ISSN:0975-9646

Privacy Preserving Data Mining Using Sanitizing Algorithm

[1] R.Hemalatha [2] M.Elamparithi

[1]*Research Scholar, Department of Computer Science*
Sree Saraswathi Thyagaraja College, Pollachi.

[2]*Assistant Professor, Department of Computer Application,*
Sree Saraswathi Thyagaraja College, Pollachi.

Abstract – The sharing of data is often beneficial in data mining applications. It has been proven useful to support both decision-making processes and to promote social goals. However, the sharing of data has also raised a number of ethical issues. Some such issues include those of privacy, data security, and intellectual property rights. We focus primarily on privacy issues in data mining, notably when data are shared before mining. Specifically, we consider some scenarios in which applications of association rule mining and data clustering require privacy safeguards. Addressing privacy preservation in such scenarios is complex. One must not only meet privacy requirements but also guarantee valid data mining results. Huge volume of detailed personal data is regularly collected and sharing of these data is proved to be beneficial for data mining application. Such data include shopping habits, criminal records, medical history, credit records etc. On one hand such data is an important asset to business organization and governments for decision making by analyzing it. On the other hand privacy regulations and other privacy concerns may prevent data owners from sharing information for data analysis. In order to share data while preserving privacy data owner must come up with a solution which achieves the dual goal of privacy preservation as well as accurate clustering result.

Keywords: Association Rule mining, Privacy preserving, Data Transformation, Sliding Window Algorithm.

1. INTRODUCTION

A database is an organized and typically large collection of detailed facts concerning some domain in the outside world. The aim of Data Mining is to examine this database for regularities that may lead to a better understanding of the domain described by the database. Data mining generally assume that the database consists of a collection of individuals. Depending on the domain, individuals can be anything from customers of a bank to molecular compounds or books in a library. For each individual, the database gives us detailed information concerning the different characteristics of the individual, such as the name and address of a customer of a bank, or the accounts owned. Over the last twenty years, there has been a extensive growth in the amount of private data collected about individuals. This data comes from a number of sources including medical, financial, library, telephone, and shopping records. Such data can be integrated and finalized digitally as it's possible due to the rapid growth in database, networking, and computing technologies. On the one hand, this has led to the development of data mining tools that aim to infer useful trends from this data. But, on the other hand, easy access to personal data poses a threat to individual privacy.

2. DATA MINING AND PRIVACY

Data mining deals with large database which can contain sensitive information. It requires data preparation which can uncover information or patterns which may compromise confidentiality and privacy obligations. Advancement of efficient data mining technique has increased the disclosure risks of sensitive data. A common way for this to occur is through data aggregation. Data aggregation is when the data are accrued, possibly from various sources, and put together so that they can be analyzed. This is not data mining per se, but a result of the preparation of data before and for the purposes of the analysis. The threat to an individual's privacy comes into play when the data, once compiled, cause the data miner, or anyone who has access to the newly compiled data set, to be able to identify specific individuals, especially when originally the data were anonymous. Data mining causes is social and ethical problem by revealing the data which should require privacy. Providing security to sensitive data against unauthorized access has been a long term goal for the database security research community and for the government statistical agencies. Hence, the security issue has become, recently, a much more important area of research in data mining. Therefore, in recent years, privacy-preserving data mining has been studied extensively.

3. MOTIVATION

Recent developments in information technology have made possible the collection and analysis of millions of transactions containing personal data. These data include shopping habits, criminal records, medical histories, and credit records, among others. This progress in the storage and analysis of data has led individuals and organizations to face the challenge of turning such data into useful information and knowledge. Data mining is a promising approach to meet this challenging requirement. The area of data mining, also called Knowledge Discovery in Databases (KDD), has received special attention since the 1990s. This new research area has emerged as a means of extracting hidden patterns or previously unknown implicit

