STATE-OF-THE-ART IN PRIVACY PRESERVING DATA MINING

by V.Verykios, E. Bertino, I. Nai Fovino, L.Parasiliti Provenza, Y.Saygin, Y. Theodoridis

Center for Education and Research in
Information Assurance and Security,
Purdue University, West Lafayette, IN 47907-2086

State-of-the-art in Privacy Preserving Data Mining*

Vassilios S. Verykios[1], Elisa Bertino[2], Igor Nai Fovino[2]
Loredana Parasiliti Provenza[2], Yucel Saygin[3], Yannis Theodoridis[1]

[1]*Academic and Research Computer Technology Institute, Athens, GREECE*
[2]*Dipartimento di Scienze dell'Informazione, Universita di Milano, Milano, ITALY*
[3]*Faculty of Engineering and Natural Sciences, SABANCI University, TURKEY*

Abstract

We provide here an overview of the new and rapidly emerging research area of privacy preserving data mining. We also propose a classification hierarchy that sets the basis for analyzing the work which has been performed in this context. A detailed review of the work accomplished in this area is also given, along with the coordinates of each work to the classification hierarchy. A brief evaluation is performed, and some initial conclusions are made.

1 Introduction

Data mining and knowledge discovery in databases are two new research areas that investigate the automatic extraction of previously unknown patterns from large amounts of data. Recent advances in data collection, data dissemination and related technologies have inaugurated a new era of research where existing data mining algorithms should be reconsidered from a different point of view, this of privacy preservation. It is well documented that this new without limits explosion of new information through the Internet and other media, has reached to a point where threats against the privacy are very common on a daily basis and they deserve serious thinking.

Privacy preserving data mining [9, 18], is a novel research direction in data mining and statistical databases [1], where data mining algorithms are analyzed for the side-effects they incur in data privacy. The main consideration in privacy preserving data mining is two fold. First, sensitive raw data like identifiers, names, addresses and the like, should be modified or trimmed out from the original database, in order for the recipient of the data not to be able to compromise another person's privacy. Second, sensitive knowledge which can be mined from a database by using data mining algorithms, should also be excluded, because such a knowledge can equally well compromise data privacy, as we will indicate. The main objective in privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process. The problem that arises when confidential information can be derived from released data by unauthorized users is also commonly called the "database inference" problem. In this report, we provide a classification and an extended description of the various techniques and methodologies that have been developed in the area of privacy preserving data mining.

2 Classification of Privacy Preserving Techniques

There are many approaches which have been adopted for privacy preserving data mining. We can classify them based on the following dimensions:

- data distribution
- data modification
- data mining algorithm
- data or rule hiding
- privacy preservation

The first dimension refers to the distribution of data. Some of the approaches have been developed for centralized data, while others refer to a distributed data scenario. Distributed data scenarios can also be classified as horizontal data distribution and vertical data distribution. Horizontal distribution refers to these cases where different database records reside in different places, while vertical data

*This work was supported by the CODMINE IST FET Project IST-2001-39151.

distribution, refers to the cases where all the values for different attributes reside in different places. The second dimension refers to the data modification scheme. In general, data modification is used in order to modify the original values of a database that needs to be released to the public and in this way to ensure high privacy protection. It is important that a data modification technique should be in concert with the privacy policy adopted by an organization. Methods of modification include:

- perturbation, which is accomplished by the alteration of an attribute value by a new value (i.e., changing a 1-value to a 0-value, or adding noise),
- blocking, which is the replacement of an existing attribute value with a "?",
- aggregation or merging which is the combination of several values into a coarser category,
- swapping that refers to interchanging values of individual records, and
- sampling, which refers to releasing data for only a sample of a population.

The third dimension refers to the data mining algorithm, for which the data modification is taking place. This is actually something that is not known beforehand, but it facilitates the analysis and design of the data hiding algorithm. We have included the problem of hiding data for a combination of data mining algorithms, into our future research agenda. For the time being, various data mining algorithms have been considered in isolation of each other. Among them, the most important ideas have been developed for classification data mining algorithms, like decision tree inducers, association rule mining algorithms, clustering algorithms, rough sets and Bayesian networks.

