FI MU

Faculty of Informatics
Masaryk University Brno

Privacy Preserving Data Mining State-of-the-Art

by

Ondřej Výborný

FI MU Report Series **FIMU-RS-2006-06**

 September 2006

Publications in the FI MU Report Series are in general accessible via WWW:

`http://www.fi.muni.cz/reports/`

Further information can obtained by contacting:

**Faculty of Informatics
Masaryk University
Botanická 68a
602 00 Brno
Czech Republic**

Privacy Preserving Data Mining State-of-the-Art

Ondřej Výborný
KD Group, Faculty of Informatics, Masaryk University
xvyborny@fi.muni.cz

September 15, 2006

Abstract

The past few years have confirmed the increasing importance of the preservation of the privacy of an individual due to the number of methods by which private information can be misused. At the same time, since vast amounts of data are easily available, data mining is taking on a greater role in the processing thereof, placing techniques that can satisfy strict legal and public requirements on the processing of sensitive data in great demand. This paper presents the state-of-the-art in privacy preserving data mining and also proposes an alternative classification of privacy preserving techniques based on several points of view on privacy[1].

1 Introduction

Since vast amounts of data have to be processed nowadays, methods and techniques which will facilitate this work are needed. The field of knowledge discovery can provide us with such tools. In particular, when searching for information or patterns that are hidden in data, data mining (DM) is precisely the tool needed. DM techniques can find previously unknown patterns in data which are useful in many fields of interest, such as medicine, insurance, banking and information analysis in general.

Privacy preserving data mining (PPDM [49]) is, as a part of the DM, assuming a more important role in the whole area of knowledge discovery because of the increasing sensitivity of data that can be used for DM purposes. Despite the complexity of

[1]This work has been partially supported by the Grant Agency of the Czech Republic under the Grant No. MSM0021622418.

solving problems associated with preserving privacy, it is nevertheless crucial that solutions be found. We must prevent the misuse of private data from mobile phones, bank accounts, etc. We also need tools for the legal processing of such data, because of strict requirements in laws.These restrictions prohibit the revealing of sensitive information that can identify any individual. PPDM techniques should therefore avoid leakage of such information.

There are many problems that have to be solved in PPDM and some of them are caused by the lack of widely accepted definitions of privacy in this area as well as the complexity of privacy issues, having, as they do, many different aspects (e.g. not only privacy of individuals but also the privacy of other entities, such as businesses). As far as we know, no work on precise definitions of privacy terms has been done, except for [12], where some suggestions on how to specify privacy constraints are proposed together with metrics that can be used for defining and measuring privacy in DM. In this paper we propose several points of view that, as we believe, can be used to better understand some specific aspects of the meaning of privacy in the DM process.

This report will continue as follows. In section 2 we discuss the meanings of the term *privacy* in DM and in section 3 we show some basic specifications that can be used to categorise existing PPDM techniques. In section 4 we use our classification on several techniques that have been proposed over the few past years and in section 5 we present current directions in PPDM research. In section 6 we conclude this paper with a look to possible future directions of research in this area.

2 Privacy in Data Mining

Privacy (from the DM point of view) is not a simple feature. It has many different aspects which have to be considered in different situations. There is also the problem of balancing between the accuracy of the particular DM technique and the privacy it preserves. These two qualities cannot be both completely squared.

We considered that there are three basic views on privacy related to the part of the DM process where exactly we preserve privacy and we will discuss them in this section.

Data Point of View

In DM, we are dealing with many different types of data and some of them are sensitive. These sensitive data can be misused when revealed and therefore we have to preserve their privacy.

There are two basic ways of ensuring privacy of data in the data mining process. The first way is to exclude sensitive data from the database (DB) before we do the DM. This can be done by modifying or trimming the sensitive raw data in the DB. There are several techniques, such as perturbation (adding noise), blocking (deleting data from the DB), aggregation (or merging of parts of the DB), swapping (some values are simply swapped) or sampling (releasing only part of the DB for the DM) that are used for such data modification.