information from large repositories of data. The fascination with the promise of analysis of large volumes of data has led to an increasing number of successful applications of data mining in recent years. Undoubtedly, these applications are very useful in many areas such as marketing, business, medical analysis, and other applications in which pattern discovery is paramount for strategic decision making. Despite its benefits in various areas, the use of data mining techniques can also result in new threats to privacy and information security. The problem is not data mining itself, but the way data mining is done. Data mining results rarely violate privacy, as they generally reveal high-level knowledge rather than disclosing instances of data. However, the concern among privacy advocates is well founded, as bringing data together to support data mining projects makes misuse easier. Thus in the absence of adequate safeguards, the use of data mining can jeopardize the privacy and autonomy of individuals. More serious is the privacy invasion occasioned by secondary usage of data when individuals are unaware of "behind the scenes" use of data mining techniques.

Even though many nations have developed privacy protection laws and regulations to guard against private use of personal information, the existing laws and their conceptual foundations have become outdated because of changes in technology. As a result, these personal data reside on thousands of file servers, largely beyond the control of existing privacy laws, leading to potential privacy invasion on a scale never before possible. Complex issues, such as those involved in privacy-preserving data mining (PPDM), can-not simply be addressed by restricting data collection or even by restricting the secondary use of information technology. Moreover, there is no exact solution that resolves privacy preservation in data mining. An approximate solution could be sufficient, depending on the application since the appropriate level of privacy can be interpreted in different contexts. In some applications (e.g., association rules, classification, or clustering), an appropriate balance between a need for privacy and knowledge discovery should be found. Preserving privacy when data are shared for mining is a challenging problem. The traditional methods in database security, such as access control and authentication that have been adopted to successfully manage the access to data present some limitations in the context of data mining. While access control and authentication protections can safe-guard against direct disclosures, they do not address disclosures based on inferences that can be drawn from released data. Preventing this type of inference detection is beyond the reach of the existing methods. Clearly, privacy issues pose new challenges for novel uses of data mining technology. These technical challenges indicate a pressing need to rethink mechanisms to address some issues of privacy and accuracy when data are either shared or exchanged before mining. Such mechanisms can lead to new privacy control methods to convert a database into a new one that conceals private information while preserving the general patterns and trends from the original database.

4. PROBLEM DEFINITION

We address the problem of transforming a database into a new one that conceals sensitive information while preserving the general patterns and trends from the original database. The sensitive information is not limited to personal data, but may reflect customer's purchasing behaviour, financial, medical, and insurance liability information and sensitive patterns. The transformation applied to the database occurs before the sharing of data for mining, as can be seen in Figure-1.

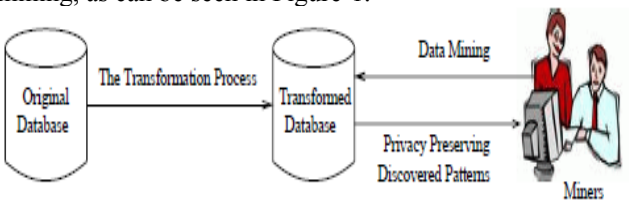

Figure-1: An example of a database transformed before the mining phase

We focus primarily on privacy preserving data mining, notably in the context of the mining tasks: a) association rules which describe interesting relationships among items grouped together in a sufficient number of examples; and b) clustering which is concerned with grouping objects into classes of similar objects.

5. RELATED WORK

5.1 Classification of Privacy Preserving Data Mining

According to [1] work done in PPDM can be classified according to different categories. These are

Data Distribution - The PPDM algorithms can be first divided into two major categories, centralized and distributed data, based on the distribution of data. In a centralized database environment, data are all stored in a single database; while, in a distributed database environment, data are stored in different databases. Distributed data scenarios can be further classified into horizontal and vertical data distributions. Horizontal distributions refer to the cases where different records of the same data attributes are resided in different places. While in a vertical data distribution, different attributes of the same record of data are resided in different places. Earlier research has been predominately focused on dealing with privacy preservation in a centralized database. The difficulties of applying PPDM algorithms to a distributed database can be attributed to: first, the data owners have privacy concerns so they may not willing to release their own data for others; second, even if they are willing to share data, the communication cost between the sites is too expensive.

