Grenze International Journal of Computer Theory and Engineering , Oct 2017

A Review of Anonymization Approach for Privacy Preservation Data Mining

Ratandeep Kaur[1], Dr. Manisha Sharma[2] and Dr. S.Taruna[3]
[1]Research Scholar, Department of Computer Science, Banasthali Vidyapith, Jaipur(Rajasthan),India
Email: Ratandeep.sgtbimit@gmail.com
[2]Associate Professor, Department of Computer Science, Banasthali Vidyapith, Jaipur(Rajasthan),India
Email: manishasharma8@gmail.com
[3]Associate Professor, Department of Computer Science, JK Lakshmipat University, Jaipur(Rajasthan),India
Email: Staruna@jklu.edu.in

***Abstract*— Social Networking sites have gained enormous popularity over the last few years. Today, internet has become an inevitable part of the lives of more than millions of people. Social networking provides the platform to share the personal information, which has raised the serious concerns related to the privacy and security of the users. A broad area of data mining is focusing on providing the privacy and introduced a field known as privacy preserving data mining(PPDM).This paper(or work) addresses the problem by presenting analysis of anonymization algorithms of privacy preserving data mining (PPDM) such as k-anonymity and l-diversity and t-closeness.**

***Index Terms*— PPDM, K-anonymity, L-diversity, t-closeness, Sensitive Attributes, Quasi-Identifiers, Data anonymization.**

I. INTRODUCTION

Privacy preserving data mining (PPDM) is gaining more popularity as it enables the sharing of data related to private and sensitive information of a user. As the purpose of Data mining is to extract useful information or Knowledge from multiple data sources whereas the main goal of privacy preservation in data mining is to preserve these kinds of data against any disclosure or loss of information. Privacy preserving data mining (PPDM) is a broad area to research in data mining as new challenges are emerging as there is huge rise in the use of social network sites. PPDM is mainly focused on the reducing the privacy risk while modifying the data in such a way that sensitive information can be protected while performing data mining operations[1]. There are various data mining algorithms which deal with the privacy issues and proposed the Privacy Preserving data mining as a two step process. First, protect the sensitive information such as user's password or bank account number from the direct access for mining process. Second, secure the sensitive mining results or outcomes from disclosure or inference that can lead to privacy violation. Privacy of an individual can be at stake when the information is shared with third parties like advertiser, researchers and application developers by the service providers. An individual can have various kinds of privacy breaches over the social network site such as [2]:

A. *Revelation of Person's Identity* – In this the user can be identified over a network and all the information related to him/her and relationship with other users can be revealed.

Grenze ID: 01.GIJCTE.3.3.503

B. *Disclosure of Link or Association* – It is the information related to association between two individual users which can be accessed by using social activities or services by the user.
C. *Access to Sensitive Attributes* – This type of disclosure can occur when someone can access user's account by link relationship and get the confidential data that can harm the user.

All the above mentioned privacy breaches become threats for users like stalking, blackmailing, financial loss and tarnishing the public image as user expect privacy and security from social network service providers. Privacy of data must be ensured before sharing the data by the service provider.

The rest of the paper is organized as follows: Section I gives the introduction of PPDM. Section II describes the anonymization technique of PPDM and their classifications. Section III describes the related work that has been done in the area of anonymization.

II. PPDM TECHNIQUE-ANONYMIZATION

PPDM proposed anonymizing of data as a technique to remove the identifying information from the original data of the user to protect the sensitive information while it is shared to others. Anonymization is classified into two methods, generalization and suppression [3]. In generalization based method , original values have been replaced by more general values to form the subsets of original records and in suppression based method, certain values of attributes are replaced by special values like an asterisk'*' in order to form the original data records.

While anonymizing the data there are three types of attributes are used [4] as shown in Fig.1:

A. *Explicit Identifier/Key Attribute*: It is the information or attribute that can directly identify the individual for e.g. Name, Voter ID.
B. *Quasi –Identifier/Pseudo-Identifier*: It is the attribute that combine with other attributes to uniquely identify the individual e.g. Date of Birth, zip, age.
C. *Sensitive Attribute*: It is the attribute which hold the sensitive or personal information about the individual e.g. Salary, bank balance.

