Improving Data Loss in Collaborative Data Publishing

Omprakash Waghmare[1], Ashwini Tikle[2]
[1,2]*Department of Computer Science and Engineering, Wainganga college of Engineering and Management, Nagpur (MS), India*

***Abstract*- Nowadays, privacy becomes important to secure the data from various probable attackers. When data is sharing for public advantage as required for Health care and researches, individual privacy become major task regarding sensitive information. So while publishing this type of data, privacy should be preserved .When collaborative data was published to multiple data provider's two types of attack takes place, outsider attack is first attack and second is the insider attack. Outsider attack means the people those are different from data providers and insider attack is by data provider those are the neighbour of each other who may use their own data records to understand the other provider's data records shared by them. This problem can be overcome by combining Record Elimination techniques with mprivacy techniques and addition secure multiparty computation protocol and trusted third party will increase the privacy effectively of system**

Index-Terms- Privacy, Security, Integrity, Protection, Distributed Databases

I. INTRODUCTION

Data mining is an young and promising field of computer science, is the computational process of mining a large data sets to retrieve a pattern which is relevant to one need involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the mining of data process means to extract relevant information from a large data set and convert it into a structure which is an understandable for further use. The actual data mining concern is the automatic or semi-automatic analysis of large number of data to extract previously not known interesting patterns such as groups of data records, unusual records and dependencies. Data mining use previous information to analyse the outcome of a particular problem or situation that may arise later. Data mining works to a extract data stored in large databases that are used to store that data that is being analysed.

Nowadays, for public advantage data is sharing among different organization. Generally data is collected from distributed databases for e.g. in Health care organization and for researches, data is collected from different providers and gathered in central network. In health care organization all information of patient is present in central network which includes details of diseases, corresponding treatment and test details.

By using anonymization technique the data is modified and then released for public use. This process known as the privacy preservation data publishing. The attributes are categorized into three types which are Key attribute, quasi identifier and sensitive attribute. Key attribute represents unique identity such as names, SSN and it always removed before publishing. Quasi-identifiers are segments of information that are identifiers correlated with an entity; they can be combined with other quasi-identifier to create a identifier to form unique. Example birth date, gender, which can be used link unionised dataset with other dataset. Last one is attributes which we called sensitive example diseases, policy detail, and salary.

II. RELATED WORK

Collaborative data publishing has received considerable attention in recent years [1]–[3]. Most work has center of interest on a single data provider setting and considered that recipient as an attacker. A large body of literature [2] assumes some 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 preserving of privacy and publishing of data. Privacy Preserving Data Publishing provides methods and tools for publishing useful information while preserving privacy of data. These methods include K-anonymity, L-diversity and δ-Presence

which encounter the attack of linkage by record, attribute linkage and table linkage respectively.

In the distributed setting that we study, since each data holder knows its own records, the records which is corrupt 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 represent high dimensional data. This LKC model gives better result than traditional k anonymization model. But LKC model consider relational data only 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 done by insider or external attackers, which may be a single or a group of internal and external bodies that wants to destroy privacy of collaborative data using background information/knowledge, also anonymized data. Privacy is corrupt if one knows anything about data. Main goal is to publish an anonymized view of incorporated data, D* in which internal or external attacks is not possible. This improves the security and privacy with combination of, mprivacy techniques and slicing technique which completed 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, to find the computation function where individual privacy is preserved. For example, computation on secretl medical or criminal data in such a way that information is not revealed. This is called secure multiparty computation where many parties wants to mutually compute some functions on their confidential inputs and through the result of such computation, parties only consider the correct output and nothing else, even if some of the parties nastily plan to extract more and 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 personal data inputs is an example of SMC which can give a secure protocol with lower probability of data leakage.

In this paper, the idea of secure sum protocol has been extend which is proposed by C. Clifton et al. [10]. Distributed K-secure sum protocol estimate the sum of individual data inputs with zero probability of data leakage when two neighbour parties plan to get the data of a middle party. Each data block is broken into k segments where k is equal to the n number of parties. Then the segments are divided 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 strengthen privacy of collaborative data publishing. This model integrate Record elimination techniques with m-privacy techniques. Slicing overcomes the limitations of generalization process and Bucketization process and preserves better utility while protecting against privacy threats.

Proposed model have 4 module .First Aggregation, Second Suppresssion, Third Record elimination ,Fourth Generalization After than comparison calculation of privacy gain and information loss, then comparison of result of new technique with old techniques.

A. Aggregation

Here data gathered from various publisher are combine into larger dataset, then preprocessing technique applied, Preprocessing technique remove all the inconsistence value from the dataset and prepare a dataset for a further process.

