International Journal of Innovative Research in Engineering Science and Technology

APRIL 2018 ISSN 2320 –981X

A SURVEY ON ANONYMIZING DATA IN GRAPH BASED MULTIFOLD MODEL

[1]Ms.G.Bhavani, [2]Dr.S.Sivakumari

1. PhD Scholar,
   Department of Computer Science and Engineering,
   Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore-641108.
   bhavanigb16@gmail.com.

2. Professor and Head,
   Department of Computer Science and Engineering,
   Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore-641108.
   prof.sivakumari@gmail.com

Abstract-Privacy preservation is an essential aspect in data mining as privacy of sensitive information should be preserved while allocating the data between different untrusted parties. Privacy preserving data mining (PPDM) protects the secrecy of sensitive data without losing the usability of the data. Transactional data with attributes of several types may be enormously suitable to secondary analysis. However, anonymization of such data is challenging because it comprises multiple types of attributes. Prevailing privacy-preserving techniques are not suitable to address this problem. A novel graph-based multi-fold model has been used to anonymize data with attributes of various types. Data are represented using graph where privacy is protected through clustering and fuzzing among sensitive attributes and renovating associations among items into an indeterminate form. Also k-anonymity and other anonymization techniques are analysed for preserving privacy.

1. INTRODUCTION

a) Anonymization: Set-valued data are transactional data which consists of sets of items that has to be preserved through anonymization. Transactional data involves multiple types of related objects. These types of datasets are called transactional datasets with multiple types attributes called AMT datasets. It contains numerous variety of sensitive information like private information of a person which should be preserved. The anonymizing approaches aims to prevent sensitive attributes leakage and also the identity and association of a person should be disclosed. This can be achieved by using association of single (relational attributes) and set (transaction attributes) values and then their relationships are analysed.

b) K-anonymity: Anonymization can be performed through k-anonymity which is used to make the data protected from others. The data released after k-anonymization should contain information for each person that cannot be distinguished from at least k-1 individuals. For some value of k the anonymization can be performed through *suppression* where certain values of the attributes are replaced by an asterisk '*' and *generalization* where individual values of attributes are replaced by with a broader category.

c) l-diversity : A group based anonymization for preserving data privacy by granularity reduction is l-diversity. This model is an extension of the k-anonymity which overcomes the drawbacks like homogeneity and knowledge discovery attacks by adding intra-group diversity for sensitive values.

*i) Homogeneity Attack*arises when all the values for a sensitive value within a set of k records are identical.

*ii) Background Knowledge Attack*arises when reduced set of possible values are obtained in association of quasi identifiers for the sensitive attribute.

*d) t-closeness*is also a group based anonymization approach which overcomes the drawbacks of l-diversity by treating the values of an attribute distinctly and by taking into account the distribution of data values for a specific attribute.

Attribute disclosure should be reduced to prevent sensitive information leaks and also l-diversity could not identify semantically close values.

2. LITERATURE SURVEY

The analyses of the existing work of anonymization are as follows:

Ercan et.al.,[1] proposed a new clustering algorithm to achieve multi relational anonymity by confirming that data cannot be related to a single individual. This paper provides a single table for multi relational anonymization.

Benjamin et.al.,[2] proposed the analysis to review and estimate diverse methodologies to perform Privacy Preserving in Data Publishing and study the provocations in real-world data publishing, elucidate the modifications and necessities that distinguish PPDP from other related problems.

Liuet.al.,[3] proposed novel techniques for addressing the effectiveness and scalability tasks. Wide trials on real-world databases show that this approach overtakes the state-of-the-art methods that embrace global generalization, local generalization, and total suppression which can be analysed through benchmarked data mining tools.

Loukides et.al.,[4] proposed the process of assessing the balance between disclosure risk and data utility obtained by standard algorithms using the R-U confidentiality map that permits the construction of high-quality anonymization results.

Comas et.al.,[5] proposed a construction based on bucketization and computational procedure to attain t-closeness and l-differential privacy in data publishing with the k-anonymity set of models on transaction data.

Poulis et.al.,[6] proposed two frameworks to offer privacy, with bounded information loss in one attribute type and minimal information loss in the other. Three cluster merging algorithms have been developed to preserve data utility and attribute disclosure.

