© 2017 IJSRSET | Volume 3 | Issue 2 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Information Security and Data Mining in Big Data

Tejas P. Adhau*1, Prof. Dr. Mahendra A. Pund*2

Department of Computer Science of Engineering/SGBAU University/PRMIT Badnera/Amravati, Maharashtra, India

ABSTRACT

The growing popularity and development of data mining technologies bring serious threat to the security of individual's sensitive information. An emerging research topic in data mining, known as privacy-preserving data mining (PPDM), has been extensively studied in recent years. The basic idea of PPDM is to modify the data in such a way so as to perform data mining algorithms effectively without compromising the security of sensitive information contained in the data. Current studies of PPDM mainly focus on how to reduce the privacy risk brought by data mining operations, while in fact, unwanted disclosure of sensitive information may also happen in the process of data collecting, data publishing, and information (i.e., the data mining results) delivering. In this paper, we view the privacy issues related to data mining from a wider perspective and investigate various approaches that can help to protect sensitive information. In particular, we identify four different types of users involved in data mining applications, namely, data provider, data collector, data miner, and decision maker. For each type of user, we focus on his privacy and how to protect sensitive information.

Keywords: Data Mining, Sensitive Information, Privacy-Preserving Data Mining Provenance, Anonymization , Privacy Auction, Antitracking.

I. INTRODUCTION

Data mining has attracted more and more attention in recent years, probably because of the popularity of the``big data'' concept. Data mining is the process of examining large pre-existing databases in order to generate new information and the result gives direction to guide future activities. Data mining process is also used for the analysis of data for relationships that have not previously been discovered. The term data warehouse is used to store a database that is used for analysis. Warehouse should be able to tell you what type of data they want to view and at what levels relationships among data items they want to be able to view it.

II. METHODS AND MATERIAL

1. The Process of KDD

Generally three of the major data mining techniques are regression, classification and clustering. Data Mining also popularly known as Knowledge Discovery in Databases (KDD) [1] [2]. KDD widely used data mining technique is a process that includes data preparation, selection, and generate result patterns. Some issues involved in the entire KDD process are:

- Identify the goal of the KDD process.
- Understand application domain involved an the knowledge that's required. Select data set on which discovery is be performed.
- Alter the data as per the requirements.
- Simplify the data sets by removing unwanted variables and missing fields
- Match KDD goals with data mining methods to
- suggest hidden patterns. Choose data mining algorithms to discover hidden patterns.
- Search for patterns of interest in a particular representational form, which include classification rules or trees, regression and clustering.
- Interpret essential knowledge from the mined patterns.

- Use the knowledge and incorporate it into another system for further action.

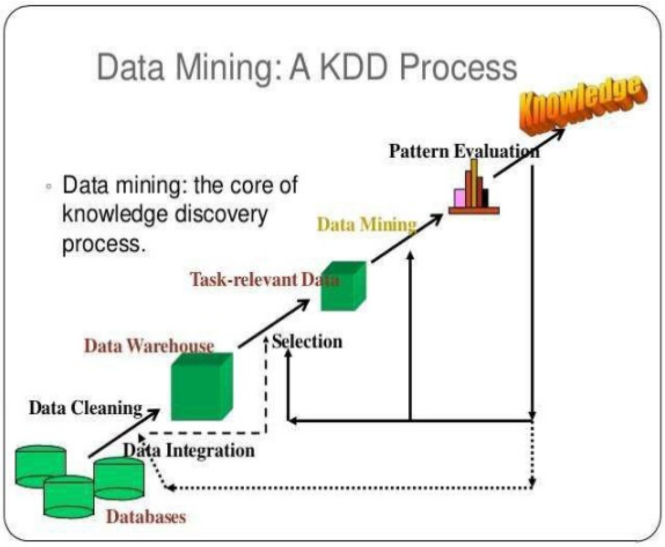


Figure 1. An overview of KDD process

To solve this issue we apply following step are performed in an iterative way.

Data cleaning Data cleansing is also known as data cleaning or data scrubbing. it is a Step in which irrelevant data and noise data are removed from the raw collection of data. Although data cleansing can involve deleting old, incomplete or duplicated data.

Data integration is the combination of analytical and technical processes used to combine data from distinct sources into meaningful and valuable information.

Data selection at this step, the data relevant to the analysis is decided on and retrieved from the data collection.

Data transformation is the process of converting data or information from one format to another, usually from the format of a source system into the required format of a new destination system.

Pattern evaluation and presentation KDD process in which discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and Interpret the data mining results.

The explosive development in KDD process leads to privacy preservation which has been one of the greater concerns in data mining and given rise to a new research field, known as Privacy Preserving Data Mining (PPDM). PPDM mainly focus on the hiding the data in which the sensitive data like person name person identity, phone number, resident address etc., In data hiding process, we alter or block such sensitive information out from the original database, in order to preserve personal sensitive information. On the other hand, the sensitive information is extracted in data mining process. To eliminate such type of sensitive information by using association mining rule algorithm [3]. To achieve the privacy of sensitive data, user should share their sensitive information in encrypted manner with the third party or distributed environment. PPDM is a new emerging research field. Many approaches were been developed in early years.