The second way is to exclude sensitive knowledge that could be possibly mined from the DB with some DM tool. In this case, we may want to exclude only a part of the mined information or the whole rules (in case of association rule mining), containing sensitive data. The second approach is much more difficult, because we need to find all possible relations between the data in our DB that can lead to revealing the sensitive data to the user and exclude all data necessary to avoid this leakage [44, 51, 14].

Owner's Point of View

When we have a DB and we want to offer its data for DM purposes, there are two possibilities how we can do it. Firstly, we can make the whole DB available for users (we do not need to keep the data secret). In this case we have no problems with privacy and users can do almost anything they want with the data. Secondly, we may need to keep the data secret but still offer the possibility of performing the DM (this means that any user can do the DM on our data, but the DB itself is not available [27]). This approach is reasonable when, e.g., we want to charge users for every access to our DB (because this cannot be done with the former approach, when the DB is available for a free use).

Another reason for preserving the privacy of the owner of the DB is that it could be possible to mine some information that will jeopardise privacy of the owner while still preserving the privacy of individuals, whose data were used for DM [24]. As an example, imagine research on several medical DBs belonging to different hospitals. It could be possible to mine information about the success rates of specific types of operations. If

any hospital has lower results (and this information does not necessary have a relation with the quality of service of that hospital), it could lead to a decrease in the number of patients who go to that hospital, which is not in the hospital's interest and they will not want to participate in such research.

User's Point of View

When users want to do DM on a DB they do not own, they may not want to reveal their queries to the DB owner (this is in some cases also sensitive information - imagine that you want to search some medical DB for information about your health without showing your queries to other people which could reveal some unwanted and maybe false information about your health). Therefore we have to preserve the privacy of their queries and, in some cases, it would be sensible to preserve even the identity of users (using some anonymization techniques). Almost anything that users have to do publicly in the DM process reveals some information about them or their interests, and we have to take this into account when trying to ensure privacy.

3 Classification of PPDM Techniques

PPDM techniques can be categorised in many ways. In this section we will show some possible characteristics which divide PPDM techniques into several groups. Some of these characteristics were described in previous work on the state-of-the-art in the PPDM [50] by V. S. Verykios et al. and we use some more which, as we believe, can help us to look at PPDM from another point of view. The categories we derived from our classification are not disjunct and we will not try to combine them to make an ideal and complete classification.

Classification According to Points of View on Privacy

PPDM techniques can be categorised according to the privacy points of view we have discussed in the previous section. These points of view are related to different parts of knowledge discovery and therefore we need to use different approaches when preserving privacy. The basic classification divides PPDM techniques into two groups – techniques that are used during the DM process and those that are used before or after the DM process:

- During the DM process
  these techniques are used to preserve the privacy from the user or the owner point of view and most of them use some cryptographic approach to achieve this goal.
- Before or after the DM process
  these techniques are related mostly to the data point of view and consist mainly of data modification techniques.

During the DM Process

Techniques used during the DM process do not change the data, but try to manipulate the DM process to avoid revealing sensitive knowledge that follows from the process itself. We deal here with privacy from the user or the owner point of view where, e.g., users don't want to reveal their queries and owners want to provide the database but not make it public. It is obvious that it is suitable to use these techniques when more than one party is involved in the DM process, i.e., when we need to do some distributed computations during the DM process.

It seems that techniques based on the cryptographic approach are sufficient to deal with this kind of task. Cryptography provides us with many useful tools which can help us to achieve our goal. Extremely useful are *secure multiparty computations* (SMCs [53, 21]). SMC is a computation between 2 or more parties based on their inputs. The goal is to conduct such computation without revealing anything else than the result and the particular input to each party. We can imagine an ideal SMC as a protocol with a trusted third party, where each participant sends its input to this third party and receives only the result of the computation. Actually, real SMCs do not necessarily need a trusted third party. There are several types of SMCs, such as secure sum, secure size of intersection, secure union or secure scalar product. These computations serve as basic blocks for larger PPDM protocols.

