International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) e-ISSN: 2395-5325
Volume 3, Issue 11 (November '2017)

SECURITY USING PRIVACY-PRESERVING DATA MINING

M.Sathiyapriya,
M.Phil Research Scholar,
P.G & Research Department of Computer Science,
Siri PSG Arts and Science College for Women,
Sankari,Tamilnadu,India.

K.Sudha,
Assistant Professor,
P.G & Research Department of Computer Science,
Siri PSG Arts and Science College for Women,
Sankari,Tamilnadu,India.

Abstract: Privacy preserving data mining (PPDM) is one of the newest trends in privacy and security research. It is driven by one of the major policy issues of the information era - *the right to privacy*. Although this research field is very new, have already seen great interests in it, the recent proliferation of PPDM techniques is evident; the interest from academia and industry has grown quickly; and separate workshops and conferences devoted to this topic have emerged in the last few years.

Keywords: Data mining, Privacy Preserving , sanitization process, Round Robin, Association Rules Mining

I.INTRODUCTION

Analyzing what right to privacy means is a fraud with problems, such as the exact definition of privacy, whether it constitutes a fundamental right, and whether people are and/or should be concerned with it. Several definitions of privacy have been given, and they vary according to context, culture, and environment. In general, privacy is viewed as a social and cultural concept. However, with the ubiquity of computers and the emergence of the Web, privacy has also become a digital problem [1]. With the Web revolution and the emergence of data mining, privacy concerns have posed technical challenges fundamentally different from those that occurred before the information era. In the information technology era, privacy refers to the right of users to conceal their personal information and have some degree of control over the use of any personal information disclosed to others [2].Clearly, the concept of privacy is often more complex than realized. In particular, in data mining, the definition of privacy preservation is still unclear, and there is very little literature related to this topic. A notable exception is the work presented in [3], in which PPDM is defined as getting valid data mining results without learning the underlying data values." However, at this point, each existing PPDM technique has its own privacy definition. Our primary concern about PPDM is that mining algorithms are analyzed for the side effects they incur in data privacy. Therefore, our definition for PPDM is close to those definitions in [4] - *PPDM encompasses the dual goal of meeting privacy requirements and providing valid data mining results*. Our definition emphasizes the dilemma of balancing privacy preservation and knowledge disclosure.

Privacy Violation in Data Mining : Understanding privacy in data mining requires understanding how privacy can be violated and the possible means for preventing privacy violation. In general, one major factor contributes to privacy violation in data mining: *the misuse of data*. Users' privacy can be violated in different ways and with different intentions. Although data mining can be extremely valuable in many applications (e.g., business, medical an analysis, etc), it can also, in the absence of adequate safeguards, violate informational privacy. Privacy can be violated if personal data are used for other purposes subsequent to the original transaction between an individual and an organization when the information was collected.

One of the sources of privacy violation is called data magnets [1]. Data magnets are techniques and tools used to collect personal data. Examples of data magnets include explicitly collecting information through on-line registration, identifying users through IP addresses, software downloads that require registration, and indirectly collecting information for secondary usage. In many cases, users may or may not be aware that information is being collected or do not know how that information is collected [5]. Worse is the privacy invasion occasioned by secondary usage of data when individuals are unaware of behind the scenes" uses of data mining techniques [9]. In particular, collected personal data can be used for secondary usage largely beyond the users' control and privacy laws. This scenario has led to an uncontrollable privacy violation not because of data mining itself, but fundamentally because of the misuse of data.

In general, privacy preservation occurs in two major dimensions: users' personal information and information concerning their collective activity. Refer to the former as *individual privacy preservation* and the latter as *collective privacy preservation*, which is related to corporate privacy in [3].

Individual privacy preservation: The primary goal of data privacy is the protection of personally identity able information. In general, information is considered personally identify able if it can be linked, directly or indirectly, to an individual person. Thus, when personal data are subjected to mining, the attribute values associated with individuals are private and must be protected from disclosure. Miners are then able to learn from global models rather than from the characteristics of a particular individual.

