RESEARCH ARTICLE OPEN ACCESS

A Survey on Security and Privacy Implications of Privacy Preserving Data Mining

R.Hemalatha [1], M.Elamparithi [2]
Research Scholar [1], Assistant Professor [2]
Department of Computer Science,
Sree Saraswathi Thyagaraja College, Pollachi
Tamil Nadu – India

ABSTRACT
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 fo r 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. 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. In particular, we address the problem of transforming a database to be shared into a new one that conceals private information while preserving the general patterns and trends from the original database. To address this challenging problem, we propose a unified framework for privacy-preserving data mining that ensures that the mining.

Keywords:- Data Cluster, Privacy preserving, K-Means algorithm.

I. INTRODUCTION

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.

II. RELATED WORK

2.1 DATA PARTITIONING TECHNIQUES

Data partitioning techniques have been applied to some scenarios in which the databases available for mining are distributed across a number of sites, with each site willing to share only data mining results, not the source data. In these cases, the data are distributed either horizontally or vertically [1]. In a horizontal partition, different entities are described with the same schema in all partitions, while in a vertical partition the attributes of the same entities are split across the partitions. The existing solutions can be classified into Cryptography-Based Techniques and Generative-Based Techniques.

2.1.1 Cryptography-Based Techniques

Cryptography-based techniques have been developed to solve problems of the following nature: two or more parties want to conduct a computation based on their private inputs. The issue here is how to conduct such a computation so that no party knows anything except its own input and the results. This problem is referred to as the secure multi-party computation problem [2, 3, 4]. Generally speaking, secure multi-party computation is the branch of cryptography that deals with the realization of distributed tasks in a secure manner; in this case, the definition of security can have different flavours, such as preserving the privacy of the data or protecting the computation against malicious attacks [5]. Typically, secure multi-party computation consists of computing some function f(x; y), where input x is in the hands of one participant and input y is in the hands of the other. For the computation to be secure, no more information is revealed to a participant than can be inferred from that participant's input and the output of the function itself.

The idea behind secure multi-party computation was introduced in [6]. The paper introduces a technique that enables the implementation of any probabilistic computation between two participants in a secure manner. Later on, this technique was generalized to the setting of multiple participants [7, 8]. However, the concept of Privacy-Preserving Data Mining (PPDM) by using secure multi-party computation was introduced in [9]. In this model, two parties owning confidential databases (e.g. confidential patient records) wish to run a data mining algorithm on the union of their databases

without revealing any unnecessary information. In particular, this paper focuses on the problem of decision tree learning and uses ID3 [10], a popular and widely used algorithm for this problem. The training set is distributed between two parties. This approach treats PPDM as a special case of secure multi-party computation, and not only aims at preserving individual privacy but also tries to preserve leakage of any information other than the final result. The solution is efficient for data partition applications and demands slow overhead of communication and reasonable bandwidth.

The solutions presented in [11, 12] aim at mining globally valid results from distributed data without revealing information that compromises the privacy of the individual sources. In particular, the work in [11] addresses secure mining of association rules over horizontally partitioned data. This approach considers the discovery of associations in transactions that are split across sites, without revealing the contents of individual transactions. In this model, the data available in all parties have the same schema, and it is assumed that three or more parties are involved to minimize the leakage of information. The solution is based on secure multi-party computation to minimize the information shared, while adding overhead to the mining task. On the other hand, the work in [12], addresses the problem of association rule mining in which transactions are distributed across sources. Each site holds some attributes of each transaction, and the sites wish to collaborate to identify globally valid association rules. In this model, two parties are involved, one party being designated as the primary, which is the initiator of the protocol. The other party is the responder. There is a join key present in both databases. The goal is to find association rules involving attributes other Than the join key. In the context of privacy-preserving data clustering, the first solution using secure multiparty computation was introduced in [13]. Specifically, a method for k-means clustering was proposed when different sites contain different attributes for a common set of entities. Each site has information for all the entities for a specific subset of attributes. In this model, it is assumed that the existence of an entity in a particular site's database may be revealed (e.g., because of join operations with other parties). However, the values associated with an entity are private. In this solution, each site learns the cluster of each entity, but learns nothing about the attributes of an entity at other sites. This work ensures reasonable privacy while limiting communication cost.