The fourth dimension refers to whether raw data or aggregated data should be hidden. The complexity for hiding aggregated data in the form of rules is of course higher, and for this reason, mostly heuristics have been developed. The lessening of the amount of public information causes the data miner to produce weaker inference rules that will not allow the inference of confidential values. This process is also known as "rule confusion".

The last dimension which is the most important, refers to the privacy preservation technique used for the selective modification of the data. Selective modification is required in order to achieve higher utility for the modified data given that the privacy is not jeopardized. The techniques that have been applied for this reason are:

- heuristic-based techniques like adaptive modification that modifies only selected values that minimize the utility loss rather than all available values
- cryptography-based techniques like secure multiparty computation where a computation is secure if at the end of the computation, no party knows anything except its own input and the results, and
- reconstruction-based techniques where the original distribution of the data is reconstructed from the randomized data.

It is important to realize that data modification results in degradation of the database performance. In order to quantify the degradation of the data, we mainly use two metrics. The first one, measures the confidential data protection, while the second measures the loss of functionality.

3 Review of Privacy Preserving Algorithms

3.1 Heuristic-Based Techniques

A number of techniques have been developed for a number of data mining techniques like classification, association rule discovery and clustering, based on the premise that selective data modification or sanitization is an NP-Hard problem, and for this reason, heuristics can be used to address the complexity issues.

3.1.1 Centralized Data Perturbation-Based Association Rule Confusion

A formal proof that the optimal sanitization is an NP-Hard problem for the hiding of sensitive large itemsets in the context of association rules discovery, have been given in [4]. The specific problem which was addressed in this work is the following one. Let D be the source database, R be a set of significant association rules that can be mined from D, and let R_h be a set of rules in R. How can we transform database D into a database D', the released database, so that all rules in R can still be mined from D', except for the rules in R_h. The heuristic proposed for the modification of the data was based on data perturbation, and in particular the procedure was to change a selected set of 1-values to 0-values, so that the support of sensitive rules is lowered in such a way that the utility of the released database is kept to some maximum value. The utility in this work is measured as the number of non-sensitive rules that were hidden based on the side-effects of the data modification process.

A subsequent work described in [10] extends the sanitization of sensitive large itemsets to the sanitization of sensitive rules. The approaches adopted in this work was either to prevent the sensitive rules from being generated by hiding the frequent itemsets from which they are derived, or to reduce the confidence of the sensitive rules by bringing it below a user-specified threshold. These two approaches led to the generation of three strategies for hiding sensitive rules. The important thing to mention regarding these three strategies were the possibility for both a 1-value in the binary database to turn into a 0-value and a 0-value to turn into a 1-value. This flexibility in data modification had the side-effect that apart from non-sensitive association rules that were becoming hidden, a non-frequent rule could become a frequent one. We refer to these rules as "ghost rules". Given that sensitive rules are hidden, both non-sensitive rules which were hidden and non-frequent rules that became frequent (ghost rules) count towards the reduced utility of the released database. For this reason, the heuristics used for this later work, must be more sensitive to the utility issues, given that the security is not compromised. A complete work which was based on this idea, can be found in [24].

The work in [19] builds on top of the work previously presented, and aims at balancing between privacy and disclosure of information by trying to minimize the impact on sanitized transactions or else to minimize the accidentally hidden and ghost rules.