Collective privacy preservation: Protecting personal data may not be enough. Sometimes, may need to protect against learning sensitive knowledge representing the activities of a group. Refer to the protection of sensitive knowledge as collective privacy preservation. The goal here is quite similar to that one for statistical databases, in which security

control mechanisms provide aggregate information about groups (population) and, at the same time, should prevent disclosure of confidential information about individuals. However, unlike as is the case for statistical databases, another objective of collective privacy preservation is to preserve (hide) strategic patterns that are paramount for strategic decisions, rather than minimizing the distortion of all statistics (e.g., bias and precision). In other words, the goal here is not only to protect personally identify able information but also some patterns and trends that are not supposed to be discovered.

II.CHARACTERIZING SCENARIOS IN PPDM

Before describe the general parameters for characterizing scenarios in PPDM, let us consider two real-life motivating examples where PPDM poses different constraints:

Scenario 1: A hospital shares some data for research purposes (e.g., concerning groupof patients who have a similar disease). The hospital's security administrator may suppress some identifiers (e.g., name, address, phone number, etc) from patient records to meet privacy requirements. However, the released data may not be fully protected. A patient record may contain other information that can be linked with other datasets to re-identify individuals or entities [6]. How can identify groups of patients with similar disease without revealing the values of the attributes associated with them?

Scenario 2: Two or more companies have a very large dataset of records on their customers' buying activities. These companies decide to cooperatively conduct association rule mining on their datasets for their mutual benefit since this collaboration brings the man advantage over other competitors. However, some of these companies may not want to share some strategic patterns hidden within their own data (also called sensitive association rules) with the other parties. They would like to transform their data in such a way that these sensitive association rules cannot be discovered but others can be. Is it possible for these companies to benefit from such collaboration by sharing their data while preserving some sensitive association rules?

Note that the above scenarios describe different privacy preservation problems. Each scenario poses a set of challenges. For instance, scenario 1 is a typical example of individual's privacy preservation, while scenario 2 refers to collective privacy preservation. How can characterize scenarios in PPDM? One alternative is to describe them in terms of general parameters. In [7], some parameters are suggested as follows:

Outcome: Refers to the desired data mining results. For instance, someone may look for association rules identifying relationships among attributes, or relationships among customers' buying behaviors as in scenario 2, or may even want to cluster data as in scenario 1.

Data Distribution: How are the data available for mining: are they centralized or distributed across many sites? In the case of data distributed throughout many sites, are the entities described with the same schema in all sites (horizontal partitions), or do different sites contain different attributes for one entity (vertical partitions)?

Privacy Preservation: What are the privacy preservation requirements? If the concern is solely that values associated with an individual entity not be released (e.g., personal information), techniques must focus on protecting such information. In other cases, the notion of what constitutes \sensitive knowledge" may not be known in advance. This would lead to human evaluation of the intermediate results before making the data available for mining.

III. LITERATURE REVIEW

R. Agrawal, *et al.* [8] In this paper requirements and challenges of data mining are studied such as handling of different types of data, efficiency and scalability of data mining algorithms, usefulness, certainty, and expressiveness of data mining results, expression of various kinds of data mining requests and results, interactive mining knowledge at multiple abstraction levels, mining information from different sources of data, protection of privacy and data security. A comprehensive overview of recently developed data mining techniques by considering the requirements and challenges of data mining is studied to understand the user behavior, to improve the services, and to increase the business opportunities. A classification of the available data mining techniques and a comparative study of each technique are also discussed.

J. Kiernan, *et al.* [9] An overview of tasks involved in knowledge discovery system and the approaches to solve these tasks are provided by the authors and they also described the software tools which are available to use for knowledge discovery tasks and also proposed a feature classification scheme which can be used to study knowledge and data mining software tools. Based on the general characteristics, database connectivity and characteristics of data mining, software tools are classified. They further investigated software products in which some are research prototypes and some are commercial packages. From their analysis, they specify features which should exist in knowledge discovery software in order to accommodate its novice users as well as experienced analysts effectively, also discussed the issues which are not addressed or not solved yet.

V. Krishnan, *et al.* [10] proposed classification of privacy preserving data mining techniques based on different dimensions such as data distribution, data modification, data mining algorithm, data or rule hiding, privacy preservation. They also discussed various methods exist in each classification of methodology based on the dimension. The existing methodologies are discussed for different privacy preserving data mining techniques such as classification, association rule mining and clustering in various dimensions. They evaluated the algorithms related to heuristic-based techniques, cryptography-based techniques, and reconstruction-based techniques for different data mining techniques.

