Volume 5, Issue 10, October-2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Survey on Privacy Preserving Data Mining

A. V. Sriharsha*, **Dr. C. Parthasarathy**

Department of Information Technology, SCSVMV University,
Tamilnadu, India

Abstract— ***The collection of digital information by governments, corporations, and individuals has created tremendous opportunities for knowledge- and information-based decision making. Driven by mutual benefits, or by regulations that require certain data to be published, there is a demand for the exchange and publication of data among various parties. Data in its original form, however, typically contains sensitive information about individuals, and publishing such data will violate individual privacy. Protecting private data is an important concern for society, several laws now require explicit consent prior to analysis of an individual's data, but its importance is not limited to individuals: corporations might also need to protect their information's privacy, even though sharing it for analysis could benefit the company. Clearly, the trade-off between sharing information for analysis and keeping it secret to preserve corporate trade secrets and customer privacy is a growing challenge. A variety of methods and approaches have been introduced and developed for privacy preserving data mining. This paper provides a basic survey of different privacy preserving data mining methods and analyses the representative techniques for privacy preserving data mining.***

Keywords— ***PPDM, Cryptography, Differential Privacy, Randomization, K-Anonymity, Data perturbation, ARM.***

I. INTRODUCTION

The scope of information technologies and the internet in the past two decades has brought a wealth of individual information into the hands of commercial companies and government agencies. As hardware costs go down, organizations and it easier than ever to keep any piece of information acquired from the ongoing activities of their clients. Data owners constantly seek to make better use of the data they possess, and utilize data mining tools to extract useful knowledge and patterns from the data.

In recent years, data mining has been viewed as a threat to privacy because of the widespread proliferation of electronic data maintained by corporations. This has lead to increased concerns about the privacy of the underlying data. In recent years, a number of techniques have been proposed for modifying or transforming the data in such a way so as to preserve privacy. Privacy-preserving data mining finds numerous applications in surveillance which are naturally supposed to be "privacy-violating" applications. The key is to design methods which continue to be effective, without compromising security. Most methods for privacy computations use some form of transformation on the data in order to perform the privacy preservation. Typically, such methods reduce the granularity of representation in order to reduce the privacy. This reduction in granularity results in some loss of effectiveness of data management or mining algorithms.

PPDM is a research area concerned with the privacy driven from personally identifiable information when considered for data mining. This work addresses the privacy problem by considering the privacy and algorithmic requirements simultaneously. The objective of this work is to implement a distortion algorithm using association rule hiding for privacy preserving data mining which would be efficient in providing confidentiality and improve the performance (Charu C. Aggarwal et al. 2008).

Privacy Violation in Data Mining Understanding privacy in data mining requires understanding how privacy can be violated and the possible means for preventing privacy violation. In general, one major factor contributes to privacy violation in data mining: the misuse of data. Users' privacy can be violated in different ways and with different intentions. Although data mining can be extremely valuable in many applications (e.g., business, medical analysis, etc), it can also, in the absence of adequate safeguards, violate informational privacy. Privacy can be violated if personal data are used for other purposes subsequent to the original transaction between an individual and an organization when the information was collected (Culnan, 1993). One of the sources of privacy violation is called data magnets (Rezgui et al., 2003). Data magnets are techniques and tools used to collect personal data. Examples of data magnets include explicitly collecting information through on-line registration, identifying users through IP addresses, software downloads that require registration, and indirectly collecting information for secondary usage. In many cases, users may or may not be aware that information is being collected or do not know how that information is collected. In particular, collected personal data can be used for secondary usage largely beyond the users' control and privacy laws. This scenario has led to an uncontrollable privacy violation not because of data mining itself, but fundamentally because of the misuse of data.

II. RANDOMIZATION

The randomization method has been traditionally used in the context of distorting data by probability distribution for methods such as surveys which have an evasive answer bias because of privacy concerns. This technique has also been extended to the problem of privacy preserving data mining.