In the traditional approach, all the sensitive information is hided. But if we see individual concern, the data which is important to one user, here hiding rule effects positively, the same data may not be seen as important to the other user; here hiding rule has negative impact. User information store in centralized and distributed data, based on the distribution of data.In a centralized database (DB) environment, data are all stored in a single database; while, in a distributed database environment, data are stored in different databases [4].

The Traditional PPDM algorithm mainly focuses on classification, association rule and clustering. In general Classification algorithms can be first divided into two step, In the first step classification based on previous data and generate the training data. In the second step, we use training data as a sample data to classify new data. Association analysis involves the discovery of associated rules, showing attribute value and conditions that occur frequently in a given set of data. Clustering Analysis means a collection of database into groups so that the data point in one group are similar to each other and are as different as possible from the data points in other groups.

2. The Privacy Concern and PPDM

With more and more information easily available and easily accessible in electronic forms and those electronic forms present on the web and with the increasing powerful data mining tools are developed and these tools are used in data in data mining process causes a threat to user privacy and data security. In this way, we believe that privacy concerns with unauthorized access to individual data especially focus on sensitive information for example health records, financial records, legal issue records, etc. The goal of PPDM is to protect sensitive information from unwanted or unauthorized access. The PPDM process work on two principals, first, sensitive information should not be directly used for data mining process. Second, sensitive mining results whose disclosure will result in privacy violation should be excluded. In the other words "The Privacy and PPDM deals with obtaining valid data mining results without disclosing the sensitive information data."

3. User Role-Based Approach

Recent models and algorithms in PPDM approach mainly focus on how to hide sensitive information from data mining process. The entire KDD process involves

multi-phase operations. In the data mining process, privacy issues may begin in the data collecting or data preprocessing.

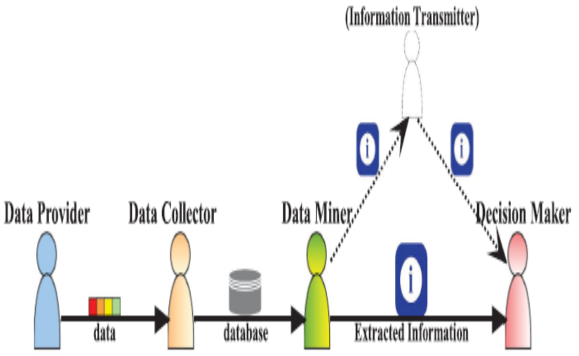


Figure 2. A Simple illustration of the Application Scenario with data mining at core.

User-role based approach to conduct the review of related studies. Based on the multi-phase operations in KDD process. we can identify four different types of users, namely four user roles,

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 goals.

In the data mining process, a user represents either a person or an organization. Also, one user can play multiple roles at once. For example, the U.S.retailer Target once received complaints from a customer who was angry that Target sent coupons for baby clothes to his teenager daughter. However, it was true that the daughter was pregnant at that time, and Target correctly inferred the fact by mining its customer data. In this story, the customer plays the role of data provider, and the retailer plays the roles of data collector, data miner and decision maker[5].

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 and appropriate solutions the

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

4. DATA PROVIDER

The major concern of a data provider is whether he can control the sensitivity of the data he provides to others. On one hand, the provider should be able to make his very private data, namely the data containing information that he does not want anyone else to know, inaccessible to the data collector [6]. On the other hand, 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 compensation for the possible loss in privacy.

5. DATA COLLECTOR

The data collected from data providers may contain individual's sensitive information. Directly releasing the data to the data miner will violate data providers' privacy, hence data modification is required. On the other hand, the data should still be useful after modification; otherwise collecting the data will be meaningless[7].Therefore, the major concern of data collector is to guarantee that the modified data contain no sensitive information but still preserve high utility.

6. 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.

7. DECISION MAKER

A decision maker can get the data mining results directly from the data miner, or from some Information Transmitter. It is likely that the information transmitter changes the mining results intentionally or unintentionally, which may cause serious loss to the decision maker. Therefore, what the decision maker concerns is whether the mining results are credible [8]. In addition to investigate the privacy-protection approaches adopted by each user role, in this paper we emphasize a common type of approach, namely game theoretical approach, that can be applied to many problems involving privacy protection in data mining. The rationality is that, in the data mining scenario, each user pursues high self-interests in terms of privacy preservation or data utility, and the interests of different users are correlated.

8. 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.

8.1 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" [9]. 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.

8.2. APPROACHES TO PRIVACY PROTECTION

8.2.1. LIMIT THE ACCESS.

Data provider can provide his data to the Data collector in active way or passive way.

Active Way: Data provider voluntarily opts in a survey initiated by the Data collector or fill in some registration form to create an account in a website.