Hiding Purposes - The PPDM algorithms can be further classified into two types, data hiding and rule hiding, according to the purposes of hiding. Data hiding refers to the cases where the sensitive data from original database like identity, name, and address that can be linked, directly or indirectly, to an individual person are hided. In contrast, in rule hiding, the sensitive knowledge (rule) derived from original database after applying data mining algorithms is removed. Majority of the PPDM algorithms used data

hiding techniques. Most PPDM algorithms hide sensitive patterns by modifying data.

Privacy Preservation Techniques - PPDM algorithms can further be divided according to privacy preservation techniques used. Four techniques – sanitation, blocking, distort, and generalization -- have been used to hide data items for a centralized data distribution. The idea behind data sanitation is to remove or modify items in a database to reduce the support of some frequently used item sets such that sensitive patterns cannot be mined. The blocking approach replaces certain attributes of the data with a question mark. In this regard, the minimum support and confidence level will be altered into a minimum interval. As long as the support and/or the confidence of a sensitive rule lie below the middle in these two ranges, the confidentiality of data is expected to be protected. Also known as data perturbation or data randomization, data distort protects privacy for individual data records through modification of its original data, in which the original distribution of the data is reconstructed from the randomized data. These techniques aim to design distortion methods after which the true value of any individual record is difficult to ascertain, but "global" properties of the data remain largely unchanged. Generalization transforms and replaces each record value with a corresponding generalized value.

5.2 Techniques of Privacy Preserving Data Mining

Most methods for privacy computations use some form of transformation on the data in order to perform the privacy preservation. Typically, such methods reduce the granularity of representation in order to reduce the privacy. This reduction in granularity results in some loss of effectiveness of data management or mining algorithms. This is the natural trade-off between information loss and privacy. Some examples of such technique as described in [2] are:

Randomization method - The randomization technique uses data distortion methods in order to create private representations of the records .In this which noise is added to the data in order to mask the attribute values of records. In most cases, the individual records cannot be recovered, but only aggregate distributions can be recovered. These aggregate distributions can be used for data mining purposes. Data mining techniques can be developed in order to work with these aggregate distributions. Two kinds of perturbation are possible with the randomization method:

Additive Perturbation - In this case, randomized noise is added to the data records. The overall data distributions can be recovered from the randomized records. Data mining and management algorithms re designed to work with these data distributions.

Multiplicative Perturbation- In this case, the random projection or random rotation techniques are used in order to perturb the records.

The k-anonymity model and l-diversity-The k-anonymity model was developed because of the possibility of indirect identification of records from public databases. This is because combinations of record attributes can be used to exactly identify individual records. In the k-anonymity method, the granularity of data representation is reduced with the use of techniques such as generalization and suppression. This granularity is reduced sufficiently that any given record maps onto at least k other records in the data. The l-diversity model was designed to handle some weaknesses in the k-anonymity model since protecting identities to the level of k-individuals is not the same as protecting the corresponding sensitive values, especially when there is homogeneity of sensitive values within a group.

Distributed privacy preservation- In many cases, individual entities may wish to derive aggregate results from data sets which are partitioned across these entities. Such partitioning may be horizontal (when the records are distributed across multiple entities) or vertical (when the attributes are distributed across multiple entities). While the individual entities may not desire to share their entire data sets, they may consent to limited information sharing with the use of a variety of protocols. The overall effect of such methods is to maintain privacy for each individual entity, while deriving aggregate results over the entire data.

Downgrading Application Effectiveness - In many cases, even though the data may not be available, the output of applications such as association rule mining, classification or query processing may result in violations of privacy. This has lead to research in downgrading the effectiveness of applications by either data or application modifications.