Before the DM Process

These techniques are used to manipulate the data used for DM to avoid revealing any sensitive information that is included in that data. They are sometimes called data obfuscation or obscuration techniques.

The first approach that can be used to achieve this goal is to mask real values of the data. The second approach includes blocking or deleting some parts of the data and

releasing only an incomplete database for DM. We present here a short survey of these techniques:

- Perturbation (adding noise)
  these techniques are used for changing values in the whole data set and are a typical example of the data obfuscation technique. Results that can be obtained after these techniques are used are not accurate but the level of uncertainty could be sufficient in many cases.

- Aggregation (merging)
  can be used to mask some values while leaving results of summarization functions unchanged.

- Swapping
  leaves values unchanged but affects their position in the DB and is therefore not usable for many DM algorithms.

- Blocking
  we can block some values and simply consider them as zeroes (or '?') in the DM process. This approach is useful when blocking of a small number of values is needed.

- Sampling
  means hiding bigger parts of the DB and revealing only the rest of it for DM purposes.

Another way of achieving privacy before the DM process is data transformation leading to *k-anonymity*[36, 45]. K-anonymity can be considered as a database property which holds when for any query to the database the answer will contain at least k items. The great advantage of this method is its good scalability. The better privacy is needed, the higher k is set.

After the DM Process

Techniques that are related to the "After the DM Process" part are mostly used for excluding sensitive information from results of the DM process. A typical example is *association rule hiding* where the whole association rule mined from DB is erased from

the result. However, this approach is not very usual and it is more often used for determining which sensitive information could be mined from the DB and what information should be excluded from the DB before the DM process starts to avoid it being revealed.

Other Types of Classification

There are also other ways of classification of PPDM techniques, such as data distribution. The data we want to use for DM can be centralised (stored in one database) or distributed, thus divided between several smaller databases. In this case our task is more difficult and we have to take into account possible threats following from this situation, such as unauthorized listening to communication between involved parties during the DM process.

The data distribution can be horizontal or vertical. In horizontal distribution each database contains a subset of rows of the original database, whereas in vertical distribution each database contains subset of columns of the original database. It is obvious that there are more problems with preserving privacy when using distributed databases and a lot of work has been done on both horizontally distributed [24] and vertically distributed [47] DM.

But the data distribution is not the only data property that influences the DM process. A different approach is also needed when dealing with different types of data, such as data streams [17]. When the temporal part of the data is important, we can treat such data as a time series which brings new relations and requirements for the analysis. Data streams also have the disadvantage of data incompleteness during computation, and this property makes the data mining task more difficult.

Another important criterion is the type of the DM algorithm we want to use - there is a big difference between, e.g., preserving privacy during association rule mining and during decision tree building. Each DM algorithm has its specific features and it is therefore much easier to make privacy preserving modification for each of them separately.

Previous work on state-of-the-art in PPDM [50] by V. S. Verykios et al. offered classification according to type of the privacy preserving method used for selective data modification. From this point of view, there are three types of PPDM techniques:

- Heuristic-based
  adaptive modification of data, minimization of utility loss of mined information after modifications.
- Cryptography-based
  mostly based on secure multiparty computations.
- Reconstruction-based
  reconstruction of original data distribution from the randomized data - obtaining a good result even if original values cannot be used for DM.

This classification is somewhat similar to our classification of PPDM techniques based on part of the KDD process where the technique is used. In our approach, techniques used before or after the DM process have a relation with heuristic-based and reconstruction-based techniques and techniques used during the DM process are similar to cryptography-based techniques.

Evaluation of PPDM Techniques

When we want to compare several PPDM techniques or algorithms, it is useful to have the ability to evaluate them. There are several criteria we can use for comparing them:

- Performance
  time requirements – computational and communication cost – a very good measure in the case of distributed computations.
- Data utility
  i.e., how accurate and useful the information can be which is mined from a DB after the application of a privacy preserving technique.
- Level of uncertainty
  or level of robustness – if we look at it another way, tells us the probability with which hidden information can still be predicted.
- resistance to different DM techniques
  can we mine any sensitive information using different DM technique after the process of privacy preservation?