	Quasi-identifier			*Key attribute*	*Sensitive Data*
#	**Zip**	**Age**	**Nationality**	**Name**	**Condition**
1	13053	28	Indian	Kumar	Heart Disease
2	13067	29	American	Bob	Heart Disease
3	13053	35	Canadian	Ivan	Viral Infection
4	13067	36	Japanese	Umeko	Cancer

Figure1. Classification of attributes [5]

Anonymization is a technique to sanitize the information by encrypting or removing that data from data set which can uniquely identify the individual.

A. K- anonymity

According to Latanya Sweeney (2002) , k- anonymity is proposed as a model for privacy preservation to provide protection against attribute linkage [6].It states that- There should be at least k tuples having the same quasi-identifier values to guarantee an individual's privacy. Every tuple in a table should be similar to at least (k-1) tuples then only the table will achieve k-anonymity. K- anonymity use generalization and suppression technique to create a data set T satisfies the anonymity of data [7] .Generalization is performed at the column or attribute level(AG) and at cell level(CG). Suppression can be performed at tuple or row(TS),Attribute(AS) or cell(CS) level. By different combination of generalization and suppression, several models of k-anon has been proposed such as[8],shown in Fig.2:

i) AG_TS: Generalization is applied at the level of attribute (column) and suppression at the level of tuple (row).
ii) AG_AS: Both generalization and suppression are applied at the level of column.
iii) AG_CS: Generalization is applied at the level of column, while suppression at the level of cell.
iv) AG: Generalization is applied at the level of column, suppression is not considered.

v) CG_CS: Both generalization and suppression are applied at the cell level. Then, for a given attribute we can have values at different levels of generalization.
vi) CG: Generalization is applied at the level of cell, suppression is not considered.
vii) TS: Suppression is applied at the tuple level, generalization is not allowed.
viii) AS: Suppression is applied at the attribute level, generalization is not allowed.
ix) CS: Suppression is applied at the cell level, generalization is not allowed.

Generalization	Suppression			
	Tuple	*Attribute*	*Cell*	*None*
Attribute	**AG_TS**	**AG_AS** ≡ AG_	**AG_CS**	**AG_** ≡ AG_AS
Cell	**CG_TS** not applicable	**CG_AS** not applicable	**CG_CS** ≡ CG_	**CG_** ≡ CG_CS
None	**_TS**	**_AS**	**_CS**	- not interesting

Figure 2. Classification of K-anonymity models[9]

Attack on K-anonymity

a. *Background Knowledge Attack*: This attack is based on the association between one or more quasi-identifier attributes with the sensitive attributes that result in the reduce set of possible values for the sensitive attributes. For example, For example, Machanavajjhala et al. (2007) showed that knowing that heart attacks occur at a reduced rate in Japanese patients could be used to narrow the range of values for a sensitive attribute of a patient's disease [10].
b. *Homogeneity Attack*: This attack is responsible where all the values of a sensitive attribute within a set of k records are identical and can be easily predicted.

K-anonymity has the advantage of preventing the linkage of records by generating large equivalence class but if in that class most of the records have similar values of sensitive attribute then the attacker can relate to those values without identifying the record of the individual.

B. L-diversity

L-diversity was proposed by Machanavajjhala et al.(2006) to preserve the privacy related to user's relationship privacy[10]. "A data set is said to satisfy l-diversity if, for each group of records sharing a combination of key attributes, there are at least l "well represented" values for each confidential attribute"[11]. Machanavajjhala et al. (2006) defines "well-represented" in three possible ways [10]:

a. *Distinct l-diversity* – The simplest definition ensures that at least *l* distinct values for the sensitive field in each equivalence class.
b. *Entropy l-diversity* – The most complex definition defines *Entropy* of an equivalent class *E* to be the negation of summation of *s* across the domain of the sensitive attribute of $p(E,s)\log(p(E,s))$ where $p(E,s)$ is the fraction of records in *E* that have the sensitive value *s*. A table has entropy *l*-diversity when for every equivalent class *E*, $Entropy(E) \geq \log(l)$.
c. *Recursive (c-l)-diversity* – A compromise definition that ensures the most common value does not appear too often while less common values are ensured to not appear too infrequently.