M. P. Armstrong, *et al.* [11] The state of the art in the area of privacy preserving data mining (PPDM) techniques is discussed by the authors. The authors presented the classification of privacy preserving techniques based on the five dimensions such as data distribution, data modification, data mining algorithm, data or rule hiding, privacy preservation. They also discussed the methodologies based on heuristic for classification, association rule mining and clustering techniques and also cryptography based techniques for vertically partitioned and horizontally partitioned databases in multi distributed environment for association rule mining and classification technique. Privacy preserving clustering problem's solution is discussed in this paper with expectation-maximization algorithm. They also studied Reconstruction-Based Techniques for Binary and Categorical Data.

Oliveira, *et al.* [12], proposed a heuristic-based framework for preserving privacy in mining frequent item sets. They focus on hiding a set of frequent patterns, containing highly sensitive knowledge. They propose a set of sanitized algorithms that only remove information from a transactional database, also known in the statistical disclosure control area as non-perturbative algorithms, unlike those algorithms, that modify the existing information by inserting noise into the data, referred to as perturbative algorithms. The first parameter is evaluated in terms of: Hiding Failure (ie) the percentage of restrictive patterns that are discovered from the sanitized database; Misses Cost (ie) the percentage of non-restrictive patterns that are hidden after the sanitization process; Artifactual Pattern, measured in terms of the percentage of discovered patterns that are artifacts.

IV.THE FRAMEWORK FOR PRIVACY-PRESERVING ASSOCIATION RULE MINING

In this section, introduce the framework to address privacy preservation in association rule mining. As depicted in Figure 1, the framework encompasses an inverted file to speed up the sanitization process, a library of sanitizing algorithms used for hiding sensitive association rules from the database, and a set of metrics to quantify not only how much private information is disclosed, but also the impact of the sanitizing algorithms on the transformed database and on valid mining results.

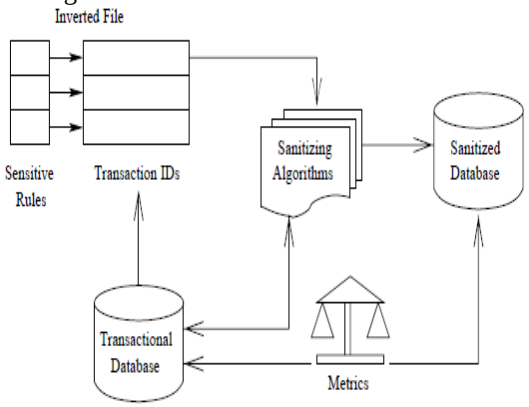


Figure 1: The sketch of the framework for privacy-preserving association rule mining.

The Inverted File: Sanitizing a transactional database consists of identifying the sensitive transactions and adjusting them. To speed up this process, scan a transactional database only once and, at the same time, build our retrieval facility (inverted file) [13]. The inverted file's vocabulary is composed of all the sensitive rules to be hidden, and for each sensitive rule there is a corresponding list of transaction IDs in which the rule is present. Figure 2(b) shows an example of an inverted file corresponding to the sample transactional database shown in Figure 2(a). For this example, assume that the sensitive rules are A,B → D and A,C → D.

TID	Items
T1	A B C D
T2	A B C
T3	A B D
T4	A C D
T5	A B C
T6	B D

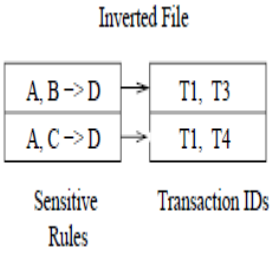


Figure 2: (a) A sample transactional database. (b) The corresponding inverted file.

Note that once the inverted file is built, a data owner will sanitize only the sensitive transactions who's IDs are stored in the inverted file. Knowing the sensitive transactions prevents a data owner from performing multiple scans in the transactional database. Consequently, the CPU time for the sanitization process is optimized. Apart from optimizing the CPU time, the inverted file provides other advantages, as follows:

- The information kept in main memory is greatly reduced since only the sensitive rules are stored in memory. The occurrences (transaction IDs) can be stored on disk when not fitted in main memory.
- Our algorithms require at most two scans regardless of the number of sensitive rules to be hidden: one scan to build the inverted file, and the other to sanitize the sensitive transactions. The previous methods require as many scans as there are rules to hide.