3.1.2 Centralized Data Blocking-Based Association Rule Confusion

One of the data modification approaches which have been used for association rule confusion is data blocking [6]. The approach of blocking is implemented by replacing certain attributes of some data items with a question mark. It is sometimes more desirable for specific applications (i.e., medical applications) to replace a real value by an unknown value instead of placing a false value. An approach which applies blocking to the association rule confusion, has been presented in [22]. The introduction of this new special value in the dataset, imposes some changes on the definition of the support and confidence of an association rule. In this regard, the minimum support and minimum confidence will be altered into a minimum support interval and a minimum confidence interval correspondingly. As long as the support and/or the confidence of a sensitive rule lies below the middle in these two ranges of values, then we expect that the confidentiality of data is not violated. Notice that for an algorithm used for rule confusion in such a case, both 1-values and 0-values should be mapped to question marks in an interleaved fashion, otherwise, the origin of the question marks, will be obvious. An extension of this work with a detailed discussion on how effective is this approach on reconstructing the confused rules, can be found in [21].

3.1.3 Centralized Data Blocking-Based Classification Rule Confusion

The work in [5] provides a new framework combining classification rule analysis and parsimonious downgrading. Notice here, that in the classification rule framework, the data administrator, has as a goal to block values for the class label. By doing this, the receiver of the information, will be unable to build informative models for the data that is not downgraded. Parsimonious downgrading is a framework for formalizing the phenomenon of trimming out information from a data set for downgrading information from a secure environment (it is referred to as High) to a public one (it is referred to as Low), given the existence of inference channels. In parsimonious downgrading a cost measure is assigned to the potential downgraded information that it is not sent to Low. The main goal to be accomplished in this work, is to find out whether the loss of functionality associated with not downgrading the data, is worth the extra confidentiality. Classification rules, and in particular decision trees are used in the parsimonious downgrading context in analyzing the potential inference channels in the data that needs to be downgraded.

The technique used for downgrading is the creation of the so called parametric base set. In particular, a parameter $\theta, 0 \leq \theta \leq 1$ is placed instead of the value that is blocked. The parameter represents a probability for one of the possible values that the attribute can get. The value of the initial entropy before the blocking and the value of the entropy after the blocking is calculated. The difference in the values of the entropy is compared to the decrease in the confidence of the rules generated from the decision tree in order to decide whether the increased security is worth the reduced utility of the data the Low will receive.

In [17] the authors presented the design of a software system, the Rational Downgrader, that is based on the parsimonious downgrading idea. The system is composed of a knowledge-based decision maker, to determine the rules that may be inferred, a "guard" to measure the amount of leaked information, and a parsimonious downgrader to modify the initial downgrading decisions. The algorithm used to downgrade the data finds which rules from those induced from the decision tree induction, are needed to classify the private data. Any data that do not support the rules found in this way, are excluded from downgrading along with all the attributes that are not represented in the rules clauses. From the remaining data, the algorithm should decide which values to transform into missing values. This is done in order to optimize the rule confusion. The "guard" system determines the acceptable level of rule confusion.

3.2 Cryptography-Based Techniques

A number of cryptography-based approaches have been developed in the context of privacy preserving data mining algorithms, to solve problems of the following nature. Two or more parties want to conduct a computation based on their private inputs, but neither party is willing to disclose its own output to anybody else. The issue here is how to conduct such a computation while preserving the privacy of the inputs. This problem is referred to as the Secure Multiparty Computation (SMC) problem. In particular, an SMS problem deals with computing a probabilistic function on any input, in a distributed network where each participant holds one of the inputs, ensuring independence of the inputs, correctness of the computation, and that no more information is revealed to a participant in the computation than that's participant's input and output.

