Survey on Improving Data Loss in Collaborative Data Publishing

Omprakash Waghmare[1], Prof. Ashwini Tikle[2]

[1]*Student, Wainganga College of Engineering & Management*

[2]*Faculty, Wainganga College of Engineering & Management*

Abstract- **In recent years, privacy takes an important role to secure the data from various probable attackers. When data need to be shared for public advantage as required for Health care and researches, individual privacy is major concern regarding sensitive information. So while publishing such data, privacy should be conserved .While publishing collaborative data to multiple data provider's two types of problem occurs, first is outsider attack and second is insider attack. Outsider attack is by the people who are not data providers and insider attack is by colluding data provider who may use their own data records to understand the data records shared by other data providers. This problem can be overcome by combining slicing techniques with m-privacy techniques and addition of protocols as secure multiparty computation and trusted third party will increase the privacy of system effectively.**

Index Terms- Privacy, security, integrity, and protection, distributed databases

I. INTRODUCTION

Data mining is an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records, unusual records and dependencies. Data mining uses information from past data to analyze the outcome of a particular problem or situation that may arise. Data mining works to analyze data stored in data warehouses that are used to store that data that is being analyzed.

In current years, for public advantage data need to be shared. Generally data is collected from distributed databases for e.g. in case of Health care and researches, data is collected from different providers and gathered in central network. In health care all information related to patient is present in central network which includes disease details, corresponding treatment and test details.

By using anonymization technique the data is modified and then released to the public. This process is known as the privacy preservation data publishing. The attributes are classified by three types which are Key attribute, quasi identifier and sensitive attribute. Key attribute represents unique identification such as names, SSN and it is always removed before publishing. Quasi-identifiers are segments of information that are not unique identifiers but well correlated with an entity; they can be combined with other quasi-identifier to create a unique identifier. Example birth date, gender, which can be used link unionized dataset with other data. Last one is sensitive attributes example diseases, policy detail, and salary.

II. RELATED WORK

Privacy preserving data analysis and publishing has received considerable attention in recent years [1]–[3]. Most work has focused on a single data provider setting and considered the data recipient as an attacker. A large body of literature [2] assumes limited background knowledge of the attacker, and defines privacy using relaxed adversarial notion [4] by considering specific types of attacks. B.C.M. Fung et al. [2] proposed the concept of privacy preserving data publishing (PPDP). PPDP provides methods and tools for publishing useful information while preserving data privacy. These methods include K-anonymity, L- diversity and δ-Presence which encounter the attack of record linkage, attribute linkage and table linkage respectively.

In the distributed setting that we study, since each data holder knows its own records, the corruption of records is an inherent element in our attack model, and is further complicated by the collusive power of the data providers. Mohammed et al. [5] proposed SMC techniques for anonymizing distributed data using the notion of LKC-privacy to address high dimensional data. This LKC model gives better result than traditional k anonymization model. But LKC model consider only relational data and healthcare data is complex, may be a combination of relational data, transaction data and textual data.

Major problem while publishing collaborative data is attacks. Attacks are executed by insider or external attackers, which may be a single or a group of internal and external bodies that wants to violate privacy of collaborative data using background information/knowledge, also anonymized data. Privacy is violated if one knows anything about data. Main goal is to publish an anonymized view of incorporated data, D* which will be resistant to internal or external attacks. This improves the security and privacy with combination of, m-privacy techniques and slicing technique which accomplish privacy verification with better performance than encryption algorithm and provider aware (base algorithm).

According to Yehuda Lindell et al. [8], the major problem related to privacy preserving is, finding the computation function where individual privacy is preserved. For example, computation on confidential medical or criminal data in such a way that information is not revealed. This is called secure multiparty computation where number of parties wants to mutually compute some functions on their confidential inputs and through the result of this computation, parties only study the correct output and nothing else, even if some of the parties nastily plan to obtain more information. Secure multiparty computation (SMC) protocol is useful in handling above discussed scenario. D.K. Mishra et al. [9] have proposed Distributed K-secure sum protocol for secure multiparty computation. Secure sum computation of private data inputs is an example of SMC which can give a secure protocol with lower probability of data leakage. Here the idea of secure sum protocol has been enhanced which is proposed by C. Clifton et al. [10]. Distributed K-secure sum protocol compute the sum of individual data inputs with zero probability of data leakage when two neighbor parties plan to know the data of a middle party. Each data block is broken into k segments where k is equal to the number of parties. Then the segments are distributed to other parties before computation. This protocol we call as dk-Secure Sum Protocol.