Attacks on L-diversity [10]:

a. *Skewness Attack*: When the overall dispersion is skewed, satisfying l-diversity does not prevent characteristic disclosure.
b. *Similarity Attack*: When the sensitive attribute values in an equivalence class are distinct but semantically similar, an attacker can learn essential information.

L-diversity prevents from homogeneity and background knowledge attack but it is insufficient to prevent the attribute disclosure.

C. t- closeness

According to Ninghui et.al.(2007) , an equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t[12]. A table is said to have t-closeness if all equivalence classes have t-closeness [7]. t-closeness is a further refinement of l-diversity group based anonymization that is used to preserve privacy

in data sets by reducing the granularity of a data representation. This reduction is a trade off that result in some loss of effectiveness of data management or mining algorithms in order to gain some privacy. The *t*-closeness model extends the *l*-diversity model by treating the values of an attribute distinctly by taking into account the distribution of data values for that attribute [7].t- closeness provide protection against the attribute disclosure but not against the identity disclosure[13].Table 1 briefs the models along with their benefits and their disadvantages.

TABLE I. ANONYMIZATION MODELS FOR PRIVACY PRESERVATION

Anonymizing techniques	Approach	Benefits	Drawbacks
k-anonymity	Generalization Suppression	Prevents identity Disclosure	Background Knowledge attack Homogeneity attack
l-Diversity	Diversification of sensitive attributes	Prevent background Knowledge attack Prevent Homogeneity attack	Skewness Attack Similarity Attack
t-closeness	Extension of l-diversity	Prevent attribute disclosure	Identity disclosure

III. RELATED WORK

In 2000, Lindell et al. proposed cryptographic technique to develop an encryption algorithm to encrypt sensitive data but this approach is not efficient for large databases and less scalable[14]. Samarati(2001) proposed that suppression can be used to achieve k-anonymity with minimal generalization and get the optimal solution[9]. In 2005 an Incognito algorithm is designed which produces all the possible k-anonymous full-domain generalizations of a relation(say T), with an optional tuple suppression threshold to locate the optimal solution but it uses the breadth first search method which takes lot of time to traverse the solution space [15]. Erkin et al.(2007) proposed an approach based upon k-means clustering approach to identify multi-party relations but it is infeasible for situations where the amount of data is large and complex[16]. A research has been conducted by Lijie et al. (2009) to study the link identification disclosure in which the more hazard attacks using link probability, *t*-confidence has been proposed [17].Further, another approach proposed by Tang et al.(2010) use edge based method to generalize social network that result in lower error rate in closeness identification[18]. Kavianpour et al. (2011) designed an integrated algorithm by combining the advantages of k-anonymity and l-diversity algorithm then evaluated the effectiveness of the combined strengths. This algorithm has been able to increase the level of privacy for social network users by anonymizing and diversifying disclosed information [19].Sowmiyaa et al.(2015) proposed a heuristic generating algorithm for privacy preservation of micro data , result in the reduction of possibilities of similarity attack and result in less distortion ratio[20]. In Table II, the various techniques of Privacy Preserving are summarized.

TABLE II. VARIOUS APPROACHES FOR PRIVACY PRESERVATION

Authors	Reference & Year	Technique	Approach	Result
Y. Lindell, B.Pinkas	[14] 2000	Cryptographic Technique	An encryption algorithm for encrypt sensitive data.	This approach is not efficient for large databases and less scalable.
L. Sweeney	[6] 2002	k- anonymity	A technique to distinguish from k-1 records from same dataset.	These approaches efficiently handle the privacy preservation.
Rizvi S., Haritsa J	[21] 2002	Probabilistic Distortion	A method based on association rule mining	This approach provides accuracy in data mining.
J.Gehrke, A. Machanavajjhala, D. Kifer and M.Venkitasubramania m	[10] 2006	l-diversity Algorithm	A technique to identify class values for sensitive attribute.	This approach prevent the limitations of k-anonymity in preserving Data mining.
Erkin et al.	[16] 2007	k-means approach	A technique based upon k-means clustering approach to identify multi-party relations.	This approach is infeasible for situations where the parties having large amount of data or there are complex functions to be evaluated.