Ghinita et.al.,[7] proposed a novel anonymization method for data correlation using real-life datasets and enables the formation of anonymized groups with low information loss through which attribute disclosure and data utility can be achieved.

Terrovitis et.al.,[8] proposed an anonymization model which depend on generalization. The developed algorithm finds the optimal solution for realistic problems with two greedy heuristics of low computational cost with near optimal solutions.

Machanavajjhala et.al.,[9] discussed two simple attacks on k-anonymity: the sensitive attributes leakage and background knowledge attack. To overcome the attacks ℓ-diversity has been proposed for applied and resourceful implementation.

Ninghui Li et.al.,[10] analysed the k-anonymity and l-diversity which does not prevent attribute disclosure and then proposed t-closeness in which distribution of sensitive attributes should be close to the distribution of overall distribution.

Cormodeet.al,[11] proposed(k,l)-groupings of anonymization, for bipartite graph data that preserve the basic graph construction flawlessly. This grouping assure resistance to various attacks and reveal that (k,l)-groupings provide solid balance among privacy and utility.

Chang et.al,[12] proposed a bipartite graph method for detecting various attacks that can re-identify users and determine their item ratings. For dealing with these attacks, formal privacy definitions are defined for recommendation data and then construct an efficient and more predictive anonymization algorithm.

Li et.al,[13] proposed an approach to attain privacy of sensitive associations between entities and retain the largest amount of non sensitive associations to provide better data utility. The accuracy of answering aggregate queries have been measured which provides solid balance among privacy and utility.

Xu et.al,[14]discussed the problem of publishing transaction data and concluded with two implications: transaction data are excellent candidates for data mining research and use of transaction data would raise serious concerns over individual privacy.

Xiao et.al,[15] proposed a novel technique called anatomy for publishing sensitive data which releases all the quasi-identifier and sensitive values in two separate tables. Linear-time algorithms for computing anatomized tables have been developed to follow the l-diversity privacy requirement which reduce the error of rebuilding the micro data.

Xue et.al.,[16] proposed a novel alternative method that places data in a generalized form through which generalized bitmaps are employed and recasts data values in a non reciprocal manner. This scheme assures popular privacy and resists attacks and gain a clear utility advantage over the previous state of the art.

Boldi et.al.,[17] proposed a new anonymization approach which is based on injecting uncertainty in social graphs and publishing the resulting uncertain graphs. Fine-grained perturbation has been developed that adds or removes edges partially to achieve obfuscation and utility.

Cao et.al.,[18] proposed ρ -uncertainty that inherently safeguards against sensitive associations without constraining the nature of an adversary's knowledge and without falsifying data. The information loss occurred has been overcome by trivial solution to suppress all sensitive items and non-trivially by a combining generalization and suppression.

Samarati et.al.,[19] addressed the problem of releasing micro data while safeguarding the anonymization through k-anonymity. K-anonymity can be employed without compromising the information integrity using generalization and suppression techniques.

Sweeney [20] proposed a formal protection model named *k*-anonymity and a set of accompanying policies for deployment. Re-identification attacks have been examined that can be realized on releases that adhere to *k*-anonymity unless accompanying policies are respected. The *k*-anonymity protection model forms the basis on which the real-world systems like Data fly, m-Argus and *k*-Similar provide guarantees of privacy protection.

Divanis et.al.,[21] proposed a novel clustering-based framework to anonymizing transaction data, which provides the basis for designing algorithms that better preserve data utility. This approach out performs the current state-of-the-art algorithms with greater utility and efficiency.

Nguyen et.al.,[22]analyzed the drawbacks in a recent uncertainty-based anonymization scheme and proposed Maximum Variance approach thatprovides better tradeoff between privacy and utility.

He et.al.,[23] proposed a top-down, partition-based approach for anonymizing set-valued data that scales linearly with the input size and scores well on an information-loss data quality metric. This technique can be applied to anonymize the infamous AOL query logs.

Terrovitis et.al.,[24] proposed the problem of protecting privacy in the publication of set-valued data. The data are classified into potential quasi-identifiers and potential sensitive data based on the knowledge adversary and efficient algorithms are applied to transform the database. An Apriori algorithm has been developed to find the optimal and generalized solution. The applications of techniques partition the database and perform anonymization to achieve reduced memory consumption.