The method of randomization can be described as follows. Consider a set of data records denoted by $\boldsymbol{X} = \{x_1 \ldots x_N\}$. For record $x_i \in \boldsymbol{X}$, we add a noise component which is drawn from the probability distribution $f_Y(y)$. These noise components are drawn independently, and are denoted $y_1 \ldots y_N$. Thus, the new set of distorted records are denoted by $x_1 + y_1 \ldots x_N + y_N$. We denote this new set of records by $z_1 \ldots z_N$. In general, it is assumed that the variance of the added noise is large enough, so that the original record values cannot be easily guessed from the distorted data. Thus, the original records cannot be recovered, but the distribution of the original records can be recovered. Thus, if $\boldsymbol{X}$ be the random variable denoting the data distribution for the original record, Y be the random variable describing the noise distribution, and $\boldsymbol{Z}$ be the random variable denoting the final record, we have:

$$Z = X + Y$$
$$X = Z - Y$$

Now, we note that N instantiations of the probability distribution $\boldsymbol{Z}$ are known, whereas the distribution $\boldsymbol{Y}$ is known publicly. For a large enough number of values of $\boldsymbol{N}$, the distribution $\boldsymbol{Z}$ can be approximated closely by using a variety of methods such as kernel density estimation. By subtracting Y from the approximated distribution of $\boldsymbol{Z}$, it is possible to approximate the original probability distribution $\boldsymbol{X}$. We note that at the end of the process, we only have a distribution containing the behavior of $\boldsymbol{X}$. Individual records is not available. Furthermore, the distributions are available only along individual dimensions. Therefore, new data mining algorithms need to be designed to work with the uni-variate distributions rather than the individual records. This can sometimes be a challenge, since many data mining algorithms are inherently dependent on statistics which can only be extracted from either the individual records or the multi-variate probability distributions associated with the records.

One key advantage of the randomization method is that it is relatively simple, and does not require knowledge of the distribution of other records in the data. This is not true of other methods such as k-anonymity which require the knowledge of other records in the data. Therefore, the randomization method can be implemented at data collection time, and does not require the use of a trusted server containing all the original records in order to perform the anonymization process. The randomization method has been extended to a variety of data mining problems.

III. PRIVACY QUANTIFICATION

The quantity used to measure privacy should indicate how closely the original value of an attribute can be estimated. A key privacy measure is based on the differential entropy of a random variable. The differential entropy h(A) of a random variable A is defined as follows:

$$h(\mathrm{A}) = - \int_{\Omega_A} f_A(a) \log_2 f_A(a)\, da$$

The privacy level may be quantified using various parameters. Such parameters include, perturbation mechanism, data privacy achieved, techniques used for anonymization, disclosure risk. The privacy level of such method is assessed by using the disclosure risk, that is, the risk that a piece of information be linked to a specific individual. There are several approaches to measure the disclosure risk. One approach is based on the computation of the distance-based record linkage. An intruder is assumed to try to link the masked dataset with the external dataset using the key variables. The distance between records in the original and the masked datasets is computed. A record in the masked dataset is labelled as “linked” or “linked to 2nd nearest” if the nearest or 2nd nearest record in the original dataset turns out to be the corresponding original record. Then the disclosure risk is computed as the percentage of “linked” and “linked to 2nd nearest”. The second approach is based on the computation of the probabilistic record linkage. The linear sum assignment model is used to ‘pair’ records in the original file and the masked file. The percentage of correctly paired records is a measure of disclosure risk.

IV. DATA MODIFICATION TECHNIQUES

Data Modification Techniques Data modification 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.

Noise Addition Techniques: 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. We categorize noise addition techniques into three groups: **(1)** data swapping techniques that interchange the values of individual records in a database (Estivill-Castro & Brankovic, 1999); **(2)** data distortion techniques that perturb the data to preserve privacy, and the distorted data maintain the general distribution of the original data (Agrawal & Srikant, 2000); and **(3)** data randomization techniques which allow one to perform the discovery of general patterns in a database with error bound, while protecting individual values. Like data swapping and data distortion techniques, randomization techniques are designed to find a good compromise between privacy protection and knowledge discovery (Evfimievski et al., 2002; Rizvi & Haritsa, 2002; Zang, Wang, & Zhao, 2004).

