yu, X. Kong, and Y. Fu. (2013). ``Privacy preserving social network publication against mutual friend attacks." [Online]. Available: http://arxiv.org/abs/1401.3201
[31] C.-H. Tai, P. S. Yu, D.-N. Yang, and M.-S. Chen, ``Privacy-preserving social network publication against friendship attacks," in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2011, pp. 1262_1270.
[32] C.-H. Tai, P. S. Yu, D.-N. Yang, and M.-S. Chen, ``Structural diversity for resisting community identi_cation in published social networks," IEEE Trans. Knowl. Data Eng., vol. 26, no. 1, pp. 235_252, Nov. 2013.
[33] M. I. Hafez Ninggal and J. Abawajy, ``Attack vector analysis and privacypreserving social network data publishing," in Proc. IEEE 10th Int. Conf. Trust, Secur. Privacy Comput. Commun. (TrustCom), Nov. 2011, pp. 847_852.
[34] Y. Wang, L. Xie, B. Zheng, and K. C. K. Lee, ``High utility k-anonymization for social network publishing," Knowl. Inf. Syst., vol. 36, no. 1, pp. 1_29, 2013.
[35] N. Medforth and K. Wang, ``Privacy risk in graph stream publishing for social network data," in Proc. IEEE 11th Int. Conf. Data Mining (ICDM), Dec. 2011, pp. 437_446.
[36] C.-H. Tai, P.-J. Tseng, P. S. Yu, and M.-S. Chen, ``Identity protection in sequential releases of dynamic networks," IEEE Trans. Knowl. Data Eng., vol. 26, no. 3, pp. 635_651, Mar. 2014.
[37] G. Ghinita, Privacy for Location-Based Services (Synthesis Lectures on Information Security, Privacy, and Trust). San Rafael, CA, USA: Morgan & Claypool, 2013.
[38] M. Wernke, P. Skvortsov, F. Dürr, and K. Rothermel, ``A classi_cation of location privacy attacks and approaches," Pers. Ubiquitous Comput., vol. 18, no. 1, pp. 163_175, Jan. 2014.
[39] M. Terrovitis and N. Mamoulis, ``Privacy preservation in the publication of trajectories," in Proc. 9th Int. Conf. Mobile Data Manage. (MDM), 2008, pp. 65_72.
[40] M. E. Nergiz, M. Atzori, and Y. Saygin, ``Towards trajectory anonymization: A generalization-based approach," in Proc. SIGSPATIAL ACM GIS Int. Workshop Secur. Privacy GIS LBS, 2008, pp. 52_61.
[41] O. Abul, F. Bonchi, and M. Nanni, ``Never walk alone: Uncertainty for anonymity in moving objects databases," in Proc. IEEE 24th Int. Conf. Data Eng. (ICDE), Apr. 2008, pp. 376_385.
[42] R. Yarovoy, F. Bonchi, L. V. S. Lakshmanan, and W. H. Wang, ``Anonymizing moving objects: Howto hide aMOBin a crowd?" in Proc.12th Int. Conf. Extending Database Technol., Adv. Database Technol., 2009, pp. 72_83.
[43] R. Chen, B. C. M. Fung, N. Mohammed, B. C. Desai, and K. Wang, ``Privacy-preserving trajectory data publishing by local suppression," Inf. Sci., vol. 231, pp. 83_97, May 2013.
[44] M. Ghasemzadeh, B. C. M. Fung, R. Chen, and A. Awasthi, ``Anonymizing trajectory data for passenger _ow analysis," Transp. Res. C, Emerg. Technol., vol. 39, pp. 63_79, Feb. 2014.
[45] A. E. Cicek, M. E. Nergiz, and Y. Saygin, ``Ensuring location diversity in privacy-preserving spatio-temporal data publishing," VLDB J., vol. 23, no. 4, pp. 1_17, 2013.
[46] G. Poulis, S. Skiadopoulos, G. Loukides, and A. Gkoulalas-Divanis, ``Distance-based k_ m-anonymization of trajectory data," in Proc. IEEE 14th Int. Conf. Mobile Data Manage. (MDM), vol. 2. Jun. 2013, pp. 57_62.
[47] F. Bonchi, L. V. S. Lakshmanan, and H.W.Wang, ``Trajectory anonymity in publishing personal mobility data," ACM SIGKDD Explorations Newslett., vol. 13, no. 1, pp. 30_42, Jun. 2011.
[48] X. Xiao and Y. Tao, ``Personalized privacy preservation," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2006, pp. 229_240.
[49] K. Qing-Jiang,W. Xiao-Hao, and Z. Jun, ``The (p, _, k) anonymity model for privacy protection of personal information in the social networks," in Proc. 6th IEEE Joint Int. Inf. Technol. Artif. Intell. Conf. (ITAIC), vol. 2. Aug. 2011, pp. 420_423.
[50] B. Wang and J. Yang, ``Personalized (_, k)-anonymity algorithm based on entropy classi_cation," J. Comput. Inf. Syst., vol. 8, no. 1, pp. 259_266, 2012.
[51] Y. Xua, X. Qin, Z. Yang, Y. Yang, and K. Li, ``A personalized k-anonymity privacy preserving method," J. Inf. Comput. Sci., vol. 10, no. 1, pp. 139_155, 2013.
[52] S. Yang, L. Lijie, Z. Jianpei, and Y. Jing, ``Method for individualized privacy preservation," Int. J. Secur. Appl., vol. 7, no. 6, p. 109, 2013.
[53] A. Halevy, A. Rajaraman, and J. Ordille, ``Data integration: The teenage years," in Proc. 32nd Int. Conf. Very Large Data Bases (VLDB), 2006, pp. 9_16.
[54] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, ``State-of-the-art in privacy preserving data mining," ACM SIGMOD Rec., vol. 33, no. 1, pp. 50_57, 2004.
[55] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu, ``Tools for privacy preserving distributed data mining," ACM SIGKDD Explorations Newslett., vol. 4, no. 2, pp. 28_34, 2002.
[56] R. Agrawal, T. Imieli«ski, and A. Swami, ``Mining association rules between sets of items in large databases," in Proc. ACM SIGMOD Rec., 1993, vol. 22, no. 2, pp. 207_216.
[57] V. S. Verykios, ``Association rule hiding methods," Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery, vol. 3, no. 1, pp. 28_36, 2013.
[58] K. Sathiyapriya and G. S. Sadasivam, ``A survey on privacy preserving association rule mining," Int. J. Data Mining Knowl. Manage. Process, vol. 3, no. 2, p. 119, 2013.
[59] D. Jain, P. Khatri, R. Soni, and B. K. Chaurasia, ``Hiding sensitive association rules without altering the support of sensitive item(s)," in Proc. 2nd Int. Conf. Adv. Comput. Sci. Inf. Technol. Netw. Commun., 2012, pp. 500_509.
[60] J.-M. Zhu, N. Zhang, and Z.-Y. Li, ``A new privacy preserving association rule mining algorithm based on hybrid partial hiding strategy," Cybern. Inf. Technol., vol. 13, pp. 41_50, Dec. 2013.
[61] H. Q. Le, S. Arch-Int, H. X. Nguyen, and N. Arch-Int, ``Association rule hiding in risk management for retail supply chain collaboration," Comput Ind., vol. 64, no. 7, pp. 776_784, Sep. 2013.