These criteria seem to be a good choice for comparing different PPDM techniques, but as far as we know, no such comparison has been done on larger scale.

4 Categorization of Existing Techniques

In the previous section, we proposed possible ways of classification of PPDM techniques and in this section we will try to make a survey of recent results in PPDM research and apply our classification on the techniques mentioned.

Privacy in the context of DM got the attention of a larger research community around 1999 [8, 9]. However, we can profit from results obtained from many research areas, such as mathematical statistics, cryptography and of course knowledge discovery itself, that have a much longer history.

Techniques Used Before or After the DM Process

These techniques are often called data obfuscation (obscuration) and modification techniques because that is exactly what they are used for. They modify the data to mask or erase the original value that should not be revealed due to its high sensitivity. But even when we use these modification techniques we are still able to obtain good results with techniques that can reconstruct the original data from our results [2, 1, 43].

Techniques that are mostly used for data modification are based on some kind of perturbation (they are often called methods for noise addition) [32, 42, 33, 30]. There are also techniques using swapping [22] or blocking [10] of specific data.

Since selective data modification or sanitization is a very hard problem [4] (in this paper a formal proof, that optimal sanitization is an NP-hard problem for the hiding of sensitive large itemsets in the context of association rule mining, has been given), some heuristics have been developed for these problems.

We can look closer on, e.g, the problem of hiding association rules [14, 44, 51]. The basic situation is as follows. We have a DB D from which we can mine a set of association rules. But some of them are sensitive. How can we hide them? We can decrease their support by changing several values in DB D (thus obtaining a new DB D′), so large itemsets from which they were generated become infrequent. This is a solution with some side effects. By changing values in DB D, we can produce new frequent large itemsets and therefore make new association rules (referred to as "ghost rules"). These changes have impact on the utility of the new DB D′ and we have to balance this utility and the privacy we obtain.

If privacy scalability is one of the main qualities required, k-anonymity [36, 45, 46, 23] can be employed. Here the values in the DB are changed so that the answer for any

query contains at least k items. If a higher level of privacy is needed, the value of k is set higher. The name "k-anonymity" means that we are unable to identify an individual by any query (for example, if we ask a medical DB for a 20-year-old man with a lung cancer, the result will contain at least k such men).

Techniques Used During the DM Process

Current PPDM approaches are focused mainly on cryptography-based techniques based on SMC algorithms which offer almost perfect privacy and accuracy, because we can use unchanged, original values in the DM process. It was shown that this approach can be applied to any function that has an efficient representation as a circuit [40]. But nothing is ideal, so this approach has several practical disadvantages too. One of them is its impracticality when too many participants are involved in computation and they have to work synchronously. We can sometimes overcome these difficulties by involving third parties but this could bring other security problems.

In current research, a big effort is being made on developing PPDM techniques for distributed data [13, 24, 30]. As a practical example, we can take system DIODA [16] for mining association rules on horizontally distributed data. This system is based only on the semi-honest model where no active malicious action is allowed and it is therefore usable only for research purposes now. Nevertheless, the issue of distributed PPDM techniques is becoming more and more important and we can expect other software available to perform such computations in the real environment in the near future.

PPDM techniques used during the DM process can also be categorised according to the different DM algorithms they use. There are techniques for association rule mining [24, 16], techniques for decision tree learning [15], clustering [48, 18], SVM (support vector machines) [55, 54] or naive bayes classifiers and computations on bayes network structure [25, 52].

Basic Protocols for PPDM

Most algorithms used nowadays are built on several basic protocols, such as Oblivious Transfer (OT) or Private Information Retrieval (PIR) and we will briefly show their main ideas here.