Space Transformation Techniques: These techniques are specifically designed to address privacy-preserving clustering. These techniques are designed to protect 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. We categorize space transformation techniques into two major groups: (1) object similarity-based representation relies on the idea behind the similarity between objects, i.e., a data owner could share some data for clustering analysis by simply computing the dissimilarity matrix (matrix of distances) between the objects and then sharing such a matrix with a third party. Many clustering algorithms in the literature operate on a dissimilarity matrix (Han & Kamber, 2001). This solution is simple to be implemented and is secure, but requires a high communication cost (Oliveira & Zaïane, 2004); (2) dimensionality reduction-based transformation can be used to address privacy-preserving clustering when the attributes of objects are available either in a central repository or vertically partitioned across many sites. By reducing the dimensionality of a dataset to a sufficiently small value, one can find a trade-off between privacy, communication cost, and accuracy. Once the dimensionality of a database is reduced, the released database preserves (or slightly modifies) the distances between data points. In tandem with the benefit of preserving the similarity between data points, this solution protects individuals' privacy since the attribute values of the objects in the transformed data are completely different from those in the original data (Oliveira & Zaïane, 2004).

V. CRYPTOGRAPHY BASED TECHNIQUES

The cryptography-based technique usually guarantees very high level of data privacy. In their work, Kantarcioglu and Clifton address the problem of secure mining of association rules over horizontally partitioned data, using cryptographic techniques to minimize the information shared. Their solution is based on the assumption that each party first encrypts its own itemsets using commutative encryption, then the already encrypted itemsets of every other party. Later on, an initiating party transmits its frequency count, plus a random value, to its neighbour, which adds its frequency count and passes it on to other parties. Finally, a secure comparison takes place between the final and initiating parties to determine if the final result is greater than the threshold plus the random value.

Another cryptography-based approach is described in the works of Tamir Tassa et. al. Such approach addresses the problem of association rule mining in vertically partitioned data. In other words, its aim is to determine the item frequency when transactions are split across different sites, without revealing the contents of individual transactions. The security of the protocol for computing the scalar product is analysed. Though cryptography-based techniques can well protect data privacy, they may not be considered good with respect to other metrics like efficiency.

In the context of PPDM over distributed data, cryptography-based techniques have been developed to solve problem 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 (SMC) problem (Goldreich, Micali, & Wigderson, 1987). The technique proposed in (Lindell & Pinkas, 2000) address privacy-preserving classification, while the techniques proposed in (Kantarcioğlu & Clifton, 2002; Vaidya & Clifton, 2002) address privacy-preserving association rule mining, and the technique in (Vaidya & Clifton, 2003) addresses privacy-preserving clustering.

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 only willing to share data mining results, not the source data. In these cases, the data are distributed either horizontally or vertically. 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.

VI. PRIVACY IN DATA MINING

PPDM is usually carried out in multiple steps. First, the data being mined are collected from their sources, which are referred as data providers. In many systems, data providers are physically distributed, forming the bottom tier of the architecture of data mining systems, as shown in Figure 1.1. It shows privacy-preserving data mining usually has multiple steps that translate to a three-tiered architecture. The bottom tier has the data providers, the data owners, which are often physically distributed. The data providers submit their private data to the data warehouse server. This server, which constitutes the middle tier, supports online analytical data processing to facilitate data mining by translating raw data from the data providers into aggregate data that the data mining servers can more quickly process (Nan Zhang et. al 2007).

In general, privacy preservation occurs in two major dimensions: users' personal information and information concerning their collective activity. We refer to the former as individual privacy preservation and the latter as collective privacy preservation, which is related to corporate privacy in (Clifton et al., 2002).

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

Collective privacy preservation: Protecting personal data may not be enough. Sometimes, we may need to protect against learning sensitive knowledge representing the activities of a group. We refer to the protection of sensitive knowledge as collective privacy preservation. The goal here is quite similar to that one for statistical databases, in which security control mechanisms provide aggregate information about groups (population) and, at the same time, prevent disclosure of confidential information about individuals. However, unlike as is the case for statistical databases, another objective of collective privacy preservation is to protect sensitive knowledge that can provide competitive advantage in the business world.

Secure multi-party computation: An alternative approach based on the multiparty computation is that every part of private data is validly known to one or more parties. Revealing private data to parties such as by whom the data is owned or the individual to whom the data refers to is not a condition of violating privacy. The problem arises when the private information is revealed to some other third parties. To deal with this problem, we use a specialized form of privacy preserving distributed data mining. Parties that each knows some of the private data participate in a protocol that generates the data mining results, that guarantees no data items is revealed to other parties. Thus the process of data mining doesn't cause, or even increase the opportunity for breach of privacy.

Sequential pattern hiding method is necessary to conceal sensitive patterns that can otherwise be extracted from published data, without seriously affecting the data and the non sensitive interesting patterns. Sequential pattern hiding is a challenging problem, because sequences have more composite semantics than item sets, and calls for efficient solutions that offer high utility.