Regarding privacy preservation in classification, one solution was proposed in [14] based on a Naive Bayes classifier. Naive Bayes is based on a Bayesian formulation of the classification problem which uses the simplifying assumption of attribute independence. This approach assumes that the data available for mining are horizontally partitioned, i.e., all parties involved 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, the parties decide on a model. The model's 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. More recently, a new approach was introduced in [15] to address privacy preservation in classification. An algorithm is proposed to protect data before a data mining process takes place. The algorithm encrypts not only the attribute values but also the attribute labels. The algorithm is reversible, thus allowing the results of the models to be translated back to the readable form, but only by the database owner. Although the data are encrypted before the data mining process, the data remain unchanged (not-distorted), and the statistics inside the data remain the same. In this way, the modeling algorithm performs equally well on the protected as on the non-protected data. This approach is especially useful when the knowledge discovery process is outsourced. It should be pointed out that although universal and general, secure multi-party computation can be very inefficient and heavy in terms of communication complexity when the inputs are large and when the function to compute is relatively complicated to describe [16, 17].

2.1.2 Generative-Based Techniques

Generative-based techniques are designed to perform distributed mining tasks. In this approach, each party shares just a small portion of its local model that is used to construct the global model. The existing solutions are built over horizontally partitioned data. The solution presented in [18] addresses privacy-preserving frequent itemsets in distributed databases. Each site Si ($3 \leq i \leq n$) sends its frequent itemsets to a combiner that finds the globally frequent item sets based on the local models. Each site uses another representation of the itemsets (ABC becomes, for instance, 0-14-28) in a way such that the combiner is not able to identify the itemsets. It is assumed that all sites use the same codification. After combining the locally frequent itemsets, the combiner sends the upper bound for the globally frequent itemsets to all sites, and each site is able to restore the original itemsets' codification. At this point, each site knows only the information concerning its frequent itemsets and the upper bound of the globally frequent itemsets. Site S1 then generates a random number for each of its itemsets. This number is then added to the support count of each itemset, and the perturbed support counts are sent to site S2. The algorithm continues in the same way as before up to the last iteration. After receiving the values of the total local counts, site Sn requests from site S1 the values of random numbers and their respective itemsets. Site Sn simply decrements each global support count by the respective number and checks which itemsets are locally frequent. It is shown that the global model generated is accurate and the communication cost requires only one round of message passing around the sites and one reduction operation to aggregate the final results.

The solution in [19] addresses privacy-preserving distributed clustering using generative models. This solution relies on Expectation Maximization (EM) based algorithms. These algorithms are guaranteed to asymptotically converge to a global model that is locally optimal as the sample size used to obtain the global model goes to infinity. The intuition behind this approach is that, rather than sharing parts of the original data or perturbed data, the parameters of suitable generative models are built at each local site. Such parameters are then transmitted to a central location. The best representative of all data is a certain "mean" model. It was empirically shown that such a model can be approximated by generating artificial samples from the underlying distributions using Markov Chain Monte Carlo techniques. This approach achieves high quality distributed clustering with acceptable privacy loss and low communication cost. This framework also encompasses a measure for quantifying privacy based on ideas from information theory.

2.2 DATA MODIFICATION TECHNIQUES

These techniques modify the original values of a database that needs to be shared, and in doing so, privacy preservation is ensured. The transformed database is made available for mining and must meet privacy requirements without losing the benefit of mining. In general, data modification techniques aim at finding an appropriate balance between privacy preservation and knowledge disclosure. Methods for data modification include noise addition techniques and space transformation techniques.

2.2.1 Noise Addition Techniques

In statistical databases, noise addition techniques are used to protect individuals' privacy, but at the expense of allowing partial disclosure, providing information with less statistical quality, and introducing biases into query responses [20]. In data mining, the major requirement of a security control mechanism (in addition to protect the privacy) is not to ensure precise and bias-free statistics but rather to preserve the high-level descriptions of knowledge discovered from large databases [21, 22]. Thus, the idea behind noise addition techniques for PPDM is that some noise (e.g., information not present in a particular tuple or transaction) is added to the original data to prevent the identification of confidential information relating to a particular individual. In other cases, noise is added to confidential attributes by randomly shuffling the attribute values to prevent the discovery of some patterns that are not supposed to be discovered. Noise addition techniques can be categorized into three groups: (1) data swapping techniques; (2) data perturbation techniques; and (3) data randomization techniques.