6. PRIVACY-PRESERVING USING ASSOCIATION RULE MINING METHOD

The sharing of association rules is often beneficial in industry, but requires privacy safe-guards. One may decide to disclose only part of the knowledge mined from databases, and protect sensitive knowledge represented by sensitive rules. These sensitive rules must re-main private since they are essential for strategic decisions. Some companies prefer to share their data for collaboration, while others prefer to share only the patterns discovered from their data. Our algorithms presented in this chapter take into account these two important aspects, i.e., the sharing of data and the sharing of patterns. The process of protecting sensitive rules in transactional databases is called data sanitization. We describe some scenarios that demonstrate the need for techniques to protect collective privacy (e.g., sensitive knowledge) in association rule mining. This framework is composed of a retrieval facility (e.g., inverted index), a set of algorithms to "sanitize" a database, and a set of metrics to measure how much private information is disclosed as well as the impact of the sanitizing algorithms on valid mining results. We introduce data sharing-based sanitizing algorithms in which the sanitization process acts on the data to remove or hide the group of sensitive association rules. After sanitizing a database, the released database is shared for association rule mining. A different approach to hide sensitive knowledge is introduced called pattern sharing-based. In this approach, the sanitizing algorithm acts on the rules mined from a database instead of the data itself. Rather than sharing the data, data owners

may prefer to mine their own data and share some discovered patterns. The sanitization removes not only all sensitive patterns but also blocks other patterns that could be used to infer the sensitive hidden ones.

6.1 The Framework For Privacy-Preserving Association Rule Mining

In this section, we introduce the framework to address privacy preservation in association rule mining. As depicted in Figure-2, the framework encompasses an inverted le 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.

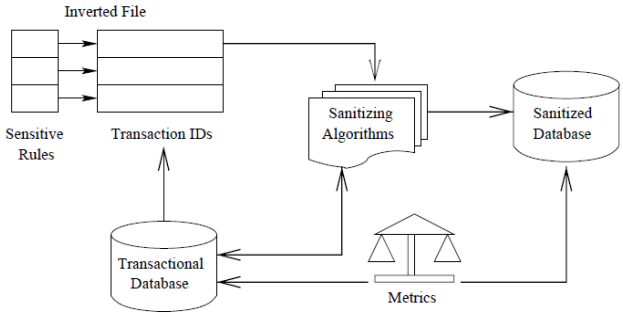

Figure-2: The sketch of the framework for privacy-preserving association rule mining.

6.2 The Sliding Window Algorithm (SWA)

The intuition behind this algorithm is that the SWA scans a group of K transactions (window size) at a time. SWA then sanitizes the set of sensitive transactions, denoted by ST, considering a disclosure threshold defined by a database owner. We applied to every group of K transactions read from the original database D. Unlike the previous sanitizing algorithms that have a unique disclosure threshold for all sensitive rules, the SWA has a disclosure threshold assigned to each sensitive association rule. We refer to the set of mappings of a sensitive association rule into its corresponding disclosure threshold as the set of mining permissions, denoted by MP, in which each mining permission mp is characterized by an ordered pair, defined as mp = < sr_i, ψ_i >, where $\forall$ i , $sr_i \varepsilon$ SR and $\psi_i \varepsilon [0\ldots1]$. The sketch of the Sliding Window algorithm is given as follows:

