ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 7, July 2014

Secure Multiparty Data Anonymization and Integration with *m*-Privacy

M.Ashok Kumar, R.Nandhakumar,

master of philosophy (CS), KMG College of Arts and Science, Gudiyatham, India.

R&D, Xtream Technologies, Chennai, India.

ABSTRACT*:* We propose secure multi-party computation protocols for collaborative data publishing with m-privacy. All protocols are extensively analyzed and their security and efficiency are formally proved. Experiments on real-life datasets suggest that our approach achieves better or comparable utility and efficiency than existing and baseline algorithms while satisfying m-privacy. We consider the collaborative data publishing problem for anonym zing horizontally partitioned data at multiple data providers. We consider a new type of "insider attack" by colluding data providers who may use their own data records (a subset of the overall data) to infer the data records contributed by other data providers. The paper addresses this new threat, and makes several contributions. First, we introduce the notion of m-privacy, which guarantees that the anonym zed data satisfies a given privacy constraint against any group of up to m colluding data providers. Second, we present heuristic algorithms exploiting the monotonicity of privacy constraints for efficiently checking m-privacy given a group of records. Third, we present a data provider-aware anonymization algorithm with adaptive m-privacy checking strategies to ensure high utility and m-privacy of anonym zed data with efficiency.

KEYWORDS: Horizontal Division, Vertical Division, Encryption, Privacy, Database.

I. INTRODUCTION

The goal of privacy preserving data mining is to develop data mining methods without increasing the risk of misuse of the data used to generate those methods. The topic of privacy preserving data mining has been extensively studied by the data mining community in recent years. Many effective Techniques for privacy preserving data mining have been proposed that use some transformation method on the original data in order to perform the privacy preservation. The transformed dataset is made available for mining and must meet privacy requirements without losing the benefit of mining. We classify them into the following three categories:

Randomization method is a popular method in current privacy preserving data mining studies. It masks the values of the records by adding noise to the original data. The noise added is sufficiently large so that the individual values of the records can no longer be recovered. However, the probability distribution of the aggregate data can be recovered and subsequently used for privacy-preservation purposes. In general, randomization method aims at finding an appropriate

II. RELATED WORK

Following method plays an important role in our project work to protect data from insider attack to improve security.

i) The Anonymization Method:

Anonymization method aims at making the individual record be indistinguishable among a group records by using techniques of generalization and suppression. The representative anonymization method is k-anonymity. The motivating factor behind the k-anonymity approach is that many attributes in the data can often be considered quasi-identifiers which can be used in conjunction with public records in order to uniquely

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 7, July 2014

identify the records. Many advanced methods, such as, p-sensitive, (a, k)-anonymity-anonymity,t-closeness, l-diversity and M-invariance, Personalized anonymity etc. have been proposed. The anonymization method can ensure that the transformed data is true, but it also results in information loss in some extent.

ii) The Encryption Method:

Encryption method mainly resolves the problems that people jointly conduct mining tasks based on the private inputs they provide. These mining tasks could occur between mutual un-trusted parties, or even between competitors, therefore, protecting privacy becomes a primary concern in distributed data mining setting. The two different approaches for distributed privacy preserving data mining are method on horizontally partitioned data and that on vertically partitioned data. The encryption method may not be so efficient but it ensures that the transformed data is exact and secure.

We consider the collaborative data publishing setting with horizontally partitioned data across multiple data providers. These contribute a subset of records Ti. Even a data provider could be the data owner themselves who contribute their own records. This is a common observed scenario in social networking and recommendation systems. Our main aim is that a data recipient including the data providers will not be able to compromise the privacy of the individual records provided by other parties by publishing an anonymized view of the integrated data such.

Attacks by Data Providers Using Anonymized Data and Their Own Data: Each data provider such as P1 in Figure 1 can also be anyonymised data T* and his own data (T1) additional information about other records. If the attacks by the external recipient in the first attack scenario are compared with those of data providers, each provider has more knowledge of their own data records. This attack scenario will be further worsened when multiple data providers collude with each other.

T_a^*

Provider	*Name*	Age	Zip	Disease
P_1	*Alice*	[20-30]	*****	Cancer
P_1	*Emily*	[20-30]	*****	Asthma
P_3	*Sara*	[20-30]	*****	Epilepsy
P_1	*Bob*	[31-35]	*****	Asthma
P_2	*John*	[31-35]	*****	Flu
P_4	*Olga*	[31-35]	*****	Cancer
P_4	*Frank*	[31-35]	*****	Asthma
P_2	*Dorothy*	[36-40]	*****	Cancer
P_2	*Mark*	[36-40]	*****	Flu
P_3	*Cecilia*	[36-40]	*****	Flu

FIGURE 1

T_b^*

Provider	*Name*	Age	Zip	Disease
P_1	*Alice*	[20-40]	*****	Cancer
P_2	*Mark*	[20-40]	*****	Flu
P_3	*Sara*	[20-40]	*****	Epilepsy
P_1	*Emily*	[20-40]	987**	Asthma
P_2	*Dorothy*	[20-40]	987**	Cancer
P_3	*Cecilia*	[20-40]	987**	Flu
P_1	*Bob*	[20-40]	123**	Asthma
P_4	*Olga*	[20-40]	123**	Cancer
P_4	*Frank*	[20-40]	123**	Asthma
P_2	*John*	[20-40]	123**	Flu

FIGURE: 2

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 7, July 2014

IV.PROBLEM STATEMENT

The proposed project work focuses on the problem of privacy for data publishing for the improvement of database and also overcome the problem of "insider attack" to provide a better security.