Passive Way: Data collector collects the Data provider’s data by the Provider’s routine activities. Data collector retrieves the data by recording provider's routine activities unaware of Data Provider.

Data provider can avoid tracking his routine activities by emptying cache, deleting cookies, clearing usage records of applications, etc. Current security tools that are developed for internet environment to protect provider's data can be categorized into three types:

i. Anti-tracking extensions: Data collector can retrieve user's sensitive data by tracking his/her routine activities. To avoid this unathourized access to the provider's data, provider can user anti-tracking tools such as Disconnect, Do Not Track Me, Ghostery, etc.
ii. Advertisement and script blockers: By adding browser extensions such as AdBlockPlus, NoScript, FlashBlock, etc. user can block advertisements on the sites and kill scripts and widgets that send user's sen-sitive data to unknown third party.
iii. Encryption tools: A user can utilise encryption tools such as MailCloak, TorChat to encrypt mails to make sure that a private communication between two par-ties cannot be intercepted by third parties.

In addition to all these tools, user can use anti-virus, anti-malware tools to protect data. Using such a tools user can limit the access of his/her sensitive data to third parties.

8.2.2. TRADE PRIVACY FOR BENEFIT.

In some cases, provider needs to make tradeoff between the loss of privacy and the benefits brought by participating in Data mining. Consider shopping website. If website tracks user's routine activities and find outs user interested products, then it will be beneficial for user also. User can fill better shopping experience in this case. Now suppose user has to enter information about salary on shopping website, then the website can show the interested item in user's bud-get. so, disclosure of sensitive information such as salary is more beneficial as it reduces the searching time of the user.

8.2.3. PROVIDE FALSE DATA.

Data providers takes efforts to hide sensitive information from data collector. Data collector takes efforts to hide sensitive information from Data miner. But, In today's internet age, internet users cannot completely stop the unwanted access to user's personal information. So, instead of trying to limit the access, the data provider can provide false information to untrustworthy Data collectors. Following methods can be used to falsify the data.

1. sockpuppets: A sockpuppet is a false online identity. By using multiple sockpuppets, the data produced by one individual's activities will be deemed as data belonging to different individuals. Data collector do not have enough knowledge to relate different sockpup-pets with one individual.So, user's true activities are unknown to others and user's sensitive information cannot be easily retrieved.

2. Clone identity:This technique can protect user's privacy by creating fake identity.This clone identity automatically makes some actions which are totally different from user's actions.So if the third party tries to retrieve user's data, then it will get data from clone identity which is completely different.By this way, user's sensitive data is unaccessible to the unwanted user.

3. MaskMe: By adding MaskMe browser extension user can hide his/her sensitive data. Wherever user perform online transaction, user has to enter his sensitive information such as email id, Bank details, etc. Using this extensions, many aliases are created. so, user's sensitive data can be secured.

9. DATA COLLECTOR

9.1. CONCERNS OF DATA COLLECTOR

As shown in Fig.2, Data collector collects the data from Data provider and provide this data to Data Miner. Data collected from Data provider may contain sensitive information of the individual. If such a data is directly send to the Data Miner, then individual's sensitive information disclosed to the unwanted third parties or Data miner. So, before sending data to the Data miner, Data collector has to check whether data contains sensitive information or not. If so then Data collector has to encrypt the data collected from Data provider and then send it to the Data miner. Data collector has to modify the data before releasing it to the Data miner. But, After using modification techniques, there will be loss in data utility. So the main concern of Data miner is that the data must retained utility after the modification. Otherwise collecting data is waste process. The data modification process adopted by Data collector with the goal of preserving privacy and utility simultaneously is called as Privacy Preserving Data Publishing (PPDP)[2].

8.2. APPROACHES TO PRIVACY PROTECTION

3.2.1. BASICS OF PPDP.

The original data is in the form of table with multiple records. Each record consists of four types of attributes.

1. Identifier (ID): Attributes that uniquely identifies user on cloud

2. Quasi-identifier (QID): Attributes that linked with the external data to re identify user.

3. Sensitive Attribute (SA):Attributes that the user wants to hide for privacy.

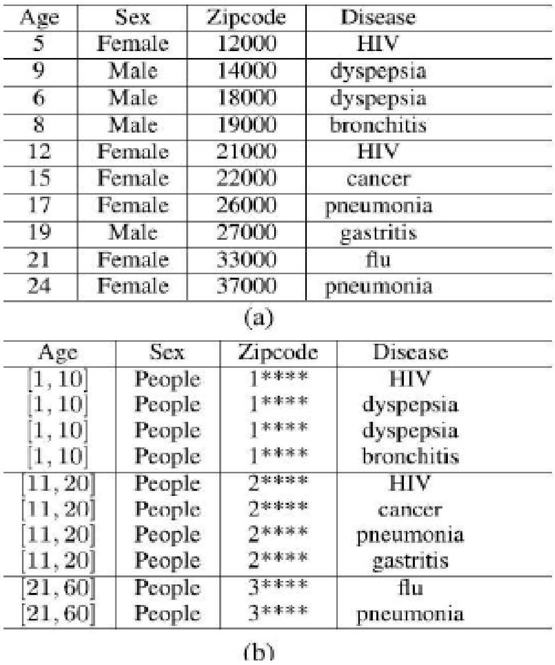
4. Non-Sensitive Attribute (NSA): Attributes that user don't matter to disclose with anyother.

