DOI: https://doi.org/10.53555/nncse.v2i1.511 Publication URL: http://nnpub.org/index.php/CSE/article/view/511

REVIEW: PRIVACY PRESERVING AND SENSITIVE DATA HIDING METHODS

Sonu Tomar*[1], Prof. Piyush singh[2]

**[1]PG scholar student RKDF Bhopal ,[2]Assit. Prof. RKDF Bhopal*

***Corresponding Author: -**

Abstract: -

*Privacy preserving data mining deals with hiding an individual’s sensitive identity without sacrificing the usability of data. It has become a very important area of concern but still this branch of research is in its infancy. People today have become well aware of the privacy intrusions of their sensitive data and are very reluctant to share their information. The major area of concern is that non-sensitive data even may deliver sensitive information, including personal information, facts or patterns. Several techniques of privacy preserving data mining have been proposed in literature***.**

Keywords: - *Privacy Preserving, Association Rules, Sensitive Rules, Minimum Support, Minimum confidence*

1. INTRODUCTION

Data Mining [1] refers to extracting or “mining” knowledge from large amounts of data. Data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. By performing data mining, interesting knowledge, regularities, or high- level information can be extracted from database and viewed or browsed from different angles. The discovered knowledge can be applied to decision making, process control, information management, and query processing. Data mining is considered one of the most important frontiers in database systems and one of the most promising interdisciplinary developments in the information industry.

The concept of Privacy-Preserving has recently been proposed in response to the concerns of preserving personal or sensible information derived from data mining algorithms. Successful applications of data mining have been demonstrated in marketing, business, medical analysis, product control, engineering design, bioinformatics and scientific exploration, among others. The current status in data mining research reveals that one of the current technical challenges is the development of techniques that incorporate security and privacy issues. The main reason is that the increasingly popular use of data mining tools has triggered great opportunities in several application areas, which also requires special attention regarding privacy protection. There have been two types of privacy concerning data mining.

The first type of privacy, called output privacy, is that the data is minimally altered so that the mining result will preserve certain privacy. The second type of privacy, input privacy, is that the data is manipulated so that the mining result is not affected or minimally.

For example, through data mining, one is able to infer sensitive information, including personal information, or even patterns from non-sensitive information or unclassified data. As a motivating example of privacy issue in data mining discussed. Consider a supermarket and two breads suppliers A and B. If the transaction database of the supermarket is released, A (or B) can mine the association rules related to his/her breads and apply the rules to the sales promotion and the goods supply. As a result, a supplier is willing to exchange a lower price of goods for the database with the supermarket. From this aspect, it is good for the supermarket to release the database. However, the conclusion can be opposite if a supplier uses the mining methods in a different way. For instance, if A finds the association rules related to B’s breads, saying that most customers who buy cheese also buy B’s breads, he/she can run a coupon that gives a 10 percent discount when buying A’s breads together with cheese. Gradually, the amount of sales on B’s breads is down and B cannot give a low price to the supermarket as before. Finally, A monopolizes the bread market and is unwilling to give a low price to the supermarket as before. From this aspect, releasing the database is bad for the supermarket. Therefore, for the supermarket, an effective way to release the database with sensitive rules hidden is required. This leads to the research of sensitive rule hiding.

II. PRIVACY PRESERVING DATA MINING

Privacy preserving [2] has originated as an important concern with reference to the success of the data mining. Privacy preserving data mining (PPDM) deals with protecting the privacy of individual data or sensitive knowledge without sacrificing the utility of the data. People have become well aware of the privacy intrusions on their personal data and are very reluctant to share their sensitive information. This may lead to the inadvertent results of the data mining. Within the constraints of privacy, several methods have been proposed but still this branch of research is in its infancy. In figure 1, framework for privacy preserving DataMining is shown [2]. Data from different data sources or operational systems are collected and are preprocessed using ETL tools. This transformed and clean data from Level 1 is stored in the data warehouse. Data in data warehouse is used for mining. In level 2, data mining algorithms are used to find patterns and discover knowledge from the historical data.