Input D, M_P, K
Output: D'.
Step-1: begin
Step-2: foreach K transactions in D do
Step-3: Identifying sensitive transactions and building index T
Step-4: foreach transaction t ε K do
Sort the items in t is alphabetic order;
Step-5: foreach sensitive association rule $sr_i \varepsilon M_P$ do
if items(sr_i) $\subseteq$ t then
$T[sr_i].tid\_list \leftarrow T[sr_i].tid\_list \bigcup TID\_of(t)$;
$T[sr_i].size\_list \leftarrow T[sr_i].size\_list \bigcup size(t)$;
$freq[item_j] \leftarrow freq[item_j] + 1$;
$v\_transac \leftarrow v\_transac \bigcup t$;
end
end
Step-6: if t is sensitive then
Sort vector freq in descending order;
foreach sensitive association rule $sr_i \varepsilon M_P$ do
Select $item_v$ such that $tem_v \varepsilon sr_{i\ and} \forall item_k \varepsilon sr_i$
$freq[item_v] \geq freq[item_k]$
if $freq[item_v] > 1$ then
$T[sr_i].victim \leftarrow T[sr_i].victim \bigcup item_v$;
else
$T[sr_i].victim \leftarrow T[sr_i].victim \bigcup Randomitem(sr_i)$;
end
end
end
end
end
Step-7: Selecting the number of sensitive transaction
foreach sensitive association rule $sr_i \varepsilon M_P$ do
$NumbTrans_{sri} \leftarrow [T[sr_i]] \times (1-\psi)$
Sort the vector in ascending order of size;
end
Step-8: $D' \leftarrow D$
foreach sensitive association rule $sr_i \varepsilon M_P$ do
for j = 1 to $NumbTrans_{sri}$ do
remove(v_transac[$T[sr_i].tid\_list[j]$, $T[sr_i].victim[j]$]);
else
if $\psi_i = 0$ then
do look_ahead(sr_i, $T[sr_i].tid\_list[j]$, $T[sr_i].victim[j]$]);
end
end
end
end

Figure-3: Sliding Window Algorithm

The inputs for the Sliding Window algorithm are a transactional database D, a set of mining permissions MP, and the window size K. The output is the sanitized database D'. The SWA has essentially four steps. In the first, the algorithm scans K transactions and stores some information in the data structure T. This data structure contains: 1) a list of sensitive transactions IDs for each sensitive rule; 2) a list with the size of the corresponding sensitive transactions; and 3) another list with the victim item for each corresponding sensitive transaction. A transaction t is sensitive if it contains all items of at least one sensitive rule. The SWA also computes the frequencies of the items of the sensitive rules that are present in each sensitive transaction. This computation will support the selection of the victim items in the next step.

In step 2, the vector with the frequencies, computed in the previous step, is sorted in descending

order. Subsequently, the victim item is selected for each sensitive transaction. The item with the highest frequency is the victim item and must be marked to be removed from the transaction. If the frequency of the items is equal to 1, any item from a sensitive association rule can be the victim item. In this case, we select the victim item randomly.

In the last step, the sensitive transactions are sanitized. If the disclosure threshold is 0 (i.e., all sensitive rules need to be hidden), we do a look ahead in the mining permissions (MP) to check whether a sensitive transaction need not be sanitized more than once. This is to improve the misses cost. The function look ahead() looks in MP from sri onward to determine whether a given transaction t is selected as a sensitive transaction for another sensitive rule r. If this is the case and, transac[T[sri].tid_list[j]] and T[sri].victim[j]] are part of the sensitive rule r, the transaction t is removed from that list since it has already just been sanitized.

7. Results And Discussions

We validated our methods for privacy-preserving clustering and privacy-preserving association rule mining using nine real datasets.

Dataset	#records	# items	Avg. Length	Shortest Record	Longest Record
BMS-Web-View-1	59,602	497	2.51	1	145
Retail	88,162	16,470	10.30	1	76
Accidents	340,183	468	33.81	18	51
Kosarak	990,573	41,270	8.10	1	1065
Reuters	7,774	26,639	46.81	1	427
Mushroom	8,124	119	23	23	23
Chess	3,196	75	37	37	37
Connect	67,557	129	43	43	43
Pumbs	49,046	2,113	74	74	74

Table-1: A summary of the datasets used in our experiments

Table-1 shows the summary of the datasets used in our experiments. The columns represent, respectively, the database name, the total number of records, the number of distinct items, the average number of items per record (transaction), the size of the shortest record, and the size of largest record. We purposely selected the sensitive rules to be sanitized based on four different scenarios, as follows:

S1: The sensitive rules selected contain only items that are mutually exclusive. In other words, there is no intersection of items over all the sensitive rules. The purpose of this scenario is to unflavored the algorithm SWA, which take advantage of rule overlaps.