Data anonymization is a type of information sanitization whose intent is privacy protection. It is the process of either encrypting or removing personally identifiable information from data sets so that identity and sensitive attribute values hidden from adversaries. Record linkage (RL) refers to the task of finding records in a data set that refer to the entity across different data sources (e.g., data files, books, websites, databases). In Attribute linkage (AL), the adversary may not precisely identify the record of the target victim, but could infer his/her sensitive values from the published data, based on the sensitive values associated to the group, that the victim belongs to. In Table linkage (TL), the attack seeks to determine the presence or absence of victim's record in the released table. Probabilistic linkage, takes a different approach to the record linkage problem by taking into account a wider range of potential identifiers, computing weighs for each identifier based on it's estimated ability to correctly identify a match or non-match, and using these weighs to calculate probability that two given records refer to the same entity. Different privacy models includes k-anonymity, l-diversity, t-closeness, epsilon-differential privacy. k-anonymity is used for record linkage.

l-diversity is used for preventing record and attribute linkage.

t-closeness is used for preventing attribute and probabilistic linkage epsilon-differential is used for preventing table and probabilistic linkage.

Among all these, k anonymity is widely used.In k-anonymity, attributes are suppressed or generalized until each row is identical with atleast k-1 other rows. Thus it prevents definite database linkages. K-anonymity guarantees that the data released is accurate.



Age	Sex	Zipcode	Disease
5	Female	12000	HIV
9	Male	14000	dyspepsia
6	Male	18000	dyspepsia
8	Male	19000	bronchitis
12	Female	21000	HIV
15	Female	22000	cancer
17	Female	26000	pneumonia
19	Male	27000	gastritis
21	Female	33000	flu
24	Female	37000	pneumonia

(a)

Age	Sex	Zipcode	Disease
[1, 10]	People	1****	HIV
[1, 10]	People	1****	dyspepsia
[1, 10]	People	1****	dyspepsia
[1, 10]	People	1****	bronchitis
[11, 20]	People	2****	HIV
[11, 20]	People	2****	cancer
[11, 20]	People	2****	pneumonia
[11, 20]	People	2****	gastritis
[21, 60]	People	3****	flu
[21, 60]	People	3****	pneumonia

(b)

Figure 3. An example of 2-anonymity where QID = Age,Sex,Zipcode.(a)Original Table (b)2-anonymous table

Consider following table which gives idea about k-anonymity. Now consider above table A and table B which denotes Raw table and anonymized table respectively. Using K-anonymous technique, Data collector can hide Identifiers and Quasi-identifier fields from third parties. As shown in fig.B, quasi-identifier fields such as age, sex zipcode are replaced by either special characters or range values or common attribute. So, by using such anonymous table, adversaries are unable to track particular individual then, the probability that the individual's record being identified by the adversary will not exceed 1/K.

To satisfy privacy model conditions, following operations can be done.

Generalization: Replace some values in the table with parent value in the taxonomy of an attribute.

Suppression: Replace some values in the table with a special character ("*"), as shown in the column "Zip-code" in table.B.

Anatomization: Deassociates the relationship between two.

Permutation: Deassociates the relationship between a quasi-identifier and a numerical sensitive attribute by partitioning a set of data records into groups and shuffling their sensitive values within each group.

Perturbation: Replace original data values with some synthetic data values.

But, All these privacy model information results into information loss.

3.2.2. PRIVACY PRESERVING PUBLISHING OF SOCIAL NETWORK DATA.

Social network data is al-ways represented in the form of graph. where vertex rep-resents an entity and edge represents the relationship between two entities. So, In case of social network PPDP deals with the anonymized graph data. Anonymizing social network data[10] is much more challenging than that of relational data.

It is much more challenging to model background knowledge of adversaries and attacks about social network data than that about relational data. On relational data, it is often assumed that a set of attributes serving as a quasiidentifier is used to associate data from multiple tables, and attacks mainly come from identifying individuals from the quasi-identifier. However, in a social network, many pieces of information can be used to identify individuals, such as labels of vertices and edges, neighborhood graphs, induced subgraphs, and their combinations. It is much more complicated and much more difficult than the relational case. it is much more challenging to measure the information loss in anonymizing social network data than that in anonymizing relational data. Typically, the information loss in an anonymized table can be measured using the sum of information loss in individual tuples. Given one tuple in the original table and the corresponding anonymized tuple in the released table, we can calculate the distance between the two tuples to measure the information loss at the tuple level. How-ever, a social network consists of a set of vertices and a set of edges. It is hard to compare two social networks by comparing the vertices and edges individually. Two social networks having the same number of vertices and the same number of edges may have very different network-wise properties such as con-nectivity, betweenness, and diameter. Thus, there can be many different ways to assess information loss and anonymization quality it is much more challenging to devise anonymization methods for social network data than for relational data. Divide-and-conquer methods are extensively applied to anonymization of relational data due to the fact that tuples in a relational tables are separable in anonymization. In other words, anonymizing a group of tuples does not affect other tuples in the table. However, anonymizing a social network is much more difficult since changing labels of vertices and edges may affect the neighborhoods of other vertices, and removing or adding vertices and edges may affect other vertices and edges as well as the properties of the network.

