A DISTINCTIVE PRIVACY BASED ALGORITHM FOR CROWDSOURCING PLATFORM

Sheeja. S [1], Gokilavani. M[2]

ABSTRACT

Crowdsourcing is modern businesses term where an individual, an institution or non-profit organization proposes to an individual or a group of individuals to accomplish a particular task. Crowdsourcing is platforms where employers are allowed interact with the employees and get their work done. Crowdsourcing environment may contain sensitive attributes, which may lead to privacy leakage and the outsiders could link SA's with other public databases to reveal individual confidential information. In recent years many techniques like randomization, generalization, k-anonymity, l-diversity have been proposed by many researchers in order to safeguard the privacy of sensitive data. In this paper we introduce an enhanced approach for crowdsourcing, which helps to overcome privacy violations to the maximum extent.

1. INTRODUCTION

Crowdsourcing[1] database is place where the employer acquire his work done from a group of persons based on his requirement. The employer and the employee accomplish their requirements from crowdsourcing platform. The employers are given the rights to upload their job details to the crowdsourcing platform and provide the work when and where needed.

Crowdsourcing platforms, such as Amazon AMT1 and Crowdflower2, use the new LaaS (Labor as a Service) [1] model. Operators process each request from the workers and publish the records in crowdsourcing platform for further processing. Many users register their curriculum vitae (CV) and many companies submit their job positions in the Websites. Hence these database consist the educational and experience summary of individuals and also personal details salary etc.,.These personal information cannot be shared with public which are considered to be sensitive information, whereas these information has to be shared with the employers. Employers require these information's to select the suitable candidate for the particular task. The primary work of operators is to gather the answers related to the queries that were previously published in the website. For example human resource agencies receive many applications from both the users and companies. Employees submit their resumes and companies submit their job positions. In modern years, it has become essential to share private data. In general data from various databases is collected, gathered and organized in central place for many reasons. Crowdsourcing platform is a place where lots of public interact. Public might share their sensitive data like their salary etc. Hence it is essential to hide the sensitive data. The attributes are classified into three types which are key attributes, quasi identifiers and sensitive attributes.

[1]Associate Professor, Department of Computer Applications, Karpagam Academy of Higher Education, Coimbatore, Tamilnadu – India

[2]Research Scholar, Karpagam University, Coimbatore, Tamilnadu - India

- Key attributes are the fields which uniquely identify the records such as names, identification number etc., which has to be always removed before publishing.
- Quasi-identifiers are the fields which do not uniquely identify the records but are well associated with an entity. For example a combination of date of birth, gender and company name can be termed as quasi identifiers which can be used identify the records. So it is very important to partially hide these identifiers before publishing it to public.
- Sensitive attributes have to be really preserved from the attackers. Example salary details of an employer should be preserved privately. Sensitive attributes may provide more knowledge to the invaders and may result in leakage of information

III. Related Work

Many techniques have been designed to preserve privacy to data and allow at the same time provide provision to analyze the data. For example, an attribute can be generalized or an attribute can be replaced with a less specific value. Date of birth of the individual can be replaced by age group.

Generalization [2][3] is a good practice for privacy preserving data mining. Generalization replaces the original data with some false values. Basically quasi-identifier values are replaced with some false values which are semantically consistent with the original values. There generalization algorithms lacks in many aspects such as heavy loss in information, especially for the high-dimensional data.

Bucketization [3], is the process of partitioning the tuples in the database into tiny individual units called buckets. Then buketization algorithms provides a separation between the sensitive attribute and non-sensitive attributes by randomly distinguishing the sensitive attribute in each bucket. The data set is processed into small buckets with sensitive attributes. Original data is partitioned into buckets and within each bucket, here we apply an independent random permutation to each column.

A Feedback based algorithm[1] is a one of the best solution is to iterate all possible K-anonymity strategies and weigh each strategy based on the sample data. A heuristic approach is derived by combining the sample-based feedbacks from many crowdsourcing platforms and the multidimensional K-Anonymity approach [1]. Each iterations will result in a set of new cells. These cells are then anonymized based on their cell ranges. The anonymized samples of crowdsourcing jobs are published to collect the feedbacks for these cells.