S2: In this scenario, the sensitive rules were selected randomly.

S3: Only sensitive rules with very high support were selected. Sanitizing such rules would maximize the differential between an original dataset and its corresponding sanitized dataset.

S4: Only sensitive rules with low support were selected. Sanitizing such rules would minimize the differential between an original dataset and its corresponding sanitized dataset.

Table-2 shows the parameters we used to mine the datasets before the selection of the sensitive rules.

Dataset	Support (%)	Confidence (%)	No. Rules	Max. Size
BMS-1	0.1	60	25,391	7 items
Retail	0.1	60	7,319	6 items
Reuters	5.5	60	16,559	10 items
Kosarak	0.2	60	349,554	13 items

Table-2: Parameters used for mining the four datasets

We evaluated the effect of the window size, for the SWA algorithm, with respect to the difference between an original dataset D and its corresponding sanitized dataset D', misses cost, and hiding failure. To do so, we varied the K (window size) from 500 to 100,000 transactions with the disclosure threshold ψ=25%. Similarly, these metrics improve after 40,000 transactions for the datasets Kosarak, Retail, and BMS-1. The results reveal that a window size representing 64.31% of the size of the Reuters dataset suffices to stabilize the misses cost and hiding failure, while a window size representing 4.04%, 45.37%, and 67.11% is necessary to stabilize the same measures in the datasets Kosarak, Retail, and BMS-1, respectively. In this example, we intentionally selected a set of 6 sensitive association rules with high support (scenario S3) to accentuate the differential between the sizes of the original database and the sanitized database and thus to better illustrate the effect of window size on the difference between D and D', misses cost, and hiding failure. Note that the distribution of the data affects the values for misses cost and hiding failure. To obtain the best results for misses cost and hiding failure, here after we set the window size K to 50,000. 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.

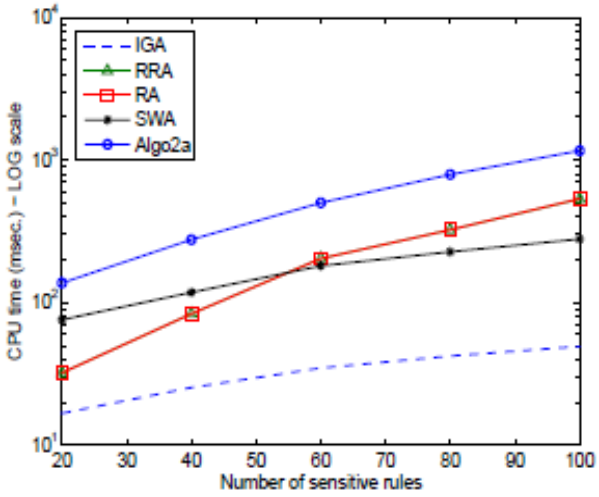

Figure-4: Results of CPU time

We 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 ψ= 0%. Figure-4 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 Item Group Algorithm requires 2 scans, it was faster than SWA in all the cases. The main reason is that the Sliding Window Algorithm performs a number of operations in main memory to fully sanitize a database.

8. CONCLUSION

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, we have already seen great interests in it: a) the recent proliferation in PPDM techniques is evident; b) the interest from academia and industry has grown quickly; and c) separate workshops and conferences devoted to this topic have emerged in the last few years. Privacy issues have posed new challenges for novel uses of data mining technology. These technical challenges cannot simply be addressed by restricting data collection or even by restricting the secondary use of information technology. An approximate solution could be sufficient, depending on the application since the appropriate level of privacy can be interpreted in different contexts. In some applications (e.g., association rules, classification, or clustering), an appropriate balance between a need for privacy and knowledge discovery should be found. We addressed the problem of transforming a database into a new one that conceals sensitive information while preserving the general patterns and trends from the original database. The sensitive information is not limited to personal data, but may reflect customers purchasing behaviour, financial, medical, and insurance liability information and sensitive patterns, considered sensitive patterns for strategic or competitive reasons by the caretaker or owner of the data. The transformation applied to the database occurs before the sharing of data for mining. We focused primarily on PPDM, notably in the context of the mining tasks: a) association rules which describe interesting relationships among items grouped together in a sufficient number of examples; and b) clustering which is concerned with grouping objects into classes of similar objects. We investigated the feasibility of achieving PPDM by data transformation.