3.2.3. ATTACK MODEL.

In Anonymized social net-work data, adversaries often rely on the background knowledge to de-anonymize individuals and learn relationships between deanonymized individuals.

”Seed and Grow” algorithm[11] invented by Peng et al. is used to identify users from an anonymized social graph, based solely on graph structure. The seed stage plants a small specially designed sub graph into undirected graph before its release. After anonymized graph is released , the attacker locates sub graph in anonymized graph.so, the vertices are readily identified and serves as the initial seeds to be grown. The grow stage is essentially comprised of a structure based vertex matching, which further identifies vertices adjacent to initial seeds. This is self reinforcing process, in which the seeds grow larger as more vertices are identified. ”Structural Attack”[12] is the attack that de-anonymize social graph data. This attack uses cumulative degree of a vertex. ”Mutual Friend Attack” is deanonymized data based on the number of social common friends of two directly connected individuals. As shown in Fig.4, The anonymization mapping f, is a random, secret map-ping.

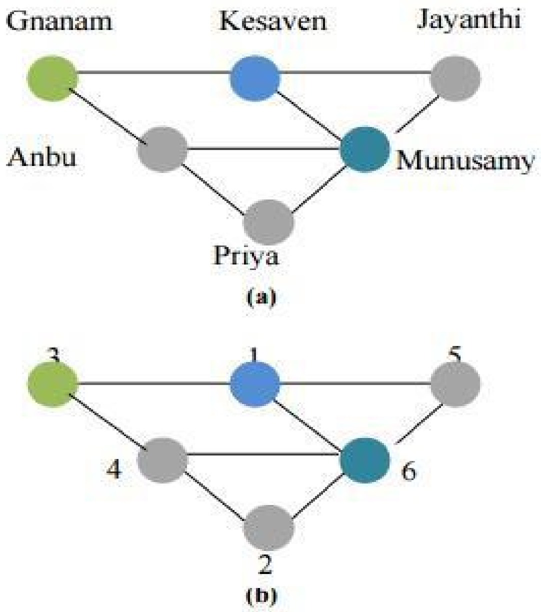
Naive anonymization prevents re-identification when adversary has no information about individual in original graph. in practice the adversary may have access to external information about the entities in the graph and their relationships. This information may be available through a public source be-yond the control of the data owner, or may be obtained by the adversarys malicious actions. For ex-ample, for the graph in Figure 1, the adversary might know that Munusamy has three or more neighbors, or that Gnanam is connected to at least two nodes, each with degree 2. Such information allows the adversary to reduce the set of candidates in the anonymized graph for each of the targeted individuals. Although an adversary may also have information about the at-tributes of nodes, the focus of this paper is structural re-identification, where the adversarys information is about graph structure. Reidentification with at-tribute knowledge has been well studied, as have techniques for resisting it. More importantly, many net-work analyses are concerned exclusively with structural properties of the graph; therefore safely publishing an unlabeled network is a legitimate goal.

1. PRIVACY MODEL No. of privacy models are pro-posed for graph data based on classic k-anonymity model.[9]

(a) k-NMF anonymity: It protects the privacy of relationship from the mutual friend attack.

(b) K2-degree anonymity: It protects information loss due to friendship attack.

c) k-structural diversity anonymization (k-SDA): It protects information loss due to degree attack.



Gnanam	3
Kesaven	1
Jayanthi	5
Anbu	4
Munusamy	6
Priya	2

(c)

Figure 4. Example of mutual friend attack: (a)Original network; (b)Naive anonymized net-work. (c)Mapping Function (f)

9.2.4. PRIVACY PUBLISING PRESERVING OF TRAJECTORY DATA.

In recent years, LBS(Location Based Services)[9] becomes very popular. Using these services user can able to find out interesting places near him/her. If he/she wants information about nearest bank. then he/she can use such a services and able to find out nearest bank location. To provide location-based services, commercial entities (e.g. a telecommunication company) and public entities (e.g. a transportation company) collect large amount of individuals' trajectory data, i.e. sequences of consecutive location readings along with time stamps. If the data collector publish such spatio-temporal data to a third party (e.g. a data-mining company), sensitive infor-mation about individuals may be disclosed. To realize a privacy-preserving publication, anonymization techniques can be applied to the trajectory data set, so that no sensitive location can be linked to a specific individual.