Two of the papers falling into this area, are rather general in nature and we describe them first. The first one [11] proposes a transformation framework that allows to systematically transform normal computations to secure multiparty computations. Among other information items, a discussion on transformation of various data mining problems to a secure multiparty computation is demonstrated. The data mining applications which are described in this domain, include data classification, data clustering, association rule mining, data generalization, data summarization and data characterization. The second paper [8] presents four secure multiparty computation based methods that can support privacy preserving data mining. The methods described include, the secure sum, the secure set union, the secure size of set intersection, and the scalar product. Secure sum, is often given as a simple example of secure multiparty computation, and we present it here as well, as an representative for the techniques used. Assume that the value $u = \sum_{l=1}^{s} u_l$ to be computed is known to lie in the range $[0, n]$. One site is designated as the master site and is given the identity 1. The remaining sites are numbered $2, \ldots, s$. Site 1 generates a random number R, uniformly chosen from $[0, n]$. Site 1 adds this number to its local value u_1 and sends the sum $R + u1 \bmod n$ to site 2. Since the value of R is chosen uniformly from $[0, n]$ the number $R + u_1 \bmod n$ is also distributed uniformly across this region, so site 2 learns nothing about the actual value of u_1. For the remaining sites $l = 2 \ldots s - 1$, the algorithm is as follows. Site l receives $V = R + \sum_{j=1}^{l-1} u_j \bmod n$. Since this value is uniformly distributed across $[0, n]$, i learns nothing. Site i then computes $R + \sum_{j=1}^{l} u_j \bmod n = (u_j + V) \bmod n$ and passes it to site $l + 1$. Site s performs the above step, and sends the result to site 1. Site 1, by knowing R, can subtract R to get the actual result. Below we present the approaches which have been developed by using the solution framework of secure multiparty computation. It should be made clear, that because of the nature of this solution methodology, the data in all of the cases that this solution is adopted, is distributed among two or more sites.

3.2.1 Vertically Partitioned Distributed Data Secure Association Rule Mining

Mining private association rules from vertically partitioned data, where the items are distributed and each itemset is split between sites, can be done by finding the support count of an itemset. If the support count of such an itemset can be securely computed, then we can check if the support is greater than the threshold, and decide whether the itemset is frequent. The key element for computing the support count of an itemset is to compute the scalar product of the vectors representing the sub-itemsets in the parties. Thus, if the scalar product can be securely computed, the support count can also be computed. The algorithm that computes the scalar product, as an algebraic solution that hides true values by placing them in equations masked with random values, is described in [23]. The security of the scalar product protocol is based on the inability of either side to solve k equations in more than k unknowns. Some of the unknowns are randomly chosen, and can safely be assumed as private. A similar approach has been proposed in [14]. Another way for computing the support count is by using the secure size of set intersection method described in [8].

3.2.2 Horizontally Partitioned Distributed Data Secure Association Rule Mining

In a horizontally distributed database, the transactions are distributed among n sites. The global support count of an itemset is the sum of all the local support counts. An itemset X is globally supported if the global support count of X is bigger than $s\%$ of the total transaction database size. A k-itemset is called a globally large k-itemset if it is globally supported. The work in [15] modifies the implementation of an algorithm proposed for distributed association rule mining [7] by using the secure union and the secure sum privacy preserving SMC operations.

3.2.3 Vertically Partitioned Distributed Data Secure Decision Tree Induction

The work described in [12] studies the building process of a decision tree classifier for a database that is vertically distributed. The protocol presented in this work, is built upon a secure scalar product protocol by using a third-party server.

3.2.4 Horizontally Partitioned Distributed Data Secure Decision Tree Induction

The work in [16] proposes a solution to the privacy preserving classification problem using a secure multiparty computation approach, the so-called oblivious transfer protocol for horizontally partitioned data. Given that a generic SMC solution is of no practical value, the authors focus on the problem of decision tree induction, and in particular the induction of ID3, a popular and widely-used algorithm for decision tree induction. The ID3 algorithm chooses the "best" predicting attribute by comparing entropies given as real numbers. Whenever the values for entropies of different attributes are close to each other, it is expected that the trees resulting from choosing either one of these attributes, have almost the same predicting capability. Formally stated, a pair of attributes has δ-equivalent information gains if the difference in the information gains is smaller than the value δ. This definition gives rise to an approximation of ID3. By denoting as $\mathcal{ID}3$, the set of all possible trees which are generated by running the ID3 algorithm, and choosing either attribute in the case that they are δ-equivalent, the work in [16] proposes a protocol for secure computation of a specific ID3$_\delta$ algorithm. The protocol for privately computing ID3$_\delta$ is composed of many invocations of smaller private computations. The most difficult computations among these reduces to the oblivious evaluation of $xlnx$ function.