After mining privacy preservation techniques are used toprotect data from unauthorized access. Sensitive data of an individual can be prevented from being misused.

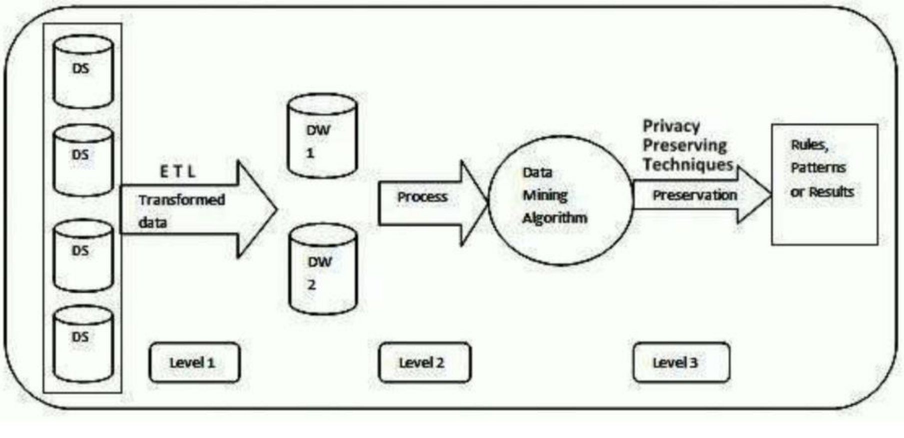


Fig. 1 Framework of privacy preserving data mining

III. RESEARCH CHALLENGES

Now-a-days, Data Mining is used in many applications. There are certain areas where data mining ifused without privacy may cause serious affects. These areas are the main research challenges and are mentioned below.

A. Cyber Terrorism, Insider Threats, and External Attacks

One of the major threats people face today is Cyber Crime [4]. Since most of our information is stored on electronic media and a lot of data is also available on internet or networks. Attacks on such areas might be dangerous and devastating for an individual. For example, consider the Banking system. If hackers attack a bank's information system and empty the accounts, the bank could lose millions of dollars. Therefore, security of information is a critical issue.

There are two types of threats –Outsider or Insider. An attack on Information System from someone outside the organization is called outsider threat, such as hackers, hacking Bank's computer systems and causing havocs. A more critical problem is the insider threat. Insider threat can be due to an intruder present in the organization. Members of an organization have studied their policies and business practices and know every bit of the information so it can affect the organization's information assets.

B. Credit Card Fraud and Identity Theft

Another area which requires attention is detecting frauds and thefts. Frauds may be credit card frauds [4]. These can be detected by identifying purchases made of enormous amounts. A similar and a more serious theft is identity theft. Here one pretends to be an identity of another person by obtaining that person's personal information and carrying out all types of transactions under the other person's name. By the time, the owner finds out it is often far too late-the victims may already have lost millions of dollars due to identity theft

2. Techniques of Privacy Preserving

2.1. Method of anonymization

When releasing micro data for research purposes, one needs to limit disclosure risks to an acceptable level while maximizing data utility. To limit disclosure risk, Samarati et al. [1]; Sweeney [2] introduced the *k*-anonymity privacy requirement, which requires each record in an anonymized table to be indistinguishable with at least *k* other records within the dataset, with respect to a set of quasi-identifier attributes. To achieve the *k*-anonymity requirement, they used both generalization and suppression for data anonymization. Unlike traditional privacy protection techniques such as data swapping and adding noise, information in a *k*anonymous table through generalization and suppression remains truthful. In particular, a table is *k*- anonymous if the Ql values of each tuple are identical, to those of at least *k* other tuples. Table3 shows an example of 2-anonymous generalization for Table. Even with the voter registration list, an adversary can only infer that Ram may be the person involved in the first 2 tuples of Table1, or equivalently, the real disease of Ram is discovered only with probability 50%.