III. RESULTS AND DISCUSSION

1. DATA MINER

1.1. CONCERNS OF DATA MINER

Data collector sends the data after modification to the Data miner. Then, the Data miner has to retrieve the important data using different data mining techniques. so,The primary concern of data miner is how to prevent sensitive information from appearing in the mining results. To per-form 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. Similar to data collector, the data miner also faces the privacy utility tradeoff problem. But in the context of PPDM, quantifications of privacy and utility are closely related to the mining algorithm employed by the data miner.

1.2. APPROACHES TO PRIVACY PROTECTION

Privacy preserving data mining approaches are classified into two main categories i.e. Approaches for centralized data mining and Approaches for Distributed data mining. Distributed Data mining again further classified as horizontally partitioned data and vertically partitioned data as shown in Fig.5. Now, most services are using distributed data mining where secure multi-party computation is used. SMS (Secure Multi-party

Computation) is a subfield of cryptography.SMC assumes that there are number of participants P1; P2; P3; ::::::; Pm with having private data D1; D2; D3; :::::; Dm respectively. The participants want to compute.The value of public function f. We can say that SMC protocal is secure if, at the end of computation, par ticipant can able to view only their own data. So, the main goal of SMC protocol is to find correct data mining results without revealing participants data with others.

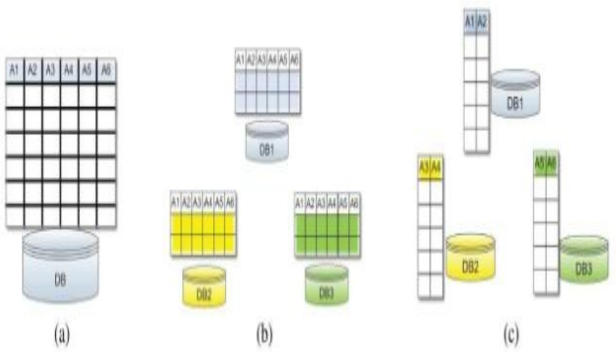


Figure 5. Data Distribution (a)Centralized Data (b)Horizontally Partitioned Data (c)Vertically Partitioned Data**.**

1.2.1. PRIVACY-PRESERVING-ASSOCIATION RULE MINING.

Association rule mining is a two-step process:

(1) Finding all frequent itemsets;
(2) Generating strong association rules from the frequent itemsets.

The purpose of privacy preserving is to discover accurate patterns to achieve specific task without precise access to the original data. The algorithm of association rule mining is to mine the association rule based on the given minimal support and minimal confidence. Therefore, the most direct method to hide association rule is to reduce the support or confidence of the association rule below the minimal support of minimal confidence. With regard to association rule mining, the proposed methodology that is effective at hiding sensitive rules is implemented mainly by depressing the support and confidence.

Various kinds of approaches have been proposed to perform association rule hiding.These approaches are clas-sified as five categorize. Heuristic Distortion Approaches: selects appropri-ate data sets for data modification. Heuristic blocking approaches: reduces the degree of support and confidence of the sensitive association rules by replacing certain attributes of some data item with specific symbol. Probabilistic distortion approaches: distorts the data through random numbers generated from predefined probability distortion function. Exact database distortion approaches: formulates the solution of the hiding problem as a constraint satisfaction problem (CSP), and apply linear programming approaches to it's solution.

Reconstruction-based approaches: generates a database from the scratch that is compatible with a given set of non-sensitive association rules.

1.2.2. PRIVACYPRESERVING CLASSIFICATION

Data classification is a two step process.

1. Step1: Learning Step: algorithm generate classification model.
2. Step2: Classification: Develops different classification models such as Decision Tree, Bayesian Model, Support Vector Machine (SMC), etc.

DECISION TREE

A decision tree[12] is defined as a predictive modeling technique from the field of machine learning and statistics that builds a simple tree-like structure to model the underlying pattern of data. Decision tree is one of the popular methods to classify is able to handle both categorical and numerical data and per-form classification with minimal computation. Decision trees are often easier to understand and compute. Decision tree is a classifier which is a directed tree with a node having no incoming edges called root. All the nodes except root have exactly one incoming edge. Each non-leaf node called internal node or splitting node contains a decision and most appropriate target value assigned to one class is represented by leaf node. Decision tree classifier is able to divide a complex process into number of simpler processes. The complex decision is sub divided into simpler decision on the basis of splitting of complex process into simple processes. It divides com[plete data set into smaller subsets. Information gain, gain ratio, gini index are three basic splitting criteria to select at-tribute as a splitting point.