B. Zhou, J. Pei, and W. Luk	[22] 2008	Anonymization Technique	An approach to identify the information loss in anonymizing social network data	This approach is not effective in context of social networking as compared to relational data.
Roy Ford, Traian Marius Truta, and Alina Campan	[23] 2009	p-sensitive k-anonymity	A greedy clustering algorithm to analyze social networks	This approach is efficiently providing identity protection in social networks and also secures them from the disclosure of sensitive information.
X. Tang and C.C. Yang	[18] 2010	KNN and EBB algorithm	An algorithm for Identifying closeness centrality measures in social network	This approach use edge based method to generalize social network result in lower error rate in closeness computation than using k-nearest neighbor method
Aaron Beach, Mike Gartrell, Richard Han	[24] 2010	q-Anon Technique	A technique to identify and access the unknown information from a social network	This technique does not provide the efficient amount of anonymous data when there is large dataset
Sanaz Kavianpour, Zuraini Ismail, and Amirhossein Mohtaseb	[25] 2011	k-anon & l-diversity algorithm	An integrated algorithm to measure the effectiveness of both the approaches i.e. k-anon & l-diversity	This approach results in the increasing level of privacy for social networking sites
Raymond Heatherly, Murat Kantarcioglu, and Bhavani Thuraisingham	[26] 2013	Na¨ıveBayes classification algorithm	An algorithm to predict the private information from user's profile	This approach provide better predictability but result in deletion of some information from user's profile
Sowmiyaa P, Tamilarasu P, Kavitha S, Rekha A, Gayathri R Krishna	[20] 2015	k-anonymity algorithm	A heuristic generating algorithm for privacy preservation of microdata	This approach leads to the reduction of possibilities of similarity attack and result in less distortion ratio

IV. CONCLUSION

Privacy Preserving Data Mining is a broad area of research and has various classifications. Data anonymization modifies the dataset to prevent the loss of sensitive information. This paper provides the brief analysis of anonymization techniques such as k-anonymity, l-diversity and t-closeness along with their benefits and drawbacks. Further various models of these classifications have also been discussed. Also, some works related to the Data anonymization and Privacy Preserving techniques have been shown. Existing approaches provides the solution for preserving the privacy by modify the original data but these techniques results in substantial information loss. In future, there is a scope of efficient techniques that include anonymizing multiple sensitive attributes, evaluation of large datasets and non homogeneous data anonymization for gaining minimum information loss and accuracy of released data.

REFERENCES

[1] Jian Wang, Yongcheng Luo, Yan Zhao and Jiajin Le, "A Survey on Privacy Preserving Data Mining" , in IEEE,2009 First International Workshop on Database Technology and Applications.

[2] Kun Liu, Kamalika Das, Tyrone Grandison, Hillol Kargupta, "Privacy-preserving data analysis on graphs and social networks," In: Next Generation of Data Mining, pp. 419-437, 2008.

[3] Mahesh Dhande, N.A.Nemade and Yogesh Kolhe, "Privacy Preserving in K- Anonymization Databases Using AES Technique",2013.

[4] Li N., Li T., Venkatasubramanian," t-Closeness: Privacy beyond k-anonymity and l-diversity", ICDE Conference, 2007.

[5] Terrovitis, Manolis, Nikos Mamoulis, and Panos Kalnis. "Privacy-preserving anonymization of set-valued data." *Proceedings of the VLDB Endowment* 1.1 (2008): 115-125.

[6] L. Sweeney," k-anonymity: a model for protecting privacy," International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

[7] Snehal M. Nargundi, Rashmi Phalnikar, "k-Anonymization using Multidimensional Suppression for Data De-identification", International Journal of Computer Applications (0975 – 8887) Volume 60– No.11, December 2012 .

[8] Kinjal Parmar,Vinita Shah(2015),"A Review on Data Anonymization in Privacy Preserving Data Mining", , In Proc. of IJARCCE,Vol. 5, Issue 2, February 2016.

[9] Samarati P (2001). "Protecting respondents' identities in micro-data release", IEEE Transactions on Knowledge and Data Engineering, 13(6):1010–1027.