Data swapping techniques replace the original database with a new one that has the same probability

distribution. Such techniques are suitable for privacy protection in knowledge discovery. The idea behind data swapping is that it interchanges the values in the records of the database in such a way that statistics about groups (e.g., frequencies, averages, etc) are preserved. The method proposed in [22] was designed for privacy preservation in classification. In this approach, a new training set, which is released to miners, is a perturbed version of the original training set. A data owner first builds a local decision tree over true data and then swaps values amongst records in a leaf node of the tree to generate randomized training data. The swapping is performed over the confidential attribute (class label) rather than other attributes in the dataset. As the class is typically a categorical attribute containing just two different values, the swapping is performed by changing the class in a small number of records. This is achieved by randomly shuffling the values of the class in the heterogeneous leaves. It has also been shown that is possible to balance statistical precision against the security level by choosing to perform the swapping in the internal nodes rather than in the leaves of the decision tree, i.e., the closer to the root, the higher the security but the lower the precision.

The work presented in [22] was extended in [23, 24]. The proposed method adds noise to datasets used for building decision trees. The method was evaluated taking into account the noise added to the class label and the noise added to the other attributes in a dataset. The authors measured the data quality by the similarity between the tree produced from the original data and a tree produced from the perturbed data. It was experimentally shown that the decision trees built on the perturbed data are very similar to the decision trees built on the original data.

Data perturbation techniques distort the data to protect individuals' privacy by introducing an error (noise) to the original data. The noise is used to generate the new (distorted) database which is subjected to mining. Miners should be able to obtain valid results (e.g., patterns and trends) from the distorted data. As opposed to statistical data analysis, miners do not aim at obtaining a definite, unbiased statistical test that answers with a probabilistic degree of confidence whether the data fit a preconceived statistical model. Data mining is not about hypothesis testing but about the generation of plausible hypotheses [25, 22]. The work presented in [26] addresses privacy preservation in classification by using data perturbation. The proposed solution aims at building a decision-tree classifier from training data in which the values of individual records have been perturbed by adding random values from a probability distribution. The resulting data records look very different from the original records, and the distribution of data values is also very different from the original distribution. While it is not possible to accurately estimate original values in individual data records, the authors proposed a novel reconstruction procedure to accurately estimate the distribution of original data values. By using these reconstructed distributions, one is able to build classifiers whose accuracy is comparable to the accuracy of classifiers built with the original data. The distribution reconstruction process naturally leads to some loss of information, but it can be acceptable in many practical situations.

A new algorithm for distribution reconstruction was introduced in [27]. This algorithm is more effective than that one proposed in [26], in terms of information loss. More specifically, the new algorithm is based on Expectation Maximization (EM) algorithms. It converges to the maximum likelihood estimate of the original distribution based on the perturbed data. When a large amount of data is available, the EM algorithm provides robust estimates of the original distribution. It was shown that the EM algorithm was in fact identical to the Bayesian reconstruction proposed in [26], except for the approximation partitioning values into intervals. Furthermore, the work in [27] introduces two new metrics, namely privacy loss and information loss to capture the amount of data in an individual record leaked to the data mining algorithm and the fidelity of the estimate respectively. As previously mentioned, the distribution reconstruction obtained by using an EM algorithm was greatly improved in [27]. However, it is shown in [28] that the number of computations each iteration is proportional to the size of the dataset and the number of intervals used in the estimate. Thus, two ways to reduce such great amounts of computation were proposed in [28]. In the first approach, the problem is

studied from a signal processing viewpoint, and algorithms are proposed to reduce the computation in the original protocol of perturbation. In particular, a Fourier series-based method is presented to compute, in one step, a good initial estimate of the distribution to reduce the number of iterations. In the second approach, a scheme for data perturbation is presented by modifying the protocol of data perturbation proposed in [27]. Unlike EM algorithms, this scheme estimates the unknown distribution in one step, and it is very simple to implement. This approach also achieves significant improvements over the previous ones [26, 27] in terms of the small privacy loss and the high fidelity in the estimate of the distribution.

2.2.2 Space Transformation Techniques

Space transformation techniques are specifically designed to address privacy-preserving clustering. These techniques aim at protecting the underlying data values subjected to clustering without jeopardizing the similarity between objects under analysis. Thus, a space transformation technique must not only meet privacy requirements but also guarantee valid clustering results.