OT Protocol

Oblivious Transfer protocol [41, 35, 40] is a secure protocol between two parties. One party (the server) holds a secret bit b and the second party (the user) can, at the end of the computation, learn this bit b with a probability of $1/2$ or learn nothing. The main thing is that the server has no knowledge about which of these two events has happened and the user does not know anything about other bits in the DB. This is a simple example of 1-out-of-2 OT protocol. There are also other modifications (1-out-of-N OT, distributed OT [34], etc.) of this protocol that are used for solving various security problems.

PIR Protocol

Private Information Retrieval protocols [6] allow users to retrieve information from a DB while keeping their query private. It can be seen that these protocols are similar to OT protocols. The difference is that in the PIR protocol a user can learn more values than just the one requested. A trivial example of the PIR protocol is sending the whole DB to the user. However, this approach has the communication complexity $c(n) = n$, where n is the size of the DB. A PIR protocol with $c(n) < n$ is called *non-trivial* and many such protocols are designed using several identical DBs (a user makes different requests on such databases and reconstructs the desired information from results). An example of the PIR protocol [11] uses this protocol for private storing users files on a remote file server. It allows a user to retrieve some encrypted files containing specific keywords while keeping these keywords secret.

5 Directions in Current PPDM

A large area aimed at by researchers nowadays is improving the efficiency of current privacy preserving algorithms, in particular the communication cost, which is a bottleneck for many PPDM algorithms [3].

A large part of data acquired nowadays has some sort of time component crucial for its meaning and usability. This component can be found in all stream data (including data concerning network traffic, mobile phone location, signals measured on different devices, etc.) and great efforts have been made to develop techniques suitable for such data [17]. Nevertheless, privacy issues regarding this type of data have not yet been

properly addressed. This has to be done due to increasing quantity [20] of such types of data and improper existing algorithms. Only a few papers on these issues have been published, such as [39, 7], where techniques for private search in a stream of documents were proposed. In both papers, the user is allowed to obtain documents corresponding to his private keywords. These keywords stay encrypted during the whole process so only the user knows them.

There are a few large projects, where PPDM could be utilised, which deal with ubiquitous data [26] or some other sources that provide vast amounts of data (mobile phone networks) [19].

However, the current research is not focused only on inventing new methods and techniques for PPDM, but there are also topics related to general research, such as the work of Oliveira and Zaïane [38], where a framework for PPDM standardization is proposed.

Internet Sources

To have an overall survey of current PPDM techniques, it is useful to know about available internet sources providing up-to-date information about this area of research.

There are several web sites about privacy preserving data mining, such as Helger Lipmaa's site "Privacy-Preserving Data mining" [28], Kun Liu's site "Privacy Preserving Data Mining Bibliography" [29] or Stanley Oliveira's "The Privacy, Security and Data Mining Site" at the University of Alberta [37], where it is possible to find links on the most important PPDM papers, researchers or conferences dealing with this area of research. Also more specialised web sites such as Nina Mishra's "A Study of Perturbation Techniques for Data Privacy" [31], which provides a good source of information about perturbation techniques or Mohamed Medhat Gaber's site "Mining Data Streams Bibliography" [17] providing a survey of papers about stream mining, can be found on the Internet.

6 Conclusions

The past few years have confirmed the increasing importance of the preservation of individuals' privacy due to the number of methods by which private information can be misused and the vast amounts of data cointaining such information that are currently available.

We presented here the state of the art in privacy preserving data mining, which provides us with techniques capable of satisfying the legal and public requirements constraining the processing of our sensitive data.

Since new types of data are rapidly emerging, adapting of existing techniques to such data is needed. One of the main common traits of current types of data is their temporal behaviour. A typical example of such data are data streams, where the meaning can be lost without knowing the precise time when particular segments of data stream were acquired. Therefore, techniques that can properly process such data are highly needed.

Also the meaning of the term "privacy" can change slightly depending on different data, and therefore understanding to all its aspects is important. In this paper, we proposed several different points of view on privacy in data mining in order to ease comprehension of this complex notion.

