International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622
International Conference on Industrial Automation and Computing (ICIAC- 12-13th April 2014)

RESEARCH ARTICLE OPEN ACCESS

Data Publishing In M-Privacy Using Anonymization and Randamization Method

Ms.Priya V. Mundafale*, Prof.M.B.Chandak**, Prof. Gurudev Sawarkar***
***(Department of Computer Science&EngineeringCOE&M, RTM Nagpur University, India*
Email: pwpriyawankhede1@gmail.com)
*** (Department of Computer Science & Engineering RamdeobabaCollege, RTM Nagpur University, India.*
Email: chandakmb@gmail.com)
**** (Department of Computer Science& EngineeringCOE&M, RTM Nagpur University, India*
Email:sawarkar92@gmail.com)

ABSTRACT
In this paper, our focus is on the study of Data mining is the extraction of interesting patterns or knowledge from huge amount of data. The goal of data mining is to extract knowledge from data. In recent years, with the explosive development in Internet, data storage and data processing technologies, privacy preservation has been one of the greater concerns in data mining. The automated prospective analysis offered by data mining move beyond the analysis of past events provided by retrospective tools typical decision support system. A number of methods and techniques have been developed for privacy preserving data mining. Privacy preserving data mining is an important issue in the areas of data mining and security on private data in the following scenario: There is an increasing need for sharing data repositories containing personal information across multiple distributed, possibly untrusted, and private databases. Such data sharing is subject to constraints imposed by privacy of data subjects as well as data confidentiality of institutions or data providers. With these recent changes in trends, sensitive data is now easily available for malicious use. The main concern is that sensitive information should not be disclosed. We developed a set of decentralized protocols that enable data sharing for horizontally partitioned and vertical partitioned databases given these constraints.
Keywords–Anonymization, Horizontal-division, Randamisation and Vertical-division,

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. To protect sensitivity or confidentiality of shared data, it often needs to be sanitized before it can be distributed and analyzed. If one person sends a file, there may be information on the file that leaves a trail to the sender. The information sent by the sender's may be traced from the data logged after the file is sent. Since , the file is anonymized, data associated with it being sent impossible traced to the sender. A number of effective methods for privacy preserving data mining have been proposed. Most methods use some form of transformation 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. Privacy preserving data mining perturbed data maintain statistical properties, but hard to guess original data from perturbed value

II. BACKGROUND

Following method plays an important role in our project work to protect data from insider attack to improve security. . We classify them into the following two methods:

2.1 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 anonymization.A process that removes or replaces identity information from a communication or record.The anonymization method can ensure that the transformed data is true, but it also results in information loss in some extent.

2.2 The Randomization Method:

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 large noise added 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. A well known method for privacy-preserving data mining is that of randomization. In randomization, we add noise to the data so that the behavior of the individual records is masked. However, the aggregate behavior of the data distribution can be reconstructed by subtracting out the noise from the data. The reconstructed distribution is often sufficient for a variety of data mining tasks such as classification. In this paper, we will provide a survey of the randomization method for privacy-preserving data mining.

In general, randomization method aims at finding an appropriate Balance between privacy preservation and knowledge discovery. Representative randomization methods Include random-noise-based perturbation and Randomized Response scheme. The randomization method is more efficient method in privacy preserving data mining. It's results in high information loss. Randomisation 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. In distributed privacy preserving data mining there are two different approaches such as the method on horizontally partitioned data and that on vertically partitioned data.

Advantages:

i. It is support to protect individual's privacy.
ii. Due to its simple working anyone can use and it is more efficient.
Iii. The randomization method is very simple method and when we collect the data which can be easily applied.

Disadvantages

i. It is not suitable when multiple attribute databases are used.
ii. It is very slow technique because when data collector collects the data from data provider the data provider adds some noise in data and to reorder that data it takes more time.

III. RELATED WORK

We consider the collaborative data publishing setting with horizontally partitioned data across multiple data providers, each contributing combining different tables. As a special case, a data provider could be the data owner itself who is contributing its own records. This is a common scenario in social networking and recommendation systems. Our aim is to publish an anonymized view of the integrated data such that a data recipient including the data providers will not be able to compromise the privacy of the individual records provided by other parties.

Attacks by Data Providers Using Anonymized Data are an obvious part of insider attacks. Compared to the attack by the external recipient, each provider has additional data knowledge of their itself records, which can help with the attacker attack. This issue can be further worsened at the time when multiple data providers collude with each other.

IV. PROBLEM STATEMENT

We focuses on the problem of privacy preserving 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 contributing combining different tables. Every 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. As a special case, a data provider could be the data owner itself who is contributing its own records. A data recipient may have access to some background knowledge data, which available publicly about released data, e.g., Census datasets. Our aim is to publish an anonymized view of the integrated data, which will be immune to attacks. Attacks are run by attackers, i.e., a single or a group (coalition) of external or internal entities that wants to privacy of data using anonymized data as well as Background knowledge. 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 independently perform data anonymization. clients or Data recipients can then query the individual anonymized databases. One main disadvantage of this approach is that data is anonymized before the integration and hence will cause the data utility to suffer. In addition, individual databases reveal their ownership of the anonymzed 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. After, clients can query the

centralized database. However, finding such a trusted third party is not always feasible.