IV. Problem Definition

Major problem while publishing crowdsourcing data are the attacks by the outsiders. External and internal attackers attack the original data for various reasons. So it becomes essential to preserve the original data from these attackers. Privacy is disrupted if anyone gets to know anything information about sensitive data. The ultimate goal of the research work is to publish an anonymized data, which will resist to internal or external

attacks. We strongly believe that when data is anonymized data is not attacked. In this modern world attackers are too brilliant to hack any of algorithm and retrieve the original data. So a combination of Slicing and m-privacy algorithm prevents the original data from internal and external attacks to a maximum extent.

V. Proposed Work

The proposed design provides a better solution to achieve privacy for crowdsourcing data publishing. This design is a combination slicing techniques with m-privacy[14] techniques. Slicing overcomes the limitations of generalization and bucketization which preserves privacy of data against the privacy breaches. m-privacy techniques guarantees that the original data is anonymised and it satisfies the privacy conditions. In this paper, we study the problem to anonymize horizontally partitioned data for multiple dataset. We consider attacks caused by insiders by plotting data providers. These insiders may not only use their own data records which is a subset of the combined data. The external background knowledge of the system and the records donated by other data providers is provided to all the insiders. The aim of our work is to publish incorporated data T in an anonymized view. Original data could be attacked from internal and external persons and reduce privacy of data. Data from multiple data providers are combined with one another and therefore an individual may have access to various databases. Here the data provider faces the different type of attack called insider attack.The data which may be tend to be attacked by the outside world is said to face outside attacker.

Definitions

B. C. M. Fung defines m-privacy as let T = ft1, t2, . . .g be a set of horizontally distributed records among n data providers P = fP1, P2, . . . , Pn., Ti _ T are set of records provided by Pi. If the records contain multiple sensitive attributes then a new sensitive attribute can be defined as a Cartesian product of all sensitive attributes. Our goal is to publish an anonymized table T* such that data cannot be hacked.

In this paper, we address the new threat and make several important contributions. We introduce the notion of distinctive privacy based approach for crowd data sourcing which will inherent data knowledge and protects anonymized data against such attacks.

Following steps are followed to attain privacy in multiple dataset

- Collect data from various data providers.
- The data collected from data providers is sliced horizontally and vertically using STA algorithm.
- Once the data is Sliced, the privacy constraint algorithms are applied to ensure privacy for all individual data.
- All the readings are noted down and the most suitable algorithm is selected

Table 1: Slicing Technique Algorithm

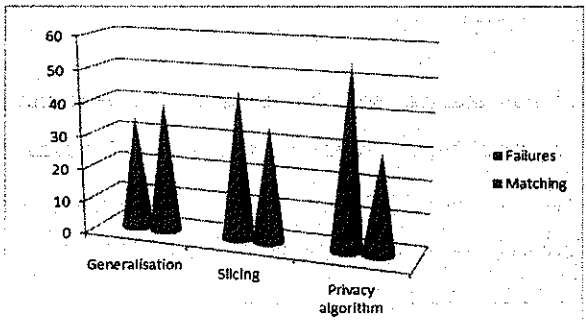
AGE	SEX	ZIPCODE	EDUCATION
34	(34,M)	(642134,4)	(34,M,4)
34	(34,M)	(642323,4)	(34,M,4)
34	(34,M)	(642323,4)	(34,M,4)
45	(45,F)	(642134,3)	(45,F,3)
45	(45,F)	(642134,3)	(45,F,3)
56	(56,F)	(644545,3)	(56,F,3)
55	(55,F)	(645534,3)	(56,F,3)

Table 2 : Distinctive privacy algorithm

AGE	SEX	ZIPCODE	EDUCATION
34	X	XXXXX	(34,M,4)
34	X	XXXXX	(34,M,4)
34	X	XXXXX	(34,M,4)
45	X	XXXXX	(45,F,3)
45	X	XXXXX	(45,F,3)
56	X	XXXXX	(56,F,3)
55	X	XXXXX	(56,F,3)

VI. Testing And Findings

The anonymization technique and many other algorithms play a vital role in preserving the privacy of published data. Each anonymization technique has its own pros and cons while preserving the privacy of data from multiple data set. Data from multiple databases are taken and they are collaborated. The original data is differentiated as Senitive attributes, Quasi identifiers and Key attributes. Many preprocessing steps are applied to data before we apply our distinct privacy preserving algorithm. Regressive experiments indicate that the algorithm works efficiently.