REFERENCES

[1] Wu Xiaodan, Chu Chao-Hsien, Wang Yunfeng, Liu Fengli, Yue Dianmin, Privacy Preserving Data Mining Research: Current Status and Key Issues, Computional Science- ICCS 2007,4489(2007), 762-772.

[2] Agarwal Charu C., Yu Philip S., Privacy Preserving Data Mining: Models and Algorithms, New York, Springer, 2008.

[3] Berkhin Pavel, A Survey of Clustering Data Mining Techniques, Springer Berlin Heidelberg, 2006.

[4] Agrawal R., Srikant R. Privacy preserving data mining. In: Proceeedings of the ACM SIGMOD Conference of Management of Data, pp. 439–450. ACM (2000).

[5] Bramer Max, Principles of Data Mining, London, Springer, 2007.

[6] Oliveira S.R.M, Zaiane Osmar R., A Privacy-Preserving Clustering Approach Toward Secure and Effective Data Analysis for Business Collaboration, In Proceedings of the International Workshop on Privacy and Security Aspects of Data Mining in conjunction with ICDM 2004, Brighton, UK, November 2004.

[7] Wang Qiang , Megalooikonomou, Vasileios, A dimensionality reduction technique for efficient time series similarity analysis, Inf. Syst. 33, 1 (Mar.2008), 115- 132.

[8] S. Agrawal, V. Krishnan, and J. R. Haritsa. On Addressing Efficiency Concerns in Privacy-Preserving Mining. In Proc. of the 9th International Conference on Database Systems for Advanced Applications (DASFAA-2004), Jeju Island, Korea, March 2004.

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

[10] 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.

[11] M. Ben-Or, S. Goldwasser, and A. Wigderson. Completeness Theorems for Non-Cryptographic Fault-Tolerant Distributed Computation. In Proc. of the 20th ACM Symposium on Theory of Computing, pages 1-10, Chicago, Illinois, USA, 1988.

[12] E. Bingham and H. Mannila. Random Projection in Dimensionality Reduction: Applications to Image and Text Data. In Proc. of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 245-250, San Francisco, CA, USA, 2001.

[13] C.L. Blake and C.J. Merz. UCI Repository of Machine Learning Databases, University of California, Irvine, Dept. of Information and Computer Sciences, 1998.

[14] L. Brankovic and V. Estivill-Castro. Privacy Issues in Knowledge Discovery and Data Mining. In Proc. of Australian Institute of Computer Ethics Conference (AICEC99), Melbourne, Victoria, Australia, July 1999.

[15] 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.

[16] A. J. Menezes, P. C. van Oorschot, and S. A. Vanstone. Handbook of Applied Cryptography. CRC Press, LLC, Fifth Printing, August 2001.

[17] B. Schneier. Applied Cryptography: Protocols, Algorithms, and Source Code in C.John Wiley & Sons, Inc., Second Edition, 1996.

[18] A. A. Veloso, W. Meira Jr., S. Parthasarathy, and M. B. Carvalho. Efficient, Accurate and Privacy-Preserving Data Mining for Frequent Itemsets in Distributed Databases. In Proc. of the 18th Brazilian Symposium on Databases, pages 281-292, Manaus, Brazil, October 2003.

[19] S. Meregu and J. Ghosh. Privacy-Preserving Distributed Clustering Using Generative Models. In Proc. of the 3rd IEEE International Conference on Data Mining (ICDM'03), pages 211-218, Melbourne, Florida, USA, November 2003.

[20] A. Evfimievski. Randomization in Privacy Preserving Data Mining. SIGKDD Explorations, 4(2):43-48, December 2002.