A. Privacy Preserving in Data Mining Results:

The first direction is to anonymize the data so that sensitive data mining patterns cannot be generated. Aggarwal et al. [2006] pointed out that simply suppressing the sensitive values chosen by individual record owners is insufficient because an attacker can use association rules learnt from the data to estimate the suppressed values. They proposed a heuristic algorithm to suppress a minimal set of values to combat such attacks. Verykios et al. [2004] proposed algorithms for hiding sensitive association rules in a transaction database. The general idea is to hide one rule at a time by either decreasing its support or its confidence, achieved by removing items from transactions. Rules satisfying a specified minimum support and minimum confidence are removed. However, in the notion of anonymity, a rule applying to a small group of individuals (i.e., low support) presents a more serious threat because record owners from a small group are more identifiable. The second direction is to directly anonymize the data mining patterns. Atzori et al. [2008] proposed the insightful suggestion that if the goal is to release data mining results, such as frequent patterns, then it is sufficient to anonymize the patterns rather than the data. Their study suggested that anonymizing the patterns yields much better information utility than performing data mining on anonymous data. This opens up a new research direction for privacy-preserving patterns publishing. Kantarcioglu et al. [2004] defined an evaluation method to measure the loss of privacy due to releasing data mining results.

B. Privacy Preserving in Distributed Data Mining:

Privacy-preserving distributed data mining (PPDDM) is a cousin to the research topic of privacy-preserving data publishing (PPDP). PPDDM assumes a scenario that multiple data holders want to collaboratively perform data mining on the union of their data without revealing their sensitive information. PPDDM usually employs cryptographic solutions. Although the ultimate goal of both PPDDM and PPDP is to perform data mining, they have very different assumptions on data ownerships, attack models, privacy models, and solutions, so PPDDM is out of the scope of this survey. We refer readers interested in PPDDM to work by [Clifton et al. 2002; Kantarcioglu 2008; Pinkas 2002; Vaidya 2008; Wright et al. 2005].

VII. SENSITIVE INFORMATION HIDING TECHNIQUES

There is a twofold consideration in Privacy Preserving Data Mining. The first is sensitive raw data that are kept secure from unauthorized access like identifiers, names, addresses should be modified from original database in order for receiver of data not be able to compromise another person's privacy. The second one is sensitive knowledge is excluded that can be mined from a database by using data mining algorithms as such type of knowledge compromises data privacy.

The purchasing pattern of one product along with another product represents an association rule. Association rules are frequently used by retail stores to support in marketing advertisement and inventory control. Where the information about the relative items is revealed from the rules. The technique that is used to reduce the guessability from another item is implemented by Data distortion. Data distortion is done by the alteration of an attribute value by a new value. There are two basic approaches for rule hiding in data distortion based technique. Reduce the confidence of rules and reduce the support of rules.

Selecting minimum support = 20% and minimum confidence = 80% and applying association rule mining algorithm, two association rules AB → C (confidence = 100%) and BC→ A (confidence= 100%) are mined. Now suppose rule AB→ C is sensitive and needs to be hidden. Decreasing the confidence of a rule AB→ C can be done by either increasing the support of AB in transactions not supporting C or by decreasing the support of C in transactions supporting both AB and C. Decreasing support of rule AB→ C can be done by decreasing the support of the corresponding large itemset ABC. The problem for finding an optimal sanitization to a database against association rule analysis has been proven to be NP-Hard. In this work proposes many ideas for the development of new research problems in privacy preserving data mining, not only using association rules but also classification and clustering.

VIII. DIFFERENTIAL PRIVACY

Differential privacy proposed by Cynthia D'Work, is a rigorous privacy guarantor that works interactively and non-interactively. Differential Privacy is a relatively new privacy ensuring mechanism, but as the number and volume of databases with private data keep on to grow, this will continue to be a powerful and important tool. The main goal of using the differential privacy techniques is to provide one of the strongest privacy grantees in the problem of private data publishing. As the usage of data mining for potential intrusive purposes using personally identifiable information increases, privately using these results will become more important. A data miner can question through a private mechanism and a database holder answers these queries in response, the multiple queries posed by the data miner the data base answers more, comparatively, the results of the queries are tested and the stability of the privacy algorithm induced in answering the query is proved. The objective in this method is to determine the most effective privacy preservation algorithm for a given database. The interactive approach is focus on question database-answering, are not gladly applicable to PPDP, where the data publisher may not have complicated database management knowledge, not want to provide an interface for database question. A data publisher, such as a hospital, has no purpose of being a database server; answering database queries is not part of its normal business.