VII. Dicussion And Future Work

The above discussed technique helps to enhance data privacy and security when data is collected from various datasources. Many k-anonymity techniques are designed to preserve the original data from hackers, but in this modern world, hackers also too brilliant to crack these algorithms. In future, our work in developing a more effective and enhanced algorithm which will reduce limit the privacy breaches to a maximum extent.

References

[1] Sai Wu, Xiaoli Wang, Shen Wang, Zhenjie Zhang and Anthony K.H. Tung, *"K-Anonymity for crowdsourcing database"* 2013.

[2] R. J. B. Jr. and R. Agrawal, *"Data privacy through optimal k-anonymization"*, in ICDE, 2005.

[3] Sweeney.L, *"Achieving k-anonymity for privacy protection using generalization and suppression,"* International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, 2002.

[4] Tiancheng Li, Ninghui Li, Jian Zhang, Lan Molloy, *"Slicing: A new approach to privacy preserving data publishing"*. 120-150, 2012.

[5] S. Goryczka, L. Xiong, and B. C. M. Fung, *"m-Privacy for collaborative data publishing,"* In Proc. of the 7th Intl. Conf. on Collaborative Computing: Networking, Applications and Work sharing, 2013.

[6] N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. Lee, *"Centralized and distributed anonymization for high-dimensional healthcare data,"* ACM Trans. on Knowl. Discovery from Data, vol. 4,no. 4, pp. 18:1–18:33, October 2010.

[7] W. Jiang and C. Clifton, *"A secure distributed framework for achieving k-anonymity"*, VLDB J., vol. 15, no. 4, pp. 316–333, 2006.

[8] Machanavajjhala, A.Gehrke J., Kifer D. and Venkitasubramaniam M. *"l-diversity: Privacy beyond k-anonymity"* In Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE).

[9] CHAWLA, S., DWORK, C.MCSHERRY, F., SMITH, A., AND WEE, H. *"Toward privacy in public databases"*. In Proceedings of the Theory of Cryptography Conference (TCC),2005.

[10] NERGIZ, M. E., CLIFTON, C., AND NERGIZ, A. E. Multirelational k-anonymity. In proceedings of the 23rd International Conference on Data Engineering (ICDE), 2007.

[11] Y. Lindell and B. Pinkas, *"Secure multiparty computation for privacy-preserving data mining,"* The Journal of Privacy and Confidentiality,vol.1, no. 1, pp. 59–98, 2009.

[12] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu, *"Tools for privacy preserving distributed data mining,"* SIGKDD Explor. Newsl., vol. 4, pp. 28–34, December 2002.

[13] R. Sheikh, B. Kumar, and D. K. Mishra, *"A distributed k-secure sum protocol for secure multi-party computations,"* J. of Computing,vol. 2, pp. 68–72, March 2010.

[14] S. Goryczka, L. Xiong, and B. C. M. Fung, *"m-Privacy for collaborative data publishing"*, In Proc. of the 7th Intl. Conf. on Collaborative Computing: Networking, Applications and Work sharing, 2011.

[15] Tiancheng Li, N inghui Li, Senior Member, IEEE, Jian Zhang, Member, IEEE, and IanMolloy *"Slicing: A New Approach for Privacy Preserving Data Publishing"* IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. MARCH 2012.

[16] S.Kiruthika and Dr. M.Mohamed Raseen *"Enhanced Slicing Models For Preserving Privacy In Data Publication"*, ICCTET, 2013.

AUTHOR'S BIOGRAPHY

Dr.S.Sheeja, She has 12 years of teaching experience. Currently, she is guiding 6 Ph.D research scholars in computer science. Her primary research interests are related to Computer Networks, Mobile Computing, Image Processing and Data mining. She has published many research papers in national and International Journals. She is working as Associate Professor in Department of Computer Applications in Karpagam University, Coimbatore.

M.Gokilavani, she is a pursuing Ph.D, in Computer Science , in Karpagam University,Coimbatore, Tamilnadu – India. Her area of interest is Image mining, Data mining, Computer Networks, Green Computing.