V. THE LIBRARY OF SANITIZING ALGORITHMS

In our framework, the sanitizing algorithms modify some transactions to hide sensitive rules based on a disclosure threshold controlled by the database owner. This threshold indirectly controls the balance between knowledge disclosure and knowledge protection by controlling the proportion of transactions to be sanitized. For instance, if ψ= 50% then half of the sensitive transactions will be sanitized, when Ψ = 0% all the sensitive transaction will be sanitized, and when ψ = 100% no sensitive transaction will be sanitized. In other words, represents the ratio of sensitive transactions that should be left untouched. The advantage of this threshold is that it enables a compromise between hiding association rules while missing non-sensitive ones, and finding all non-sensitive association rules but uncovering sensitive ones.

As can be seen in Figure 1, the sanitizing algorithms are applied to the original database to produce the sanitized one. Classify our algorithms into two major groups: *data sharing-based algorithms* and *pattern sharing-based algorithms*, as can be seen in Figure 3.

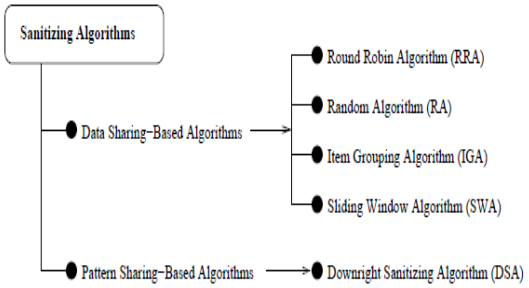
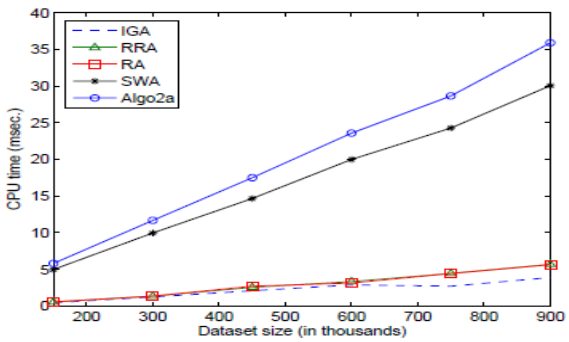


Figure 3: A taxonomy of sanitizing algorithms.

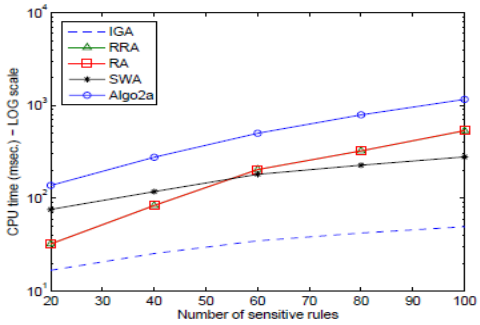
In the former, the sanitization process acts on the data to remove or hide the group of sensitive association rules representing the sensitive knowledge. To accomplish this, small numbers of transactions that participate in the generation of the sensitive rules have to be modified by deleting one or more items from them. In doing so, the algorithms hide sensitive rules by reducing either their support or confidence below a privacy threshold (disclosure threshold). In the latter, the sanitizing algorithm acts on the rules mined from database, instead of the data itself. The algorithm removes all sensitive rules before the sharing process. In Section 5.3, introduce our data sharing-based sanitizing algorithms, and in Section 5.4 present our pattern sharing-based sanitizing algorithms.

VI.CPU TIME FOR THE SANITIZATION PROCESS

Tested the scalability of the sanitization algorithms vis-a-vis the size of the database as well as the number of rules to hide. To do so, selected the Kosarak dataset since it is the largest one used in our experiments. Our comparison study also includes the algorithm Algo2a. Varied the size of the original database *D* from 150K transactions to 900K transactions, while fixing the disclosure threshold 0% and keeping the set of sensitive rules constant (6 original sensitive rules that are mutually exclusive). Figure 4(a) shows that our algorithms scale well with the database size. The algorithms IGA, RRA and RA yielded lower CPU time than that for SWA and Algo2a.



(a)



(b)

Figure 4: Results of CPU time for the sanitization process.

In particular, Algo2a requires six scans over the original database (one to hide each sensitive rule), while the algorithms IGA, RRA and RA require only two. Although the algorithm SWA requires only one scan, it performs many operations in memory (e.g., sorting transactions in ascending order of size for each window), which demands more CPU time as the dataset increases. Even though IGA, RRA, and RA require two scans, they are faster than SWA. The main reason is that these algorithms perform a sort in memory only once.