[10] 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.
[11] M.E. Nergiz, C. Clifton, and A.E. Nergiz , “Multi-relational k-anonymity. Knowledge and Data Engineering”, IEEE Transactions on, 21(8):1104 –1117, aug. 2009.
[12] Ninghui Li Tiancheng Li,Suresh Venkatasubramanian,” t-Closeness: Privacy Beyond k-Anonymity and l-diversity”, ICDE 2007, pp. 106–115 .
[13] Presswala, Freny, Amit Thakkar, and Nirav Bhatt. "Survey on Anonymization in Privacy Preserving Data Mining.", In Proc. of IJIERE,Vol. 2, Issue 2, 2015.
[14] Y. Lindell, B.Pinkas,” Privacy preserving data mining,”,in proceedings of Journal of Cryptology, 5(3), 2000.
[15] Kristen LeFevre David J. DeWitt Raghu Ramakrishnan “Incognito: Efficient Full Domain KAnonymity”. SIGMOD '05 Proceedings of the 2005 ACM SIGMOD international conference on Management of data ,Pages 49-60 .
[16] Erkin Z., Piva A., Katzenbeisser S., Lagendijk R., Shokrollahi J., Neven G., and Barni M., “Protection and Retrieval of Encrypted Multimedia Content: When Cryptography meets Signal Processing”, EURASIP Journal of Information Security, vol. 7, no. 17, pp. 1 - 20, 2007.
[17] Z. Lijie and Z. Weining ,”Edge Anonymity in Social Network Graphs”, in Proc. of International Conference on Computational Science and Engineering CSE ,2009, pp 1-8, 2009.
[18] X.Tang and C.C. Yang , “Generalizing Terrorist Social Networks with K-Nearest Neighbour and Edge Betweenness for Social Network Integration and Privacy Preservation”, In Proc. of IEEE International Conference on Intelligence and Security Informatics, 2010.
[19] Sanaz Kavianpour, Zuraini Ismail, and Amirhossein Mohtaseb, “Preserving Identity Of Users In Social Network Sites By Integrating Anonymization And Diversification Algorithms”, In: International Journal of Digital Information and Wireless Communications (IJDIWC), Hongkong, Vol. 1, Issue 1, pp 32-40, 2011.
[20] Sowmiyaa P , Tamilarasu P2 , Kavitha S 3 , Rekha A 4 , Gayathri R Krishna5,” Privacy Preservation for Micro-data by using KAnonymity Algorthim”, In Proc. of IJARCCE,Vol. 4, Issue 4, April 2015.
[21] Rizvi S., Haritsa J.: “Maintaining Data Privacy in Association Rule Mining”, VLDB Conference, 2002.
[22] B. Zhou, J. Pei, andW. Luk, “A brief survey on anonymization techniques for privacy preserving publishing of social network data”, ACM SIGKDD Explorations Newslett., vol. 10, no. 2, pp. 12_22, 2008.
[23] Roy Ford, Traian Marius Truta, and Alina Campan, “P-Sensitive K-Anonymity for Social Networks”, In Proc. of International Conference on Data Mining,USA,2009.
[24] Aaron Beach, Mike Gartrell, Richard Han, “q-Anon: Rethinking Anonymity for Social Networks”, In Proc. of IEEE Second International Conference on Social Computing (SocialCom), Minneapolis, MN, pp 185 – 192, 2010.
[25] Sanaz Kavianpour, Zuraini Ismail, and Amirhossein Mohtaseb, “Preserving Identity Of Users In Social Network Sites By Integrating Anonymization And Diversification Algorithms”, In: International Journal of Digital Information and Wireless Communications (IJDIWC), Hongkong, Vol. 1, Issue 1, pp 32-40, 2011.
[26] Raymond Heatherly, Murat Kantarcioglu, and Bhavani Thuraisingham, “Preventing Private Information Inference Attacks on Social Networks", In: IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 8, pp 1849-1862, 2013.
[27] Charu C. Aggarwal, Philip S. Yu, “Privacy-Preserving Data Mining: Models And Algorithms”, Kluwer Academic Publishers, Boston/Dordrecht/London.