It seems that importance of privacy preserving methods will not decrease since people are becoming more concerned about the misuse of their private information. This trend implies that research in this area is quite forward-looking.

References

[1] Dakshi Agrawal and Charu C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the 20th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 247–255, Santa Barbara, California, USA, 2001.

[2] Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, pages 439–450, Dallas, Texas, May 2000. ACM Press.

[3] Shipra Agrawal, Vijay Krishnan, and Jayant R. Haritsa. On addressing efficiency concerns in privacy-preserving mining. In Yoon-Joon Lee, Jianzhong Li, Kyu-Young Whang, and Doheon Lee, editors, *Proceedings of the 9th International Conference on Database Systems for Advanced Applications DASFAA*, volume 2973 of *Lecture Notes in Computer Science*, pages 113–124. Springer, 2004.

[4] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. Verykios. Disclosure limitation of sensitive rules. In *Proceedings of 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99)*, pages 45–52, Chicago, IL., November 1999.

[5] Roger S. Barga and Xiaofang Zhou, editors. *Proceedings of the 22nd International Conference on Data Engineering Workshops, ICDE 2006, 3-7 April 2006, Atlanta, GA, USA*. IEEE Computer Society, 2006.

[6] A. Beimel. Private information retrieval. `http://www.cs.bgu.ac.il/~beimel/Research/PIR.html`, July 2006.

[7] John Bethencourt, Dawn Song, and Brent Waters. New constructions and practical applications for private stream searching (extended abstract). In *Proceedings of Security and Privacy, 2006 IEEE Symposium on*, pages 132–139, 2006.

[8] L. Brankovic and V. Estivill-Castro. Privacy issues in knowledge discovery and data mining. In *Proceedings of the Australian Institute of Computer Ethics Conference (AICEC99)*, Melbourne, Australia, July 1999.

[9] A. J. Broder. Data mining, the internet, and privacy. In *Proceedings of the Web Usage Analysis and User Profiling, International WEBKDD'99 Workshop*, San Diego, California, USA, August 1999.

[10] LiWu Chang and Ira S. Moskowitz. An integrated framework for database privacy protection. In *Proceedings of the IFIP Workshop on Database Security*, pages 161–172, The Netherlands, 2000.

[11] Y. Chang and M. Mitzenmacher. Privacy preserving keyword searches on remote encrypted data. Technical report, Cryptology ePrint Archive, Report 2004., 2004.

[12] C. Clifton, M. Kantarcioglu, and J. Vaidya. Defining privacy for data mining. In *Proceedings of the National Science Foundation Workshop on Next Generation Data Mining*, pages 126–133, Baltimore, MD., November 2002. Invited paper.

[13] Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya, Xiaodong Lin, and Michael Y. Zhu. Tools for privacy preserving data mining. *SIGKDD Explorations*, 4(2):28–34, 2002.

[14] Elena Dasseni, Vassilios S. Verykios, Ahmed K. Elmagarmid, and Elisa Bertino. Hiding association rules by using confidence and support. In *Proceedings of the Information Hiding: 4th International Workshop, IHW 2001*, volume 2137 of *Lecture Notes in Computer Science*, Pittsburgh, PA, USA, 2001.

[15] W. Du and Z. Zhan. Building decision tree classifier on private data. In *Proceedings of the IEEE International Conference on Data Mining, Workshop on Privacy, Security, and Data Mining, ICDM'02*, pages 1–8, Maebashi City, Japan, December 2002.

[16] Jan Frieser and Lubomír Popelínský. Secure mining in horizontally partitioned data. In *Proceedings of the Workshop on Privacy and Security Issues in Data Mining, ECML/PKDD*, pages 10–20, Pisa, Italy, 2004.

[17] Mohamed Medhat Gaber. Mining data streams bibliography. `http://www.csse.monash.edu.au/~mgaber/WResources.htm`, July 2006.