3.2.5 Privacy Preserving Clustering

An algorithm for secure clustering by using the Expectation-Maximization algorithm is presented in [8]. The algorithm proposed is an iterative algorithm that makes use of the secure sum SMC protocol.

3.3 Reconstruction-Based Techniques

A number of recently proposed techniques address the issue of privacy preservation by perturbing the data and reconstructing the distributions at an aggregate level in order to perform the mining. Below, we list and classify some of these techniques.

3.3.1 Reconstruction-Based Techniques for Numerical Data

The work presented in [3] addresses the problem of building a decision tree classifier from training data in which the values of individual records have been perturbed. While it is not possible to accurately estimate original values in individual data records, the authors propose a reconstruction procedure to accurately estimate the distribution of original data values. By using the reconstructed distributions, they are able to build classifiers whose accuracy is comparable to the accuracy of classifiers built with the original data. For the distortion of values, the authors have considered a discretization approach and a value distortion approach. For reconstructing the original distribution, they have considered a Bayesian approach and they proposed three algorithms for building accurate decision trees that rely on reconstructed distributions.

The work presented in [2] proposes an improvement over the Bayesian-based reconstruction procedure by using an Expectation Maximization (EM) algorithm for distribution reconstruction. More specifically, the authors prove that the EM algorithm converges to the maximum likelihood estimate of the original distribution based on the perturbed data. They also show that when a large amount of data is available , the EM algorithm provides robust estimates of the original distribution. It is also shown, that the privacy estimates of [3] had to be lowered when the additional knowledge that the miner obtains from the reconstructed aggregate distribution was included in the problem formulation.

3.3.2 Reconstruction-Based Techniques for Binary and Categorical Data

The work presented in [20] and [13] deal with binary and categorical data in the context of association rule mining. Both papers consider randomization techniques that offer privacy while they maintain high utility for the data set.

4 Evaluation of Privacy Preserving Algorithms

An important aspect in the development and assessment of algorithms and tools, for privacy preserving data mining is the identification of suitable evaluation criteria and the development of related benchmarks. It is often the case that no privacy preserving algorithm exists that outperforms all the others on all possible criteria. Rather, an algorithm may perform better that another one on specific criteria, such as performance and/or data utility. It is thus important to provide users with a set of metrics which will enable them to select the most appropriate privacy preserving technique for the data at hand, with respect to some specific parameters they are interested in optimizing.

A preliminary list of evaluation parameters to be used for assessing the quality of privacy preserving data mining algorithms, is given below:

- the *performance* of the proposed algorithms in terms of time requirements, that is the time needed by each algorithm to hide a specified set of sensitive information;
- the *data utility* after the application of the privacy preserving technique, which is equivalent with the minimization of the information loss or else the loss in the functionality of the data;
- the *level of uncertainty* with which the sensitive information that have been hidden can still be predicted;
- the *resistance* accomplished by the privacy algorithms, to different data mining techniques.

Below we refer to each one of these evaluation parameters and we analyze them.

4.1 Performance of the proposed algorithms

A first approach in the assessment of the time requirements of a privacy preserving algorithm is to evaluate the computational cost. In this case, it is straightforward that an algorithm having a $O(n^2)$ polynomial complexity is more efficient than another one with $O(e^n)$ exponential complexity.

An alternative approach would be to evaluate the time requirements in terms of the average number of operations, needed to reduce the frequency of appearance of specific sensitive information below a specified threshold. This values, perhaps, does not provide an absolute measure, but it can be considered in order to perform a fast comparison among different algorithms.

The *communication cost* incurred during the exchange of information among a number of collaborating sites, should also be considered. It is imperative that this cost must be kept to a minimum for a distributed privacy preserving data mining algorithm.