A hybrid geometric data transformation method was proposed in [29] to meet privacy requirements as well as guarantee valid clustering results. This method distorts numerical attributes by translations, scaling, and rotations or even by the combination of these geometric transformations. The viability of using either a specific or the combination of all transformations (hybrid) for privacy preserving clustering was extensively studied. The key finding was that by transforming a data matrix by rotations only, one would attain both accuracy and a reasonable level of privacy. In contrast, transformation by translations is feasible if one is interested in accuracy since distortion by translations does not provide any level of privacy. The experiments also revealed that the scaling, hybrid (all transformations), and the Additive Data Perturbation (ADP) method, the latter of which is widely used in statistical databases, offer some level of privacy, but they do not preserve the distances between data points after the transformation process. Therefore, they are not recommended for privacy-preserving clustering because they are non-isometric transformations. As a consequence, they jeopardize the similarity between data points compromising the clustering results.

A more accurate investigation on privacy-preserving clustering using geometric transformation is presented in [30]. In particular, it is shown that distorting attribute pairs in a database by using only rotations is a promising approach. In this work, a spatial data transformation method is introduced for privacy-preserving clustering, called Rotation-Based Transformation (RBT). The method is designed to protect the underlying attribute values subjected to clustering without jeopardizing the similarity between data objects under analysis. RBT can be seen as a technique that is similar to obfuscation since the transformation process makes the original data difficult to perceive or understand, and preserves all the information for clustering analysis.

III. PRIVACY-PRESERVING ASSOCIATION RULE MINING

As depicted in Figure-1, 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.

Sanitizing a transactional database consists of identifying the sensitive transactions and adjusting them. To speed up this process, we scan a transactional database only once and, at the same time, we build our retrieval facility (inverted file). 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.

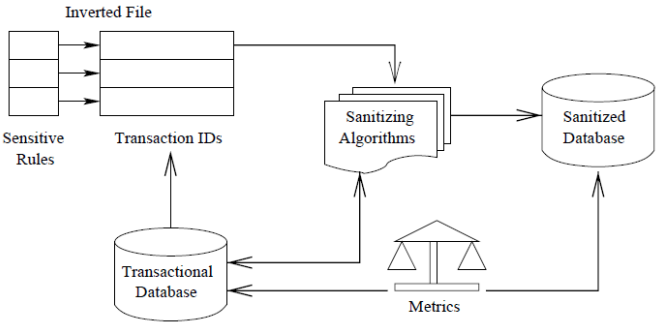

Figure-1: The sketch of the framework for privacy-preserving association rule mining.

Figure-2(b) shows an example of an inverted file corresponding to the sample transactional database shown in Figure-2(a). For this example, we 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

Inverted file

A,B → D	→	T1, T2
A,C → D	→	T1, T4

Sensitive Rules Transaction IDs

Figure-2(b) : The corresponding inverted file

Note that once the inverted file is built, a data owner will sanitize only the sensitive transactions whose 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.

IV. CONCLUSION

Privacy-preserving data mining 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 Privacy-preserving data mining 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.

REFERENCES

[1] 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, 2002.

[2] O. Goldreich, S. Micali, and A. Wigderson. How to Play Any Mental Game – A Completeness Theorem for Protocols with Honest Majority. In Proc. of the 19th Annual ACM Symposium on Theory of Computing, pages 218-229, New York City, USA, 1987.

[3] W. Du and M. J. Atallah. Secure Multi-Party Computation Problems and their Applications: A Review and Open Problems. In Proc. of 10th ACM/SIGSAC 2001 New Security Paradigms Workshop, pages 13-22, Cloudcroft, New Mexico, September 2001.

[4] B. Pinkas. Cryptographic Techniques For Privacy-Preserving Data Mining. SIGKDD Explorations, 4(2):12-19, December 2002.

[5] S. Goldwasser. Multi-party Computations: Past and Present. In Proc. of the 16th Annual ACM Symposium on Principles of Distributed Computing, pages 1-6, SantaBarbara, CA, August 1997.

[6] A.C.-C. Yao. How to Generate and Exchange Secrets. In Proc. of the 27th IEEE Symposium of Foundations of Computer Science, pages 162-167, Toronto, Ontario, Canada, October 1986.

[7] 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.