In general, *k* anonymity guarantees that an individual can be associated with his real tuple with a probability at Most1/k

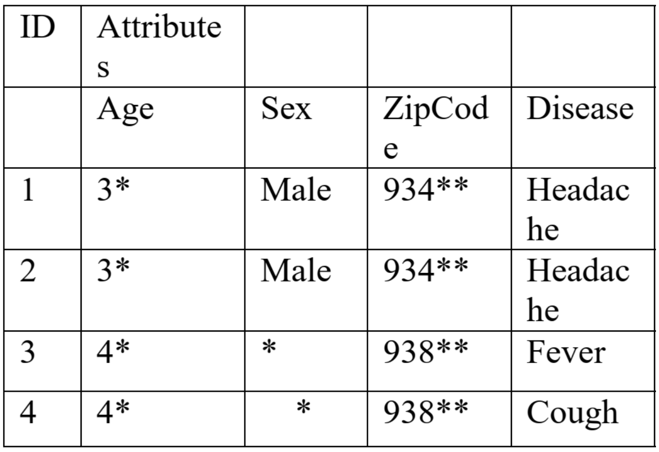
TABLE: -1 MICRODATA

ID	Attributes			
	Age	Sex	ZipCode	Disease
1	36	Male	93461	Headache
2	34	Male	93434	Headache
3	41	Male	93867	fever
4	49	Female	93849	Cough

TABLE: -2VOTER REGISTRATION LIST

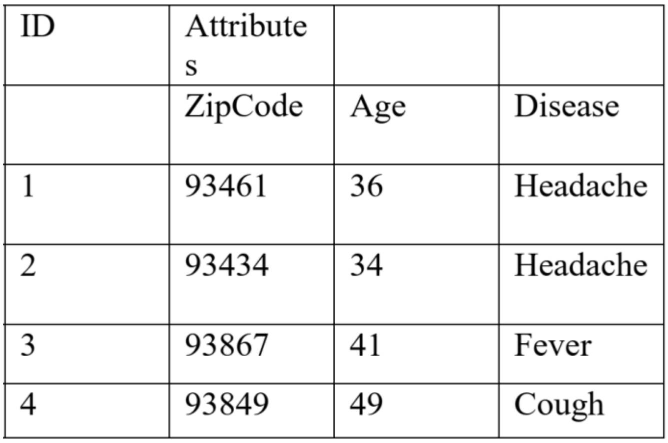
ID	Attributes			
	Name	Age	Sex	ZipCode
1	Ram	36	Male	93461
2	Manu	34	Male	93434
3	Ranu	41	Male	93867
4	Sonu	49	Female	93849

TABLE: -3 A-2 ANONYMOUS TABLE



ID	Attributes			
	Age	Sex	ZipCode	Disease
1	3*	Male	934**	Headache
2	3*	Male	934**	Headache
3	4*	*	938**	Fever
4	4*	*	938**	Cough