As can be observed, the algorithms IGA, RRA, and RA increase CPU linearly, even though their complexity in main memory is not linear. If increase the number of sensitive rules or even if selects a group of sensitive rules with very high support, these algorithms may not scale linearly. However, there is no compelling need for sanitization to be a fast operation since it can be done offline. The I/O time (scans over the dataset) is also considered in these figures. This demonstrates good scalability with the cardinality of the transactional database. Also varied the number of sensitive rules to hide from approximately 20 to 100 selected randomly, while fixing the size of the dataset Kosarak and fixing the support and disclosure thresholds to 0%. Figure 4(b) shows that our algorithms scale well with the number of rules to hide. The values are plotted in logarithmic scale because the algorithm Algo2a requires one scan for each rule to hide. Although IGA requires 2 scans, it was faster than SWA in all the cases.

VII.CONCLUSION

The main reason is that the SWA performs a number of operations in main memory to fully sanitize a database. The IGA requires one scan to build an inverted index where the vocabulary contains the sensitive rules and the occurrences contain the transaction IDs. In the second scan, IGA sanitizes only the transactions marked in the occurrences. Another interesting result observed was that over 40 rules, the SWA performed better than the algorithms RRA and RA. The reason is that the heuristic behind the SWA is optimized especially when there are rules with the intersection of items. Note that when the number of sensitive rules increases, the intersection of items among the rules tends to increase as well. In this case, the SWA touches fewer transactions than RRA and RA. As a result, SWA improves the performance as the number of rules to

hide increases since the number of sorts in memory is the same (one by window size) for the dataset.

VIII. REFERENCE

[1]. T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets. Using Association Rules for Product Assortment Decisions: A Case Study. In *Knowledge Discovery and Data Mining*, pages 254{260, 1999.

[2]. S. Cockcroft and P. Clutterbuck. Attitudes Towards Information Privacy. In *Proc. of the 12th Australasian Conference on Information Systems*, Co_s Harbour, NSW, Australia, December 2001.

[3]. C. Clifton, M. Kantarcio_glu, and J. Vaidya. De_ning Privacy For Data Mining. In *Proc. of the National Science Foundation Workshop on Next Generation Data Mining*, pages 126{133, Baltimore, MD, USA, November 2002.

[4]. M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. Verykios. Disclosure Limitation of Sensitive Rules. In *Proc. of IEEE Knowledge and Data Engineering Workshop*, pages 45{52, Chicago, Illinois, November 1999.

[5]. M. J. Culnan. How Did They Get My Name?: An Exploratory Investigation of Consumer Attitudes Toward Secondary Information. *MIS Quartely*, 17(3):341{363, September 1993.

[6]. S. Castano, M. Fugini, G. Martella, and P. Samarati. *Database Security*. Addison- Wesley Longman Limited, England, 1995.

[7]. C. Clifton, M. Kantarcio_glu, J. Vaidya, X. Lin, and M. Y. Zhu. Tools For Privacy Preserving Distributed Data Mining. *SIGKDD Explorations*, 4(2):28{34, December 2002.

[8]. R. Agrawal, T. Imielinski, and A. N. Swami. Mining Association Rules Between Sets of Items in Large Databases. In *Proc. of ACM SIGMOD International Conference on Management of Data*, pages 207{216, Washington, D.C., May 1993.

[9]. R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu. Hippocratic Databases. In *Proc. Of the 28th Conference on Very Large Data Bases*, Hong Kong, China, August 2002.

[10]. R. Agrawal and R. Srikant. Privacy-Preserving Data Mining. In *Proc. of the 2000 ACM SIGMOD International Conference on Management of Data*, pages 439{450, Dallas, Texas, May 2000.

[11]. M. P. Armstrong, G. Rushton, and D. L. Zimmerman. Geographically Masking Health Data to Preserve Con_dentiality. *Statistics in Medicine*, 18:497{525, 1999.

[12]. S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic Itemset Counting and Implication Rules for Market Basket Data. In *Proc. of ACM SIGMOD International Conference on Management of Data*, pages 255{264, Tucson, Arizona, USA, May 1997.

[13]. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Limited, England, 1999.