In a non-interactive approach, a database holder first anonymized the raw data and then releases the anonymized version for data analysis. Just the once the data are published. The data holder has no more control over the published data. The non- interactive means nothing but the data are sanitized and then release the data. The non-interactive approach also referred in a privacy preserving distributed data mining (PPDDM). In non-interactive approach in PPDDM, allows anonymizing data from different source of data release without revealing the sensitive information. The non-interactive algorithm is to securely integrate horizontally partitioned data from multiple data holders without disclosing data from one party to another. The non- interactive query model is a statistical disclosure control, in which the data recipient can scan and submit one query to the system. This kind of non-interactive query model may not fully deal with the information needs of data recipients because, in some other cases, it is very complicated for a data recipient to exactly construct a query for a data mining task in one shot.

IX. ATTACKS ON PERTURBATIONS

Data perturbation represents one common approach in privacy preserving data mining (PPDM). The technique of perturbation is from the history having made significant studies and experiments in statistical disclosure control and statistical databases, where the original dataset is perturbed and the result is released for data analysis. Typically, a privacy/accuracy trade-off is faced, on the one hand, perturbation must not allow the original data records to be adequately recovered and on other hand it must allow patterns in the original data to be mined. Facing this challenge, data perturbation includes a wide variety of techniques including additive, multiplicative, matrix multiplicative, k-anonymization, micro-aggregation, categorical dat perturbation, data swapping, resampling, data shuffling and many more. SVD (Singular-Value Decomposition) filtering, PCA (Principal Component Analyses) Filtering are a kind of a method that can guess the values in the quasi attributes of the anonymized results. MAP Estimation Attack uses Eigen-analysis by considering prior and posterior knowledge via Bayes theorem to estimate the original dataset. Distribution Analysis Attack estimates the original data set from the similarity of distributions in the perturbed data and the guessed noise. The possible improvement on additive perturbation is to use colored noise with similar correlation structure to the original data, where the guessing of the original data will be difficult. Euclidean distances are used to attack on the data perturbed using matrix multiplicative perturbation methods, where many important data mining algorithms are applied to the perturbed data and produce results very similar to, or exactly the same as those produced by the original algorithm applied to the original data.

X. SUMMARY AND FUTURE RESEARCH DIRECTIONS

Information sharing has become part of the routine activity of many individuals, companies, organizations, and government agencies. Privacy-preserving data publishing is a promising approach to information sharing, while preserving individual privacy and protecting sensitive information. In this survey, we reviewed the recent developments in the field. The general objective is to transform the original data into some anonymous form to prevent from inferring its record owners' sensitive information. We presented our views on the difference between privacy-preserving data publishing and privacy-preserving data mining, and gave a list of desirable properties of a privacy-preserving data publishing method. We reviewed and compared existing methods in terms of privacy models, anonymization operations, information metrics, and anonymization algorithms. Most of these approaches assumed a single release from a single publisher, and thus only protected the data up to the first release or the first recipient. We also reviewed several works on more challenging publishing scenarios, including multiple release publishing, sequential release publishing, continuous data publishing, and collaborative data publishing. Privacy protection is a complex social issue, which involves policy-making, technology, psychology, and politics. Privacy protection research in computer science can provide only technical solutions to the problem. Successful application of privacy preserving technology will rely on the cooperation of policy makers in governments and decision makers in companies and organizations. Unfortunately, while the deployment of privacy-threatening technology, such as RFID and social networks, grows quickly, the implementation of privacy-preserving technology in real-life applications is very limited. As the gap becomes larger, we foresee that the number of incidents and the scope of privacy breach will increase in the near future. Below, we identify a few potential research directions in privacy preservation, together with some desirable properties that could facilitate the general public, decision makers, and systems engineers to adopt privacy-preserving technology.

Finally, we emphasize that privacy-preserving technology solves only one side of the problem. It is equally important to identify and overcome the non-technical difficulties faced by decision makers when they deploy a privacy-preserving technology. Their typical concerns include the degradation of data/service quality, loss of valuable information, increased costs, and increased complexity. We believe that cross-disciplinary research is the key to remove these obstacles, and urge computer scientists in the privacy protection field to conduct cross-disciplinary research with social scientists in sociology, psychology, and public policy studies. Having a better understanding of the privacy problem from different perspectives can help realize successful applications of privacy-preserving technology.