[18] Krishnan Pillaipakkamnatt Geetha Jagannathan and Rebecca N. Wright. A new privacy-preserving distributed k-clustering algorithm. In *Proceedings of the 2006 SIAM International Conference on Data Mining (SDM)*, 2006.

[19] GeoPKDD - geographic privacy-aware knowledge discovery and delivery. `http://www.di.unipi.it/~rinziv/prin/index.php`, July 2006.

[20] Bobi Gilburd, Assaf Schuster, and Ran Wolff. A new privacy model and association-rule mining algorithm for large-scale distributed environments. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, Seattle, WA, 2004.

[21] O. Goldreich, S. Micali, and A. Wigderson. How to play any mental game. In *STOC '87: Proceedings of the nineteenth annual ACM conference on Theory of computing*, pages 218–229, New York, NY, USA, 1987. ACM Press.

[22] Shanti Gomatam, Alan F. Karr, and Ashish P. Sanil. Data swapping as a decision problem. Technical report, National Institute of Statistical Sciences, October 2004.

[23] Wei Jiang and Chris Clifton. Privacy-preserving distributed k-anonymity. In Sushil Jajodia and Duminda Wijesekera, editors, *Proceedings of the DBSec*, volume 3654 of *Lecture Notes in Computer Science*, pages 166–177. Springer, 2005.

[24] Murat Kantarcioglu and Chris Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. *IEEE Transactions on Knowledge and Data Engeneering*, 16(9):1026–1037, 2004.

[25] Murat Kantarcioglu and Jaideep Vaidya. Privacy preserving naive bayes classifier for horizontally partitioned data. In *Proceedings of the IEEE ICDM Workshop on Privacy Preserving Data Mining*, pages 3–9, Melbourne, Florida, USA, November 2003.

[26] KDubiq - Knowledge Discovery in Ubiquitous Environments. `http://www.kdubiq.org`, September 2006.

[27] Sven Laur, Helger Lipmaa, and Taneli Mielikäinen. Private itemset support counting. In Sihan Qing, Wenbo Mao, Javier Lopez, and Guilin Wang, editors, *Proceedings of the ICICS*, volume 3783 of *Lecture Notes in Computer Science*, pages 97–111. Springer, 2005.

[28] Helger Lipmaa. Privacy-preserving data mining. `http://www.cs.ut.ee/~lipmaa/crypto/link/data_mining/`, July 2006.

[29] Kun Liu. Privacy preserving data mining bibliography. `http://www.csee.umbc.edu/~kunliu1/research/privacy_review.html`, July 2006.

[30] Kun Liu, Hillol Kargupta, and Jessica Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on Knowledge and Data Engeneering*, 18(1):92–106, 2006.

[31] Nina Mishra. A study of perturbation techniques for data privacy. `http://theory.stanford.edu/~nmishra/cs369-2004.html`, January 2006.

[32] Krishnamurty Muralidhar and Rathindra Sarathy. Security of random data perturbation methods. *ACM Trans. Database Syst.*, 24(4):487–493, 1999.

[33] Krishnamurty Muralidhar and Rathindra Sarathy. A theoretical basis for perturbation methods. *Statistics and Computing*, 13(4):329–335, 2003.

[34] Moni Naor and Benny Pinkas. Distributed oblivious transfer. In Tatsuaki Okamoto, editor, *Proceedings of the ASIACRYPT*, volume 1976 of *Lecture Notes in Computer Science*, pages 205–219. Springer, 2000.

[35] Moni Naor and Benny Pinkas. Computationally secure oblivious transfer. *Cryptology*, 18(1):1–35, 2005.

[36] M. Ercan Nergiz and Chris Clifton. Thoughts on k-anonymization. In Barga and Zhou [5], page 96.

[37] Stanley Oliveira. The privacy, security and data mining site. `http://www.cs.ualberta.ca/~oliveira/psdm/psdm_index.html`, January 2006.