Decision trees can be built from historical data they are often used for explanatory analysis as well as a form of supervision learn-ng. The algorithm is designed in such a way that it works on all the data that is available and as perfect as possible. According to Breiman et al. the tree complexity has a crucial effect on its accuracy performance. The tree complexity is explicitly controlled by the pruning method employed and the stopping criteria used. Usually, the tree complexity is measured by one of the following metrics:

- The total number of nodes Total number of leaves
- Tree depth
- Number of attributes used

Decision tree induction is closely related to rule induction. Each path from the root of a decision tree to one of its leaves can be transformed into a rule simply by conjoining the tests along the path to form the antecedent part, and taking the leafs class prediction as the class value. The resulting rule set can then be simplified to improve its accuracy and comprehensibility to a human user

2. NAIVE BAYESIAN CLASSIFICATION

The Naive bayesian classifier[9] is a simple but efficient baseline classifier. This classifier used for text classification. Naive bayesian is based on a bayesian formulation of the classification problem which uses the simplifying assumption of attribute independence. It is simple to compute and computation calculates good results. Thus, preliminary evaluation is carried out using the Naive Bayesian classifier to serve both as a baseline and to decide whether more sophisticated solutions are required. The problem of secure distributed classification is an important one. The goal is to have a simple, efficient, easy to compute and privacy-preserving classifier. The ideal would be for all parties to decide on a model. Jointly select/discover the appropriate parameters for the model and then use the model locally as and when necessary. We discuss the specifics in the context of the Naive Bayesian classifier later. in this, data is assumed to be horizontally partitioned. This means that many par-ties collect the same set of information about different entities. Parties want to improve classification accuracy as much as possible by leveraging other parties data. They do not want to reveal their own instances or the instance to be classified. Thus, what we have is a collaboration for their own advantage. One way to solve this is to decide on a model. The model parameters are generated jointly from the local data. Classification is performed individually without involving the other parties. Thus, the parties decide on sharing the model, but not the training set nor the instance to be classified. This is quite realistic. For example, consider banks which decide to leverage all data to identify fraudulent credit card usage, or insurance companies which jointly try to identify high-risk customers. In this paper, we use / extend several existing cryptographic techniques to create a privacy preserving Naive Bayesian Classifier for horizontally partitioned data.

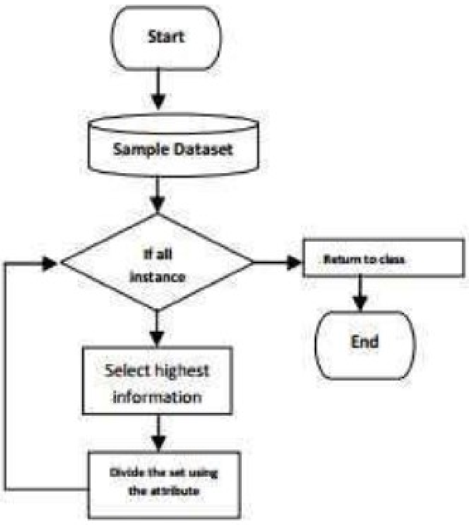

Start
Sample Dataset
If all instance
Return to class
End
Select highest information
Divide the set using the attribute

Figure 6. Flowchart for Decision Tree Based Classification.

3. SUPPORT VECTOR MACHINE

Support vector Machine (SVM)[4] is one of the most developed classification methodology in data mining. It provides properties such as the margin maximiza-tion and nonlinear classification via kernel tricks and has proven to be effective in many real world applica-tions.Privacy preserving SVM classification solution, PP-SVM which constructs the global SVM classifica-tion model from the data distributed from a multiple parties.The data may be partitioned horizontally, ver-tically or in an arbitry manner between the parties. The data of each party is kept private, while the fi-nal model is constructed at an independent site, This independent site then performs classification of new instances. of sense in many different contexts. For example, consider a clearing house for a consortium of banks. The different banks collect data of their customers. The features collected such as age, gen-der, balance, average monthly income, etc. are the same for all ban k.Thus,the data is horizontally distributed. the clearing house is an independent en-tity, unrelated to any of the banks. The classification model is constructed at the clearing house while pre-serving the privacy of the individual data from each of the banks. When a bank has a new instance it wants to classify, it goes through a secure protocol with the clearing house to classify just this instance. The clear-ing house learns nothing. This would allow all of the banks to leverage the global data without compromis-ing on privacy at all.

3. DECISION MAKER

3.1. CONCERNS OF DECISION MAKER

The final goal of data mining process is to provide useful information to the decision maker, so that the decision maker can choose a result which is better way to achieve his objective. As we can see, Data provider sends data to Data collector, Data collector sends data to the Data miner and finally Data miner sends data to the Decision Maker. So, we can say that Decision maker is less respon-sible for the data security. The data mining results pro-vided 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 suf-fer 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 skeptical about the credibility of the results, in case that the results have been distorted.

Therefore, the privacy concerns of the decision maker are twofold: how to prevent unwanted disclosure of sensitive mining results, and how to evaluate the credibility of the received mining results.