We consider the collaborative data publishing setting with horizontally distributed data across multiple data providers. Each data provider contributes subset of records Ti.As also each record has an owner, whose identity shall be protected. Each record attribute is either a sensitive attribute, which carries sensitive information about data owners, an identifier, which directly identifies the owner, or a quasi- identifier (QID), which may identify the owner if joined with a publicly known dataset. A data provider could also be the data owner itself who is contributing its own records. wants to breach privacy of data using background knowledge, as well as anonymized data. Privacy is breached if one learns anything about data.

Privacy preserving data publishing for a single database has been extensively studied in recent years. A large body of work contributes to data anonymization that transforms a dataset to meet a privacy principle such as k- anonymity using techniques such as generalization or suppression (removal) so that it does not contain individually identifiable information There are a number of potential approaches one may apply to enable privacy preserving data publishing for distributed databases. A naive approach is for each data custodian to perform data anonymization independently. Data recipients or clients can then query the individual anonymized databases or its integrated view. One main drawback is that data is nonymized before the integration and hence will cause the data utility to suffer. In addition, individual databases reveal their ownership of the anonymized data. An alternative approach assumes an existence of third party that can be trusted by each of the data owners. In this scenario, data owners send their data to this trusted third party where data integration and anonymization are performed. Then, clients can query the centralized database. However, finding such a trusted third party is not always feasible.

V. CONCLUSION & FUTURE WORK

We carried out a wide survey of the different approaches for privacy preserving data mining, and analyzed the major algorithms available for each method and pointed out the existing drawback. All the purposed methods are able to achieve our goal of privacy preservation. Hence there is a need to further perfect those approaches or develop some well-organized methods.

For this, we recognize that the following problems should be concentrated on.

1) Privacy and accuracy is a pair of contradiction; improving one usually incurs a cost in the other. How to apply various optimizations to achieve a trade-off should be deeply researched.
2) Side-effects are unavoidable in data sanitization process. How to measure and reduce their Negative impact on privacy preserving needs to be considered carefully and define some metrics for measuring them.
3) In distributed privacy preserving data mining areas, we should try to develop more efficient algorithms and look for abalance between disclosure cost, computation cost and communication cost.
4) How to deploy privacy-preserving techniques into practical applications also needs to be further studied.

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. Finally, we emphasize that privacy-preserving technology solves only one side of the problem. It is equally important to identify and overcome the nontechnical difficulties faced by decision makers when they deploy a privacy-preserving technology. Their typical concerns include the degradation of data/service quality, loss of valuable information, increased costs, and increased complexity. We believe that cross- disciplinary research is the key to remove these obstacles, and urge computer scientists in the privacy protection field to conduct cross-disciplinary research with social scientists in sociology, psychology, and public policy studies. In future it is used for Improvement of algorithm for integrated databases, like combination of Oracle,MySQLand MS-SQL databases. Making the project OS independent.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 7, July 2014

REFERENCES

1 S. Goryczka, L. Xiong, and B. C. M. Fung, "m-privacy for collaborative data Publishing," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING VOL: PP NO: 99 YEAR 2013

2 N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. Lee, "Centralizedand distributed anonymization for high- dimensional healthcare data," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 4, no. 4, pp.18:1–18:33, October 2010.

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

4 L. Sweeney, "Datafly: A system for providing anonymity in medical data," in Proc. of the IFIP TC11 WG11.3 Eleventh Intl. Conf. on Database Security XI: Status and Prospects, 1998, pp. 356–381.

5 W. Jiang and C. Clifton, "Privacy-preserving distributed k-anonymity," in Data and Applications Security XIX, ser. Lecture Notes in Computer Science,2005, vol. 3654, pp. 924–924.

6 N. Mohammed, B. C. M. Fung, K. Wang, and P. C. K. Hung, "Privacy preserving Data mashup," in Proc. of the 12th Intl. Conf. on Extending Database Technology, 2009, pp. 228–239.

7 P. Jurczyk and L. Xiong, "Distributed anonymization: Achieving privacy for both data subjects and data providers," in DBSec, 2009, pp. 191–207.

8 I. Mironov, O. Pandey, O. Reingold, and S. Vadhan, "Computational differential privacy," in Advances in Cryptology CRYPTO 2009, ser. Lecture Notes in Computer Science, vol. 5677, 2009, pp. 126–142.

9 K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: efficient full-domain k anonymity," in Proc. of the 2005 ACM SIGMOD Intl. Conf. on Management of Data, 2005, pp. 49–60.

BIOGRAPHY

M. Ashok Kumar is a Research Scholar in computer science Department in KMG college of Arts and Science, Gudiyatham, Tamilnadu, India. Affiliated by Thiruvalluvar University. He receiver master of computer Application(MCA) degree in 2013 from SNCET, Tirupattur. His interest in data mining.

R. Nandha Kumar is a Research and Development, Xtream Technologies, Chennai, TamilNadu, India. He received Master of Technology (M.TECH) degree in Information Technology from Manonmaniyam University in 2011. His interests include Data Mining and Big data, Neural Networks, Networking, Cloud Computing etc.