[38] Stanley Oliveira and Osmar R. Zaïane. Toward standardization in privacy-preserving data mining. In *Proceeding of the 3rd Workshop on Data Mining Standards (DM-SSP 2004), in conjunction with KDD 2004*, Seattle, WA, USA, August 2004.

[39] Rafail Ostrovsky and William E. Skeith III. Private searching on streaming data. In Victor Shoup, editor, *Proceedings of the CRYPTO*, volume 3621 of *Lecture Notes in Computer Science*, pages 223–240. Springer, 2005.

[40] Benny Pinkas. Cryptographic techniques for privacy-preserving data mining. *SIGKDD Explorations*, 4(2):12–19, 2002.

[41] Benny Pinkas. Oblivious transfer. `http://www.pinkas.net/ot.html`, January 2006.

[42] H. Polat and W. Du. Privacy-preserving collaborative filtering using randomized perturbation techniques. In *Proceedings of the 3rd IEEE International Conference on Data Mining ICDM'03*, Melbourne, FL, November 2003.

[43] Shariq Rizvi and Jayant R. Haritsa. Maintaining data privacy in association rule mining. In *Proceedings of the VLDB*, pages 682–693. Morgan Kaufmann, 2002.

[44] Yücel Saygin, Vassilios S. Verykios, and Chris Clifton. Using unknowns to prevent discovery of association rules. *SIGMOD Record*, 30(4):45–54, 2001.

[45] Latanya Sweene. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):571–588, 2002.

[46] Traian Marius Truta and Bindu Vinay. Privacy protection: p-sensitive k-anonymity property. In Barga and Zhou [5], page 94.

[47] Jaideep Vaidya. *Privacy Preserving Data Mining Over Vertically Partitioned Data*. PhD thesis, Purdue University, 2004. UMI Order Number: AAI3154746.

[48] Jaideep Vaidya and Chris Clifton. Privacy-preserving k-means clustering over vertically partitioned data. In Lise Getoor, Ted E. Senator, Pedro Domingos, and Christos Faloutsos, editors, *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 206–215, Washington, D.C., August 24 - 27 2003. ACM.

[49] Jaideep Vaidya, Christopher W. Clifton, and Yu Michael Zhu. *Privacy Preserving Data Mining*. Springer, 1st edition, November 2005.

[50] Vassilios S. Verykios, Elisa Bertino, Igor Nai Fovino, Loredana Parasiliti Provenza, Yücel Saygin, and Yannis Theodoridis. State-of-the-art in privacy preserving data mining. *SIGMOD Record*, 33(1):50–57, March 2004.

[51] Vassilios S. Verykios, Ahmed K. Elmagarmid, Elisa Bertino, Yücel Saygin, and Elena Dasseni. Association rule hiding. *IEEE Transactions on Knowledge and Data Engeneering*, 16(4):434–447, April 2004.

[52] Rebecca N. Wright and Zhiqiang Yang. Privacy-preserving bayesian network structure computation on distributed heterogeneous data. In Won Kim, Ron Kohavi,

Johannes Gehrke, and William DuMouchel, editors, *Proceedings of the KDD*, pages 713–718. ACM, 2004.

[53] Andrew Chi-Chih Yao. How to generate and exchange secrets (extended abstract). In *Proceedings of the FOCS*, pages 162–167. IEEE, 1986.

[54] Hwanjo Yu, Xiaoqian Jiang, and Jaideep Vaidya. Privacy-preserving SVM using nonlinear kernels on horizontally partitioned data. In *SAC '06: Proceedings of the 2006 ACM symposium on Applied computing*, pages 603–610, New York, NY, USA, 2006. ACM Press.

[55] Hwanjo Yu, Jaideep Vaidya, and Xiaoqian Jiang. Privacy-preserving SVM classification on vertically partitioned data. In Wee Keong Ng, Masaru Kitsuregawa, Jianzhong Li, and Kuiyu Chang, editors, *PAKDD*, volume 3918 of *Lecture Notes in Computer Science*, pages 647–656. Springer, 2006.