3. CONCLUSION

Thus the privacy preserving data mining concepts through anonymization have been discussed. The different types of anonymization namely k-anonymity, l-diversity and t-closeness were deliberated with the various existing works in this domain.

REFERENCE

[1] Nergiz ME, Clifton C, Nergiz AE. Multirelational k-anonymity. IEEE Trans Knowl Data Eng 2009;21(8):1104–17.

[2] Fung BCM, Wang K, Chen R, Yu PS. Privacy-preserving data publishing: a survey of recent developments. ACM ComputSurv 2010;42.

[3] Liu J,Wang K. Anonymizing transaction data by integrating suppression and generalization. In: PAKDD 2010. Hyderabad, India: 2010. p. 171–80.

[4] Loukides G, Gkoulalas-Divanis A, Shao J. Assessing disclosure risk and data utility trade-off in transaction data anonymization. Int J Softw Inform 2012;6(3):399–417.

[5] Soria-Comas J, Domingo-Ferrert J. Differential privacy via t-closeness in data publishing privacy. In: 2013 eleventh annual international conference on security and trust (PST). 2013. p. 27–35.

[6] Poulis G, Loukides G, Gkoulalas-Divanis A, Skiadopoulos S. Anonymizing data with relational and transaction attributes.In: Machine learning and knowledge discovery in databases.Springer; 2013. p. 353–69.

[7] Ghinita G, Tao Y, Kalnis P. On the anonymization of sparse high dimensional data. In: ICDE. 2008. p. 715–24.

[8] Terrovitis M, Mamoulis N, Kalnis P. Privacy-preserving anonymization of set-valued data. In: Proceedings of VLDB. 2008. p. 115–25.

[9] Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. l-diversity: privacy beyond k-anonymity. ACMTransKnowlDiscov Data 2007;1(1):1–52.

[10] Li N, Li T, Venkatasubramanian S. T-closeness: privacy beyond *k*-anonymity and *l*-diversity. In: IEEE 23rd international conference on data engineering, 2007. ICDE 2007.IEEE; 2007.

[11] Cormode G, Srivastava D, Yu T, Zhang Q. Anonymizing bipartite graph data using safe groupings. VLDB J 2010; 19:115–39.

[12] Chang CC, Thompson B,Wang H, Yao D. Towards publishing recommendation data with predictive anonymization. In: 5th ACM symposium on information, computer and communications security. 2010. p. 24–35.

[13]Wang L-E, Li X. Personalized privacy protection for transactionaldata. In: Advanced data mining and applications. Springer International Publishing; 2014b. p. 253–66.

[14] XuY,Wang K, Fu AWC, Yu PS. Anonymizing transaction databases for publication. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, Las Vegas, Nevada, USA, August 24–27, 2008. p. 767–75.

[15] Xiao X, Tao Y. Anatomy: simple and effective privacy preservation. In: Proceedings of VLDB. 2006. p. 139–50.

[16] Xue M, Karras P, Raissi C, Vaidya J, Tan KL. Anonymizing set valued data by nonreciprocal recoding. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. 2012. p. 1050–8.

[17] Boldi P, Bonchi F, Gionis A, Tassa T. Injecting uncertainty in graphs for identity obfuscation. Proceedings VLDB Endowment 2012; 5(11):1376–87.

[18] Cao J, Karras P, Raïssi C, Tan K-L. ρ-uncertainty: inference-proof transaction anonymization. Proceedings VLDB Endowment 2010;3(1–2):1033–44.

[19] Samarati P. Protecting respondents' identities in microdata release. IEEE Trans Knowl Data Eng 2001;13(6):1010–27.

[20] Sweeney L. k -Anonymity: a model for protecting privacy. IJUFKBS 2002;10(5):557–70.

[21] Gkoulalas-Divanis A, Loukides G. Utility-guided clustering-based transaction data anonymization. Trans Data Priv 2012;5(1):223–51.

[22] Nguyen HH, Imine A, Rusinowitch M. A maximum variance approach for graph anonymization. In: Foundations and practice of security. Springer; 2014. p. 49–64.

[23] He Y, Naughton J. Anonymization of set-valued data via topdown, local generalization. Proceedings VLDB Endowment 2009;2:934–46.

[24] Terrovitis M, Mamoulis N, Kalnis P. Local and global recording methods for anonymizing set-valued data. VLDB J 2011;20:83– 106.