TABLE: -4 ORIGINAL PAITENTS TABLE



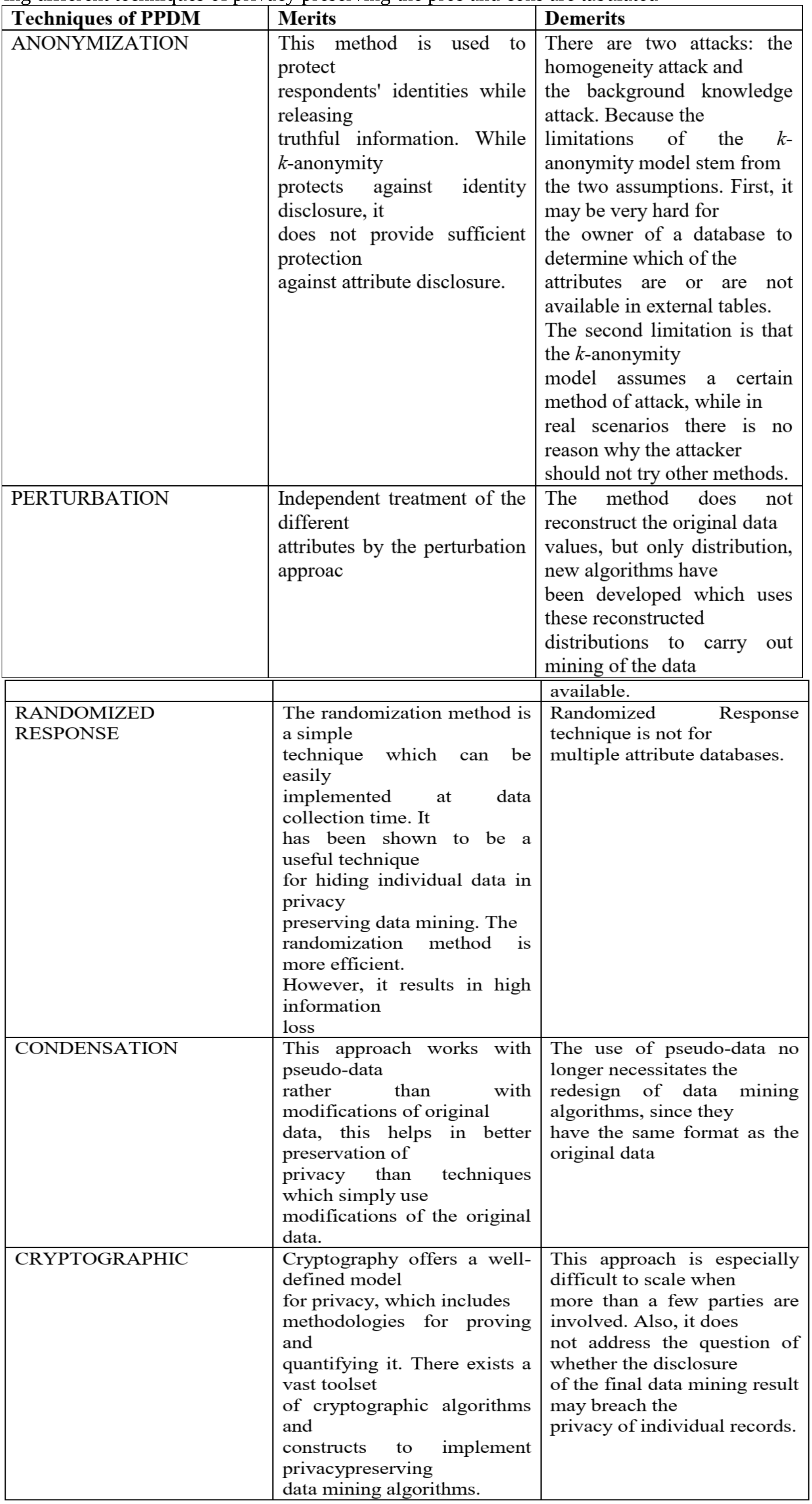
ID	Attributes		
	ZipCode	Age	Disease
1	93461	36	Headache
2	93434	34	Headache
3	93867	41	Fever
4	93849	49	Cough

TABLE: -5 ANONYMOUS VERSIONS OF TABLE1

ID	Attributes		
	Zipcode	Age	Disease
1	934**	3*	Headache
2	934**	3*	Headache
3	938**	4*	Fever
4	938**	4*	Cough

Merits and Demerits of different techniques of PPDM

After reviewing different techniques of privacy preserving the pros and cons are tabulated



Techniques of PPDM	Merits	Demerits
ANONYMIZATION	This method is used to protect respondents' identities while releasing truthful information. While *k*-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure.	There are two attacks: the homogeneity attack and the background knowledge attack. Because the limitations of the *k*-anonymity model stem from the two assumptions. First, it may be very hard for the owner of a database to determine which of the attributes are or are not available in external tables. The second limitation is that the *k*-anonymity model assumes a certain method of attack, while in real scenarios there is no reason why the attacker should not try other methods.
PERTURBATION	Independent treatment of the different attributes by the perturbation approac	The method does not reconstruct the original data values, but only distribution, new algorithms have been developed which uses these reconstructed distributions to carry out mining of the data available.
RANDOMIZED RESPONSE	The randomization method is a simple technique which can be easily implemented at data collection time. It has been shown to be a useful technique for hiding individual data in privacy preserving data mining. The randomization method is more efficient. However, it results in high information loss	Randomized Response technique is not for multiple attribute databases.
CONDENSATION	This approach works with pseudo-data rather than with modifications of original data, this helps in better preservation of privacy than techniques which simply use modifications of the original data.	The use of pseudo-data no longer necessitates the redesign of data mining algorithms, since they have the same format as the original data
CRYPTOGRAPHIC	Cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it. There exists a vast toolset of cryptographic algorithms and constructs to implement privacypreserving data mining algorithms.	This approach is especially difficult to scale when more than a few parties are involved. Also, it does not address the question of whether the disclosure of the final data mining result may breach the privacy of individual records.