In this proposed work we are dealing with three table namely city , country and country language and we are first trying to implement the horizontal partitioning of the data for the city and the country table. We are using the logic based on the anonymisation in which we are taking the rowise merging of the city and the country table .This horizontal technique will try to provide some amount of privacy as the data is seen to be mismatched .The fig1.below shows the city table , fig2. shows the country tale and the last fig3. Show the horizontal anynymisation performed using the two said tables.

1	Kabul	AFG	Kabol	1780000
2	Qandahar	AFG	Qandahar	237500
3	Herat	AFG	Herat	186800
4	Mazar-e-Sharif	AFG	Balkh	127800
5	Amsterdam	NLD	Noord-Holland	731200
6	Rotterdam	NLD	Zuid-Holland	593321
7	Haag	NLD	Zuid-Holland	440900
8	Utrecht	NLD	Utrecht	234323
9	Eindhoven	NLD	Noord-Brabant	201843
10	Tilburg	NLD	Noord-Brabant	193238
11	Groningen	NLD	Groningen	172701
12	Breda	NLD	Noord-Brabant	160398

Fig. 1 . City table

Code	Name	Continent	Region	SurfaceArea	IndepYear	Population	LifeExpt	GNP	GNPOld	LocalName
ABW	Aruba	North America	Caribbean	193.00		103000	78.4	828.00	793.00	Aruba
AFG	Afghanistan	Asia	Southern and Central Asia	652090.00	1919	22720000	45.9	5976.00		Afganistan/Afqanestan
AGO	Angola	Africa	Central Africa	1246700.00	1975	12878000	38.3	6648.00	7984.00	Angola
AIA	Anguilla	North America	Caribbean	96.00		8000	76.1	63.20		Anguilla
ALB	Albania	Europe	Southern Europe	28748.00	1912	3401200	71.6	3205.00	2500.00	Shqipëria
AND	Andorra	Europe	Southern Europe	468.00	1278	78000	83.5	1630.00		Andorra
ANT	Netherlands Antilles	North America	Caribbean	800.00		217000	74.7	1941.00		Nederlandse Antillen
ARE	United Arab Emirates	Asia	Middle East	83600.00	1971	2441000	74.1	37966.00	36846.00	Al-Imarat al- Arabiya al-
ARG	Argentina	South America	South America	2780400.00	1816	37032000	75.1	340238.00	323310.00	Argentina
ARM	Armenia	Asia	Middle East	29800.00	1991	3520000	66.4	1813.00	1627.00	Hajastan
ASM	American Samoa	Oceania	Polynesia	199.00		68000	75.1	334.00		Amerika Samoa

Fig 2. Country table

```
Kabul,AFG,Kabol,1780000.0
ABW,Aruba,North America,Caribbean,193.0,0,103000.0,78.4,828.0,793.0,Nonmetropolitan Territory of The Netherlands,Beatrix
Qandahar,AFG,Qandahar,237500.0
AFG,Afghanistan,Asia,Southern and Central Asia,652090.0,1919,2.272E7,45.9,5976.0,0.0,Afganistan/Afqanestan,Islamic Emirate,Mohamm
Herat,AFG,Herat,186800.0
AGO,Angola,Africa,Central Africa,1246700.0,1975,1.2878E7,38.3,6648.0,7984.0, Republic,José Eduardo dos Santos,56,AO
Mazar-e-Sharif,AFG,Balkh,127800.0
AIA,Anguilla,North America,Caribbean,96.0,0,8000.0,76.1,63.2,0.0,Dependent Territory of the UK,Elisabeth II,62,AI
Amsterdam,NLD,Noord-Holland,731200.0
ALB,Albania,Europe,Southern Europe,28748.0,1912,3401200.0,71.6,3205.0,2500.0,Shqipëria,Republic,Rexhep Mejdani,34,AL
Rotterdam,NLD,Zuid-Holland,593321.0
AND,Andorra,Europe,Southern Europe,468.0,1278,78000.0,83.5,1630.0,0.0, Parliamentary Coprincipality,55,AD
Haag,NLD,Zuid-Holland,440900.0
ANT,Netherlands Antilles,North America,Caribbean,800.0,0,217000.0,74.7,1941.0,0.0,Nederlandse Antillen,Nonmetropolitan Territory of
Utrecht,NLD,Utrecht,234323.0
ARE,United Arab Emirates,Asia,Middle East,83600.0,1971,2441000.0,74.1,37966.0,36846.0,Al-Imarat al-'Arabiya al-Muttahida,Emirate Fe
Eindhoven,NLD,Noord-Brabant,201843.0
ARG,Argentina,South America,South America,2780400.0,1816,3.7032E7,75.1,340238.0,323310.0, Federal Republic,Fernando
```

Fig 3. Horizontally partitioning data.

V. CONCLUSION AND 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 only approximate to aim goal of privacy preservation. So now we need to perfect those approaches further or develop some efficient methods.

For this, the following problems should be recognized and concentrated on.

1) Accuracy and Privacy are contradiction; improving one usually incurs a cost in the other. How to apply various optimizations is that 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. Some metrics for measuring these are also needed.
3) In distributed privacy preserving data mining areas, we should try to develop more efficient algorithms and look for a balance 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, MySQL and MS-SQL databases. Making the project OS independent

REFERENCES

Journal Papers:

[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.*

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

[9] 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.*

Proceedings Papers:

[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.*

[8] 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.*

[10] 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.*