[8] D. Chaum, C. Crepeau, and I. Damgard. Multiparty Unconditionally Secure Protocols. In Proc. of the 20th ACM Symposium on Theory of Computing, pages 11-19, Chicago, Illinois, USA, 1988.

[9] Y. Lindell and B. Pinkas. Privacy Preserving Data Mining. In Crypto 2000, Springer-Verlag (LNCS 1880), pages 36-54, Santa Barbara, CA, August 2000.

[10] J. R. Quinlan. Learning E_cient Classi_cation Procedures and Their Application to Chess end Games. In R.S. Michalski, J.G. Carbonell, and T.M. Mitchell, eds., Machine Learning - An Artificial Intelligence Approach, pages 463-482, Tioga, Palo Alto, CA, 1983.

[11] M. Kantarcioglu and C. Clifton. Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data. In Proc. of The ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Madison, Wisconsin, June 2002.

[12] J. Vaidya and C. Clifton. Privacy Preserving Association Rule Mining in Vertically Partitioned Data. In Proc. of the 8th ACM SIGKDD Intl. Conf. on Knowlegde Discovery and Data Mining, pages 639-644, Edmonton, AB, Canada, July 2002.

[13] J. Vaidya and C. Clifton. Privacy-Preserving K-Means Clustering Over Vertically Partitioned Data. In Proc. of the 9th ACM SIGKDD Intl. Conf. on Knowlegde Discovery and Data Mining, pages 206-215, Washington, DC, USA, August 2003.

[14] M. Kantarcioglu and J. Vaidya. Privacy Preserving Naive Bayes Classifier for Horizontally Partitioned Data. In Proc. of the IEEE ICDM Workshop on Privacy Preserving Data Mining, pages 3-9, Melbourne, FL, USA, November 2003.

[15] B. Brumen, I. Golob, T. Welzer, I. Rozman, M. Dru_zovec, and H. Jaakkola. An Algorithm for Protecting Knowledge Discovery Data. INFORMATICA, 14(3):277- 288, December 2003.

[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] P. Tendick and N. S. Matloff. Recent Results on the Noise Addition Method for Database Security. In Proc. of the 1987 Joint Meetings, American Statistical Association / Institute of Mathematical Statistics (ASA/IMA), pages 406-409, Washington, DC, USA, 1987.

[21] 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.

[22] V. Estivill-Castro and L. Brankovic. Data Swapping: Balancing Privacy Against Precision in Mining for Logic Rules. In Proc. of Data Warehousing and Knowledge Discovery DaWaK-99, pages 389-398, Florence, Italy, August 1999.

[23] Md. Z. Islam, P. M. Barnaghi, and L. Brankovic. Measuring Data Quality: Predictive Accuracy vs. Similarity of Decision Trees. In Proc. of the 6th International Conference on Computer And Information Technology (ICCIT 2003), Dhaka, Bangladesh, December 2003.

[24] Md. Z. Islam and L. Brankovic. Noise Addition for Protecting Privacy in Data Mining. In Proc. of the 6th Engineering Mathematics and Applications Conference (EMAC 2003), Sydney, Australia, 2003.

[25] D. Hand, H. Mannila, and P. Smyth. Principles of Data Mining. The MIT Press,Cambridge, Massachusetts, 2001.

[26] 26. 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.

[27] 27. D. Agrawal and C. C. Aggarwal. On the Design and Quantification of Privacy Preserving Data Mining Algorithms. In Proc. of ACM SIGMOD/PODS, pages 247-255, Santa Barbara, CA, May 2001.

[28] 28. C. W. Wu. Privacy Preserving Data Mining: A Signal Processing Perspective and a Simple Data Perturbation Protocol. In Proc. of the IEEE ICDM Workshop on Privacy Preserving Data Mining, pages 10-17, Melbourne, FL, USA, November 2003.

[29] 29. S. R. M. Oliveira and O. R. Zaiane. Privacy Preserving Clustering By Data Transformation. In Proc. of the 18th Brazilian Symposium on Databases, pages 304-318, Manaus, Brazil, October 2003.

[30] 30. S. R. M. Oliveira and O. R. Zaiane. Achieving Privacy Preservation When Sharing Data For Clustering. In Proc. of the Workshop on Secure Data Management in a Connected World (SDM'04) in conjunction with VLDB'2004, pages 67-82, Toronto, Ontario, Canada, August 2004.