Various Techniques Used By Different Authors

Y.Lindell, B.Pinkas et.al. [11] Author Used Cryptographic Technique in 2000 year CryptographicTechnique is A technique through which sensitivedata can be encrypted. There is also aproper toolset for algorithms ofcryptography. And **Result and Accuracyof** This approach is especially difficult toscale when more than a few parties are involved. Also it does not hold good for large databases.

L. Sweeney et. al.[22] Author Used K- Anonymitydechniques in 2002 .A record from a dataset cannot be distinguished from at least k-1 records whose data is also in the dataset .Result and Accuracy of this approach is K- Anonymity Approach is able to preserve privacy.

J. Vaidya and C.Clifton et. al.[20] Authors used Association Rule in 2002 year and association Rule is Distribution of data vertically intosegments.Result and accuracy of this approach is Distribution BasedAssociation Rule DataMining provides privacy.

HillolKar gupta et. al.[7] Author used Data Perturbation in 2003 They tried to preserve data privacy by adding random noise, while making sure that the random noise still preserves the “signal” from the data so that the patterns can still be accurately estimated. And Result and accuracy of this techniques is Randomization-based Techniques are used to generate random matrices.

Charu C. Aggarwal et.al. [12] Authors Used CondensationApproach in 2004 This approach works with pseudo-datarather than with modifications oforiginal data, this helps in betterpreservation of privacy thantechniques which simply usemodifications of the original data. And Result and accuracy of this approach is The use of pseudo-data no longer necessitates theredesign of data miningalgorithms, since theyhave the same format as the original data.

A. Machanavajjhala et. al.[24] Authors used L-DiversityAlgorithm in 2006 in this algorithm If there are ‘l’ ‘well represented’values for sensitive attribute then thatclass is said to have L- Diversity. Result and accuracy is better than KAnonymityin preservingData mining.

Slava Kisilevich et .al. [21] Authors used Anonymization techniques in 2010. Anonymization is a technique forhiding individual’s sensitive data fromowner’s record. Kanonymity is usedfor generalization and suppression fordata hiding. Result and Accuracy Background Knowledge and Homogeneity attacks of K-Anonymity Algorithm do not preserve sensitivity of an individual.

P. Deivanai et. al. [3] used Hybrid Approach in 2011 Hybrid Approach is a combination ofdifferent techniques which combine to give an integrated result. It uses Anonymization and suppression to preserve data.

George Mathew et.al. [25] Authors used Decision Tree in 2011 An approach which is technical,methodological and should givejudgmental knowledge. Result and accuracy of this
Approach is A graph-based frameworkfor preserving patient’ssensitive information.

Anita Parmar[10] used Blocking Based Technique in 2011 Finding sensitive attribute and thenthey replace known sensitive values with unknown values. Finally thesanitized dataset is generated fromwhich sensitive classification rules areno longer mined. Result and Accuracy of this is Unknown Values help inpreserving privacy butreconstruction of originaldata set is quite difficult.

Sara Mumtaz et. al.[16]Distortion Based Perturbation Technique in OLAP Data CubeData perturbation technique which isalso called uniformly adjusted distortion is proposed which initiallydistorts one cell of a cube and thendistortion occurs in whole cube. Result of This distribution of distortion technique notonly preserves, but also provides utmost accuracy with range sum queries and high availability.