Srno,	Zipcode	Age	Salary	disease
1	47677	29	50000	AIDS
2	47602	28	40000	TB
3	47678	29	50000	AIDS
4	47905	36	60000	TB
5	47909	52	110000	Fever
6	47906	36	80000	Cancer
7	47605	30	70000	Pneumonia
8	47673	36	90000	Cancer
9	47607	32	100000	Cancer

Table I: Published data

Process Flow Diagram:

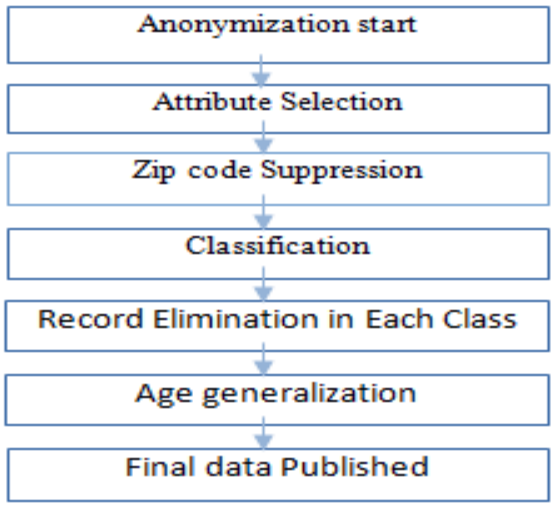


Figure: Flow Diagram of Record Elimination Process.

B. Suppression

In suppression first find out the longest common substring in quasi identifier zipcode, then replaces the common element by symbol. And form the group with common matching substring as shown in figure.

Srno	Zipcode	Age	Salary	disease	Group
1	4760*	28	40000	TB	C1
2	4760*	30	70000	Pneumonia	
3	4760*	32	100000	Cancer	
4	4790*	36	60000	TB	C2
5	4790*	52	110000	Fever	
6	4790*	36	80000	Cancer	
7	4767*	29	50000	AIDS	C3
8	4767*	29	50000	AIDS	
9	4767*	36	90000	Cancer	

C. Record Elimination

The output dataset which is form from above process, applying record elimination technique in which common record which appear more than one is eliminate from the group.

From above group, Since, Record 8 and Record 7 are identical, Therefore We can eliminate One record from group C3.

8	4767	29	50000	AIDS

So, new group is formed as shown in figure

Srno	Zipcode	Age	Salary	disease	Group
1	4760*	28	40000	TB	C1
2	4760*	30	70000	Pneumonia	
3	4760*	32	100000	Cancer	
4	4790*	36	60000	TB	C2
5	4790*	52	110000	Fever	
6	4790*	36	80000	Cancer	
7	4767*	29	50000	AIDS	C3
9	4767*	36	90000	Cancer	

D. Generalization

The output dataset which is form from above process in that find out the lower value in the group assign as minimum value, Similarly find out the lower value in the group assign as maximum value and form the pair as min<=max.as shown in figure.

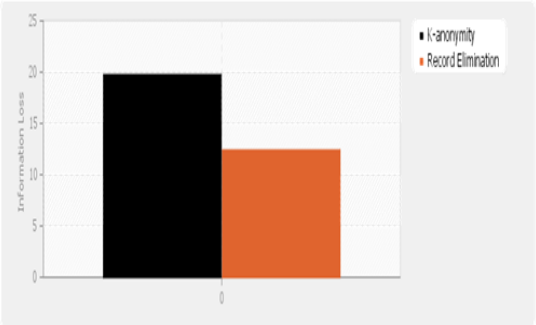
Srno	Zipcode	Age	Salary	disease
1	4760*	28<=32	40000	TB
2	4760*	28<=32	70000	PNEUMONIA
3	4760*	28<=32	100000	CANCER
4	4790*	36<=52	60000	TB
5	4790*	36<=52	110000	Fever
6	4790*	36<=52	80000	Cancer
7	4767*	29<=36	50000	AIDS
8	4767*	29<=36	90000	Cancer

The performance of the proposed algorithm is evaluated in terms of two data metrics namely information loss and privacy gain. The proposed method and three existing methods namely k-anonymity (k=3), ℓ-diversity(l=3) and t-closeness are experimented with the same data set and their performance were compared in terms of information loss and privacy gain. The following formulae are used to measure information loss ILoss[9] and privacy gain PG[11,12].

ILOSS(vg) = |vg|-1/da

where; |vg| is the number of domain values that are descendants of vg. DAis the number of domain values in the attribute A of vg.

Figure showing the comparison of Information loss between k-anomity and record elimination technique



IV. CONCLUSION AND FUTURE WORK

In this information age, data published in web pages are growing enormously every year. While utilizing the data for research purpose, privacy of the

individuals whose data are published should not be challenged. The proposed method attempts at static micro data only which contain numeric quasi identifiers. Further research is in progress to include various extended data publishing scenarios such as multiple view publishing, Anonymizing sequential release with new attributes and incrementally update data records as well as non-numeric quasi identifiers..

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.