4.2 Data Utility

The utility of the data, at the end of the privacy preserving process, is an important issue, because in order for sensitive information to be hidden, the database is essentially modified through the insertion of false information (swapping of values is a side effect in this case)or through the blocking of data values. We should notice here that some of privacy preserving techniques, like the use of sampling, do not modify the information stored in the database, but still, the utility of the data falls, since the information is not complete in this case. It is obvious that the more the changes are made to the database, the less the database reflects the domain of interest. Therefore, an evaluation parameter for the data utility should be the amount of information that is lost after the application of privacy preserving process. Of course, the measure used to evaluate the information loss depends on the specific data mining technique with respect to which a privacy algorithm is performed.

For example, information loss in the context of association rule mining will be measured either in terms of the number of rules that were both remaining and lost in the database after sanitization, or even in terms on the reduction/increase in the support and confidence of all the rules. For the case of classification, we can use metrics similar to those used for association rules. Finally, for clustering, the variance of the distances among the clustered items in the original database and the sanitized database, can be the basis for evaluating information loss in this case.

4.3 Uncertainty Level

The privacy preservation strategies, operate by downgrading the information that we want to protect below certain thresholds. The hidden information, however, can still be inferred even though with some uncertainty level. A sanitization algorithm then, can be evaluated on the basis of the uncertainty that it introduces during the reconstruction of the hidden information. From an operational point of view, a scenario would be to set a maximum to the perturbation of information, and then consider the degree of uncertainty achieved by each sanitization algorithm under this constraint. We expect that the algorithm that will attain the maximum uncertainty level, will be the one which will be preferred over all the rest .

4.4 Endurance of Resistance to different Data Mining techniques

The ultimate aim of hiding algorithms is the protection of sensitive information against unauthorized disclosure. In this case, it is important not to forget, that intruders and data terrorists will try to compromise information by using various data mining algorithms. Consequently, a sanitization algorithm developed against a particular data mining technique that assures privacy of information, may not attain similar protection against all possible data mining algorithms.

In order to provide for a complete evaluation of sanitization algorithms, we need to measure its endurance against data mining techniques which are different from the technique that a sanitization algorithm has been developed for. We call such a parameter the *transversal endurance*. The evaluation of this parameter, needs the consideration of a class of data mining algorithms which are significant for our test. Alternatively, we may need to develop a formal framework that upon testing of a sanitization algorithm against pre-selected data sets, we can transitively prove privacy assurance for the whole class of sanitization algorithms.

5 Conclusions

We have presented a classification and an extended description and clustering of various privacy preserving data mining algorithms. The work presented in here, indicates the ever increasing interest of researchers in the area of securing sensitive data and knowledge from malicious users. The conclusions that we have reached from reviewing this area, manifest that privacy issues can be effectively considered only within the limits of certain data mining algorithms. The inability to generalize the results for classes of categories of data mining algorithms might be a tentative threat for disclosing information.

References

[1] Nabil Adam and John C. Wortmann, *Security-Control Methods for Statistical Databases: A Comparison Study*, ACM Computing Surveys **21** (1989), no. 4, 515–556.

[2] Dakshi Agrawal and Charu C. Aggarwal, *On the design and quantification of privacy preserving data mining algorithms*, In Proceedings of the 20th ACM Symposium on Principles of Database Systems (2001), 247–255.

[3] Rakesh Agrawal and Ramakrishnan Srikant, *Privacy-preserving data mining*, In Proceedings of the ACM SIGMOD Conference on Management of Data (2000), 439–450.

[4] Mike J. Atallah, Elisa Bertino, Ahmed K. Elmagarmid, Mohamed Ibrahim, and Vassilios S. Verykios, *Disclosure Limitation of Sensitive Rules*, In Proceedings of the IEEE Knolwedge and Data Engineering Workshop (1999), 45–52.

[5] LiWu Chang and Ira S. Moskowitz, *Parsimonious downgrading and decision trees applied to the inference problem*, In Proceedings of the 1998 New Security Paradigms Workshop (1998), 82–89.