Hsiang-Cheh Huang[17]used Histogram Based Reversible Data hiding in 2011 A concept of reversibility which states that an original data can easily be hidden and the hidden data can also be recovered perfectly. Sensitive data is embedded into medical images which is very good technique for hiding secret data. Histogram technique is basically used for X-Ray or CT medical images and it has the potential to be integrated into databases for managing the medical images in the hospital.

Jinfei Liu et. at. [5] used Rating Based Privacy Preservation in 2011 A novel algorithm which overcomes the curse of dimensionality and provides privacy. It is better than K Anonymity and L Diversity.

KhaledAlotaibi et. at. [6] Multi Dimensional Scaling in 2012 A non linear dimensionality reduction technique used to project data on lower dimensional space. The application of nonmetric MDS transformation works efficiently and hence produces better results.

ElaheGhasemi Komishani et. al.[8] used Trajectory data in 2012 Approach for privacy Preservation in trajectory data publishing in which trajectories and sensitive attributes are generalized with respect to different privacy requirements of moving objects. It is able to provide personalized privacy preservation in trajectory data publishing, but also it is resistant to all three identity linkage, attribute linkage, and similarity attacks.

ThanveerJahan et.al.[15] used Data Perturbation Using SSVD in 2012 An analyzing system used to transform original dataset into distorted data set using Sparsified Singular Value Decomposition. Use of Sparsified SVD than SVD is more successful.

D. Karthikeswarant et. al.[19] used Association Rule in 2012 Sanitizes datasets using Sliding Window Algorithm and preserves data. A novel approach that modifies the database to hide sensitive rules.

M. N. Kumbhar et. al. [18] used Association Rule By Horizontal and Vertical Distribution in 2012 Different approaches in the field of Association rule are reviewed. The performance of all models is analyzed in terms of privacy, security and communications.

Savita Lohiya et. al. [9] used Hybrid Approach in 2012 A combination of K- Anonymity and Randomization. It has a better accuracy and original data can b reconstructed.

Martin Beck et. al.[26] used Anonymizing Demonstratorin 2012 Making a demonstrator with user friendly interface and performs Anonymization. Swapping and Recording can be applied to enhance the utility.

VI. CONCLUSION

In today's world, privacy is the major concern to protect the sensitive data. People are very much concerned about their sensitive information which they don't want to share. Our survey in this paper focuses on the existing literature present in the field of Privacy Preserving Data Mining. From our analysis, we have found that there is no single technique that is consistent in all domains. All methods perform in a different way depending on the type of data as well as the type of application or domain. But still from our analysis, we can conclude that Cryptography and Random Data Perturbation methods perform better than the other existing

Methods. Cryptography is best technique for encryption of sensitive data. On the other hand Data Perturbation will help to preserve data and hence sensitivity is maintained. In future, we want to propose a hybrid approach of these techniques.

REFERENCES

[1]. J. Han and M. Kamber , "*Data Mining: Concepts and Techniques*", 2nd ed. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor 2006.

[2]. M. B. Malik, M. A. Ghazi and R. Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", in *proceedings of Third International Conference on Computer and Communication Technology*, IEEE 2012.

[3]. P.Deivanai, J. Jesu Vedha Nayahi and V.Kavitha," A Hybrid Data Anonymization integrated with Suppression for Preserving Privacy in mining multi party data" in *proceedings of International Conference on Recent Trends in Information Technology,* IEEE 2011.

[4]. M. Prakash, G. Singaravel, "A New Model for Privacy Preserving Sensitive Data Mining", in *proceedings of ICCCNT Coimbatore, India,* IEEE 2012.

[5]. J. Liu, J. Luo and J. Z. Huang, "Rating: Privacy Preservation for Multiple Attributes with Different Sensitivity requirements", in *proceedings of 11th IEEE International Conference on Data Mining Workshops*, IEEE 2011.

[6]. K. Alotaibi, V. J. Rayward-Smith, W. Wang and Beatriz de la Iglesia, "Non-linear Dimensionality Reduction for Privacy-Preserving Data Classification" in *proceedings of 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security,Risk and Trust* , IEEE 2012.

[7]. H. Kargupta and S. Datta, Q. Wang and K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques", in *proceedings of the Third IEEE International Conference on Data Mining,* IEEE 2003.