3.2. APPROACHES TO PRIVACY PROTECTION

3.2.1. DATA PROVENANCE.

Usually, Decision maker receives data from data miner but, in some cases if the decision maker does not get the data mining results directly from the data miner i.e. receives data from other sources ,then he wants to know how the results are delivered to him and what kind of modifications are applied to the results, so that he can decide whether the results are trusted or not. This is why ”provenance” is needed. The term provenance [originally refers to the custody of the data. In computer science, data provenance refers to the information that helps determine the derivation history of the data, starting from the original source. Two kinds of information can be found in the provenance of the data: the ancestral data from which current data evolved, and the transformations applied to ancestral data that helped to produce current data. With such information, people can better understand the data and judge the credibility of the data. data provenance has been extensively studied in the fields of databases and work flows. Several surveys are now available. The following five aspects are used to capture the characteristics of a provenance system:

1. Application of provenance. Provenance systems may be applied in many fields to support a number of uses, such as estimate data quality and data reliability, trace the audit trail of data, repeat the derivation of data, etc.
2. Subject of provenance. Provenance information can be collected about different sources and at various levels of detail.
3. Representation of provenance. There are mainly two types of methods to represent provenance in-formation, one is annotation and the other is inversion. The annotation uses metadata. Using inversion method, derivations are inverted to find out inputs to the derivations.
4. Provenance storage. Provenance is tightly coupled with the data it describes and located in the same data storage system or even be embedded within the data. Alternatively, provenance can be stored separately with other metadata or simply by itself.
5. Provenance dissemination. A provenance system can use different ways to provide the provenance information, such as providing a derivation graph that users can browse and inspect.

3.2.2. WEB INFORMATION CREDIBILITY.

Because of the lack of publishing barriers, the low cost of dissemination, and the lax control of quality, credibility of web information has become a serious issue. Tudjman et al. identify the following five criteria that can be employed by Internet users to differentiate false information from the truth:

1. Authority: the real author of false information is usually unclear.
2. Accuracy: false information dose not contain accurate data or approved facts.
3. Objectivity: false information is often prejudicial.
4. Currency: for false information, the data about its source, time and place of its origin is incomplete, out of date, or missing.
5. Coverage: false information usually contains no effective links to other information online.

IV. CONCLUSION

In this paper, We discuss about security concerns and privacy preserving techniques of each user such as Data Provider, Data Collector, Data Miner and Decision Maker. For Data Provider, can secure his data by three ways: he can limit the access to his online activities or data by using anti-tracking extensions, advertisement and script blockers or by using encryp-tion tools to encrypt emails between two private par-ties.Data Provider can also demand for high price to disclose his private data with others.Nowadays, whatever you try, but the hackers can get your se-cure information.So, Data provider can provide false data to misguide such a hackers.Using sockpuppet, Data provider can make different sockpuppets. Data provider can use MaskMe to mask his sensitive infor-mation For Data Collector, He receives data from Data provider and sends that data to the Date Miner. Be-fore sending this data to the Data miner, Data col-lector has to check whether data contains any private information or not. data Collector has to develop dif-ferent attack models to check whether data contains any private information about data provider

For Data Miner, He has to retrieve the important data using different data mining techniques. so,The pri-mary concern of data miner is how to prevent sen-sitive information from appearing in the mining re-sults. 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.

Similar to data collec-tor, the data miner also faces the privacy utility trade-off problem.By using different algorithm techniques such as Decision tree, Support Vector Machine, Naive Bayesian Techniques Data collector modifies data and sends it to the Decision maker.

For Decision Maker, the privacy concerns are twofold: how to prevent unwanted disclosure of sensitive mining results.

V. REFERENCES

[1]. L. BrankovicRE and V. Estivill-Castro, "Privacy issues in knowledge discovery and data mining," in Proc. Austral. Inst.Comput. Ethics Conf., 1999, pp. 89–99.

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

[3]. L. Sweeney, "k-anonymity: A model for protecting privacy," Int. J. Uncertainty, Fuzziness Knowl.-Based Syst., vol. 10, no. 5,pp. 557–570, 2002.

[4]. Verizon Communications Inc. (2013). 2013 Data Breach Investigations Report. Online]. Available: http://www.verizonenterprise.com/resources/reports/rpdatabreachinvestigations-report-2013_en_xg.pdf

[5]. Lei Xu; Chunxiao Jiang; Jian Wang; Jian Yuan; Yong Ren, "Information Security in Big Data: Privacy and Data Mining,"Access, IEEE , vol.2, no., pp.1149,1176, 2014

[6]. 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.

[7]. 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.

[8]. R. Gibbons, A Primer in Game Theory. Hertfordshire, U.K.: Harvester Wheatsheaf, 1992.

[9]. 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.

[10]. 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.

[11]. 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.

[12]. E. Rasmusen, Games and Information: An Introduction to Game Theory, vol. 2. Cambridge, MA, USA: Blackwell, 1994.

[13]. 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.

[14]. D. C. Parkes, "Iterative combinatorial auctions: Achieving economic and computational efficiency," Ph.D. dissertation,Univ. Pennsylvania, Philadelphia, PA, USA, 2001.