[6] LiWu Chang and Ira S. Moskowitz, *An integrated framework for database inference and privacy protection*, Data and Applications Security (2000), 161–172, Kluwer, IFIP WG 11.3, The Netherlands.

[7] David W. Cheung, Jiawei Han, Vincent T. Ng, Ada W. Fu, and Yongjian Fu, *A fast distributed algorithm for mining association rules*, In Proceedings of the 1996 International Conference on Parallel and Distributed Information Systems (1996).

[8] Chris Clifton, Murat Kantarcioglou, Xiadong Lin, and Michael Y. Zhu, *Tools for privacy preserving distributed data mining*, SIGKDD Explorations **4** (2002), no. 2.

[9] Chris Clifton and Donald Marks, *Security and privacy implications of data mining*, In Proceedings of the ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (1996), 15–19.

[10] Elena Dasseni, Vassilios S. Verykios, Ahmed K. Elmagarmid, and Elisa Bertino, *Hiding Association Rules by using Confidence and Support*, In Proceedings of the 4th Information Hiding Workshop (2001), 369–383.

[11] Wenliang Du and Mikhail J. Attalah, *Secure multi-problem computation problems and their applications: A review and open problems*, Tech.

Report CERIAS Tech Report 2001-51, Center for Education and Research in Information Assurance and Security and Department of Computer Sciences, Purdue University, West Lafayette, IN 47906, 2001.

[12] Wenliang Du and Zhijun Zhan, *Building decision tree classifier on private data*, In Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining (2002).

[13] Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, and Johannes Gehrke, *Privacy preserving mining of association rules*, In Proceedings of the 8th ACM SIGKDDD International Conference on Knowledge Discovery and Data Mining (2002).

[14] Ioannis Ioannidis, Ananth Grama, and Mikhail Atallah, *A secure protocol for computing dot products in clustered and distributed environments*, In Proceedings of the International Conference on Parallel Processing (2002).

[15] Murat Kantarcioglou and Chris Clifton, *Privacy-preserving distributed mining of association rules on horizontally partitioned data*, In Proceedings of the ACM SIGMOD Workshop on Research Isuues in Data Mining and Knowledge Discovery (2002), 24–31.

[16] Yehuda Lindell and Benny Pinkas, *Privacy preserving data mining*, In Advances in Cryptology - CRYPTO 2000 (2000), 36–54.

[17] Ira S. Moskowitz and LiWu Chang, *A decision theoretical based system for information downgrading*, In Proceedings of the 5th Joint Conference on Information Sciences (2000).

[18] Daniel E. O'Leary, *Knowledge Discovery as a Threat to Database Security*, In Proceedings of the 1st International Conference on Knowledge Discovery and Databases (1991), 107–516.

[19] Stanley R. M. Oliveira and Osmar R. Zaiane, *Privacy preserving frequent itemset mining*, In Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining (2002), 43–54.

[20] Shariq J. Rizvi and Jayant R. Haritsa, *Maintaing data privacy in association rule mining*, In Proceedings of the 28th International Conference on Very Large Databases (2002).

[21] Yucel Saygin, Vassilios Verykios, and Chris Clifton, *Using unknowns to prevent discovery of association rules*, SIGMOD Record **30** (2001), no. 4, 45–54.

[22] Yucel Saygin, Vassilios S. Verykios, and Ahmed K. Elmagarmid, *Privacy preserving association rule mining*, In Proceedings of the 12th International Workshop on Research Issues in Data Engineering (2002), 151–158.

[23] Jaideep Vaidya and Chris Clifton, *Privacy preserving association rule mining in vertically partitioned data*, In the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2002), 639–644.

[24] Vassilios S. Verykios, Ahmed K. Elmagarmid, Bertino Elisa, Yucel Saygin, and Dasseni Elena, *Association Rule Hiding*, IEEE Transactions on Knowledge and Data Engineering (2003), Accepted.