[8]. E. G. Komishani and M. Abadi, "A Generalization-Based Approach for Personalized Privacy Preservation in Trajectory Data Publishing", *in proceedings of 6'th International Symposium on Telecommunications (IST'2012),* IEEE 2012.

[9]. S. Lohiya and L. Ragha, "Privacy Preserving in Data Mining Using Hybrid Approach", in *proceedings of 2012 Fourth International Conference on Computational Intelligence and Communication Networks* , IEEE 2012.

[10]. A. Parmar, U. P. Rao, D. R. Patel, "Blocking based approach for classification Rule hiding to Preserve the Privacy in Database" , in *proceedings of International Symposium on Computer Science and Society,* IEEE 2011.

[11]. Y. Lindell, B.Pinkas, "Privacy preserving data mining", in *proceedings of Journal of Cryptology,* 5(3), 2000.

[12]. C. Aggarwal , P.S. Yu, "A condensation approach to privacy preserving data mining", in *proceedings of International Conference on Extending Database Technology (EDBT),* pp. 183–199, 2004. 746

[13]. R. Agrawal and A. Srikant, " Privacy-preserving data mining", in *proceedings of SIGMOD00,* pp. 439-450.

[14]. Evfimievski, A.Srikant, R.Agrawal, and Gehrke , "Privacy preserving mining of association rules", in proceedings of KDD02, pp. 217-228.

[15]. T. Jahan, G.Narsimha and C.V Guru Rao, “Data Perturbation and Features Selection in Preserving Privacy” in *proceedings of 978-1-4673-1989-8*/12, IEEE 2012.

[16]. S. Mumtaz, A. Rauf and S. Khusro, “*A Distortion Based Technique for Preserving Privacy in OLAP Data Cube*”, in proceedings of 978-1-61284-941-6/11/$26.00, IEEE 2011.

[17]. H.C. Huang, W.C. Fang, “*Integrity Preservation and Privacy Protection for Medical Images with Histogram-Based Reversible Data Hiding*”, in proceedings of 978-14577-0422-2/11/$26.00_c, IEEE 2011.

[18]. M. N. Kumbhar and R. Kharat, “*Privacy Preserving Mining of Association Rules on horizontally and Vertically Partitioned Data: A Review Paper*”, in proceedings of 978-1-4673-5116-4/12/$31.00_c, IEEE 2012.

[19]. D.Karthikeswarant, V.M.Sudha, V.M.Suresh and A.J. Sultan, “A Pattern based framework for privacy preservation through Association rule Mining” in *proceedings of International Conference On Advances In Engineering, Science And Management (ICAESM -2012),* IEEE 2012.

[20]. J. Vaidya and C. Clifton, “Privacy preserving association rule mining in vertically partitioned data”, in *The Eighth ACM SIGKDD International conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, CA, July 2002*, IEEE 2002.

[21]. Slava Kisilevich, Lior Rokach, Yuval Elovici, Bracha Shapira, “Efficient MultiDimensional Suppression for K-Anonymity”, *in proceedings of IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, No. 3. (March 2010), pp. 334-347, IEEE 2010.

[22]. L. Sweeney, "k-Anonymity: A Model for Protecting Privacy”, in *proceedings of Int'l Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 2002.

[23]. The free dictionary.Homepage on Privacy [Online]. Available: http://www.thefreedictionary.com/privacy.

[24]. A. Machanavajjhala, J.Gehrke, D. Kifer and M. Venkitasubramaniam, "I-Diversity: Privacy Beyond k-Anonymity", *Proc. Int'l Con! Data Eng. (ICDE),* p. 24, 2006.

[25]. G. Mathew, Z. Obradovic,” A Privacy-Preserving Framework for Distributed Clinical Decision Support”, *in proceedings of 978-1-61284-852-5/11/$26.00 ©2011 IEEE*.

[26]. Martin Beck and Michael Marh¨ofer,” Privacy-Preserving Data Mining Demonstrator”, in *proceedings of 16th International Conference on Intelligence in Next Generation Networks,* IEEE 2012.