REFERENCES

[1] R. Agrawal, R. Srikant. "*Privacy-preserving data mining*". In Proceedings of the ACM SIGMOD Conference on Management of Data, pages 439–450, Dallas, TX, May 2000.

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

[3] Charu C. Aggarwal. "*On k-anonymity and the curse of dimensionality*". In Proceedings of the 31st VLDB Conference, pages 901–909, Trondheim, Norway, 2005.

[4] K. Liu. "*Multiplicative Data Perturbation for Privacy Preserving Data Mining*". PhD thesis, University of Maryland, Baltimore County, Baltimore, MD, January 2007.

[5] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. "***l***-diversity: *Privacy beyond k-anonymity*". ACM Transactions on Knowledge Discovery from Data, 1(1), 2006.

[6] K. Muralidhar and R. Sarathy. "*Data shuffling - a new masking approach for numerical data*". Management Science, 52(5):658–670, May 2006.

[7] G.W. Stewart and Ji-Guang Sun. "*Matrix Perturbation Theory*". Academic Press, 1990.

[8] Bayardo, R., Agrawal, R. "*Data privacy through optimal k-anonymization*". In: Proc. of the 21st International Conference on Data Engineering (2005).

[9] Domingo-Ferrer, J., Torra, V.: "*A quantitative comparison of disclosure control methods for microdata*". In: L. Zayatz, P. Doyle, J. Theeuwes, J. Lane (eds.) Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 113–134. North-Holland (2002).

[10] Dwork, C., Nissim, K.: "*Privacy preserving data mining in vertically partitioned database*". In: CRYPTO 2004, vol. 3152, pp. 528–544 (2004).

[11] Evfimievski, A.: "*Randomization in privacy preserving data mining*". SIGKDD Explor. Newsl. 4(2), 43–48 (2002)

[12] Iyengar, V.: Transforming data to satisfy privacy constraints. In: Proc., the Eigth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, pp. 279–288 (2002)

[13] Kifer, D., Gehrke, J.: "*Injecting utility into anonymized datasets*". In: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, pp. 217–228. ACM Press, Chicago, IL, USA (2006)

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

[15] Vaidya, J., Clifton, C.: "*Privacy preserving association rule mining in vertically partitioned data*". In: 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 639–644. ACM Press (2002)

[16] Verykios, V.S., Bertino, E., Nai Fovino, I., Parasiliti, L., Saygin, Y., Theodoridis, Y.: "*State-of-the-art in privacy preserving data mining*". SIGMOD Record 33(1), 50–57 (2004).

[17] Wang, R.Y., Strong, D.M.: "*Beyond accuracy: what data quality means to data consumers*". Journal of Management Information Systems 12(4), 5–34 (1996).

[18] Benjamin C. Fung M. and Ke Wang (2010), "*Privacy-Preserving Data Publishing: A Survey of Recent Developments*", ACM Computing Surveys, Vol. 42, No. 4, pp.322-435.

[19] Bertino E., Nai Fovino and Parasiliti Provenza (2005), "*A Framework for Evaluating Privacy Preserving Data Mining Algorithms*", Journal of Data Mining and Knowledge Discovery, pp. 78-87.

[20] Chris Clifton and Murat Kantarcioglu and Jaideep Vaidya (2002), "*Defining Privacy for Data Mining*", Proceedings of the National Science Foundation Workshop on Next Generation Data Mining, pp.274-281.

[21] Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya (2004), "*Tools for privacy preserving data mining*", Explorations Journal, volume 4, pages 28-34.

[22] Igor Fovino and Marcelo Masera (2008), "*Privacy Preserving Data Mining: A Data Quality Approach*", JRC Scientific and Technical Reports, Vol. 2, pp. 28-37.

[23] N.R. Adam and J.C. Wortmann, "*Security-Control Methods for Statistical Databases: A Comparative Study*" ACM Computing Surveys, Vol. 21, no. 4, 1989, pp. 515 - 556.

[24] Y. Lindell, B. Pinkas, "*Privacy Preserving Data Mining*", Journal of Cryptology, vol. 15, no. 3, 2002, pp. 177-206.