III. PROPOSED SYSTEM

The proposed model provides a competent approach to achieve enhanced privacy for collaborative data publishing. This model combines slicing techniques with m-privacy techniques. Slicing overcomes the limitations of generalization and Bucketization and preserves better utility while protecting against privacy threats.

M-privacy techniques assure that the anonymized data fulfils a given privacy constraint against any range of m-colluding data providers (where m can be varied between certain ranges 1 to m), additionally it's using monotonicity constraints for efficiently checking m-privacy. Model uses Slicing for partitioning the data records and then follows m-privacy techniques and its related algorithms.

A. Design Considerations:

- We present heuristic algorithms for efficiently checking m-privacy.

- Adaptive ordering technique.

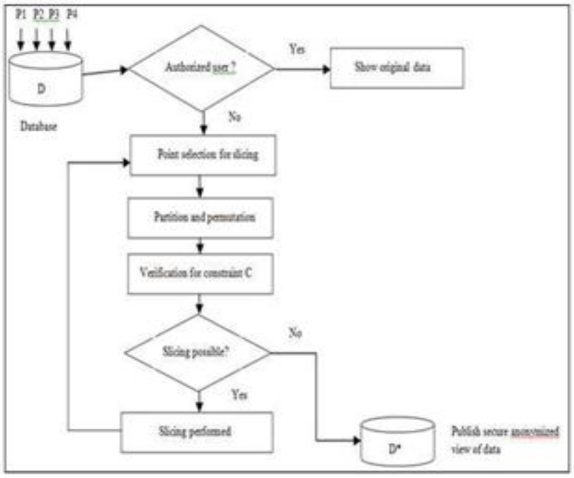
P1 P2 P3 P4
D
Database
Authorized user ?
Yes
Show original data
No
Point selection for slicing
Partition and permutation
Verification for constraint C
Slicing possible?
No
Yes
Slicing performed
D*
Publish secure anonymized view of data

Figure1. System Flow Diagram

B. Description of the Proposed Algorithm:

In Privacy for collaborative data publishing, main focus is on insider attacks. This problem can be solved by using various approaches as m-privacy, Heuristic algorithms, Data provider aware anonymization Algorithm and SMC/TTP protocols. M-Privacy helps in protecting anonymized data against m-adversary with respect to privacy constraint as K-anonymity and L-diversity. M-Privacy can also be sure when there are duplicate records; it also contains syntactic privacy constraint, monotonicity of privacy constraints and differential privacy constraint.

IV. CONCLUSION AND FUTURE WORK

In this paper we considered a new type of potential attackers in collaborative data publishing – a coalition of data providers, called m-adversary. Privacy threats introduced by m-adversaries are modeled by a new privacy notion, m-privacy, defined with respect to a constraint C. We presented heuristics to verify m-privacy w.r.t. C. A few of them check m-privacy for EG monotonic C, and use adaptive ordering techniques for higher efficiency. We also presented a provider-aware anonymization algorithm with an adaptive verification strategy to ensure high utility and m-privacy of anonymized data. Experimental results confirmed that our heuristics perform better or comparable with existing algorithms in terms of efficiency and utility.

Above discussed approaches help to enhance the data privacy and security when data is collected from various resources and output should be in collaborative style. In future, this system can consider for data, which are distributed in ad hoc grid computing. Also the system can be considered for set valued data. The consumption of various protocols can address various data publishing paradigms. The consumption of these protocols can make collaborative data publishing more effective and enhanced using m-privacy.

REFERENCES

[1] C. Dwork, "Differential privacy: a survey of results," in Proc. of the 5th Intl. Conf. on Theory and Applications of Models of Computation, 2008, pp. 1–19.

[2] B. C. M. Fung, K.Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," ACM Comput. Surv., vol. 42, pp. 14:1–14:53, June 2010.

[3] C. Dwork, "A firm foundation for private data analysis," Commun. ACM, vol. 54, pp. 86–95, January 2011

[4] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-Diversity: Privacy beyond k- anonymity," in ICDE, 2006, p. 24.

[5] 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.

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

[7] 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).

[8] 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.

[9] 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.

[10] 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.

[11] CHAWLA, S., DWORK, C.MCSHER of Cryptography Conference (TCC), 2005.

[12] RY, F., SMITH, A., AND WEE, H. "Toward privacy in public databases". In Proceedings of the Theory

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