



**INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH
TECHNOLOGY**

**PRIVACY PRESERVING MULTIPARTY COLLABORATIVE DATA MINING FOR
MULTIPLE SERVICE PROVIDERS**

Shrishti pawar ^{*1}, Hare Ram Shah ²

^{*1}Research Scholar, Gyan Ganga Institute of Technology And Science, Jabalpur, M.P, India.

²Associate Professor, Gyan Ganga Institute of Technology And Science, Jabalpur, M.P, India.

^{*}Department of Computer Science & Engg.

ABSTRACT

we present a new multiple service provider model of operation for the Internet delivery of data mining services. So the collaboration becomes especially important because of the mutual benefit it brings. For this kind of collaboration, data's privacy becomes extremely important: all the parties of the collaboration promise to provide their private data to the collaboration, but neither of them wants each other or any third party to learn much about their private data. One of the major problems that accompany with the huge collection or repository of data is confidentiality. The need for privacy is sometimes due to law or can be motivated by business interests. Performance of privacy preserving collaborative data using secure multiparty computation is evaluated with attack resistance rate measured in terms of time, number of session and participants and memory for privacy preservation.

Many anonymization techniques, such as bucketization and generalization, have been designed for privacy preserving publishing. Present work has shown that generalization loses considerable amount of information, especially for high-dimensional data. Bucketization, not clearly prevent membership disclosure and do not clear separation between sensitive and quasi-identifying attributes.

KEYWORDS: PPDM, collaborative data mining, secure multiparty computation

INTRODUCTION

Data mining or knowledge discovery techniques like association rule mining, classification, clustering, sequence mining, etc. have already been most in-demand in today information world [1]. Successful application of these techniques has become demonstrated in many areas like marketing, research, business, product control and a few other locations that social, humanitarian activities and social. Privacy Preserving Data Mining is a vital feature which each mining system must support. This selection actually secures the private and sensitive information that the database owners don't want to reveal. The sensitive data could be anything like Identification Number, Name, Address, and Disease etc. [2]

Privacy preserving data mining work required as follows:

- Privacy Preserving Data Publishing:
- Modifying the record values to preserve privacy
- Query Auditing

Privacy Preserving Data Publishing:

Privacy preserving data publishing techniques try to study different techniques associated with privacy. These techniques consist of:

THE RANDOMIZATION METHOD:

- In this technique, any random value is added to the original value of the facts to mask the values of the data. The sound is added in large amount so that the original data value is not recovered [3].
- The K-Anonymity Model and L-Diversity: In K-Anonymity, the techniques like generalization and suppression were introduced to normalize data representation. In order to decrease the identification threat, every tuple in the database must be indistinguishable. The L-Diversity technique was introduced to overcome some weakness of K-Anonymity. The novel concept of intra group variety of sensitive and private values

within anonymization scheme was discovered [4].

- Distributed Privacy Preserving: Sometimes, some users do not desire to release their information to other users. But the individual users are interested in achieving the aggregate results from the data set which are divided among the users. [5]

MODIFYING THE RECORD VALUES TO PRESERVE PRIVACY

Using these methods, the association rules are encrypted in order to secure the data. Above technique, Association Rule Hiding methods were used to preserve privacy.

QUERY AUDITING

Query auditing technique, either the result of the query is modified or the result of the query is restricted. Many perturbation methods are applied to achieve this. [6]

Privacy-preserving data publishing (PPDP) is, a task of the most importance is to develop methods and tools for publishing data in a more hostile environment, so that the published data remains practically useful while individual privacy is preserved. In Figure 1 a typical scenario for data collection and publishing is described. In the data collection phase, the data publisher collects data from record owners (e.g. Ravi, Seema, Lokesh and Dilip). In the data publishing phase, the facts publisher releases the together data to a data miner or to the public, called the data recipient, who will then conduct data mining on the published data. data mining has a broad sense, not unavoidably limited to pattern mining or model building, In this survey.

For example, a organization (hospital) collects data from patients and publishes the long-suffering records to an external medical center. In this example, the organization (hospital) is the data publisher, patients are verification owners, and the health check center is the data beneficiary. The data mining conduct at the health check center could be anything from a simple count of the number of men with diabetes to a sophisticated cluster analysis.

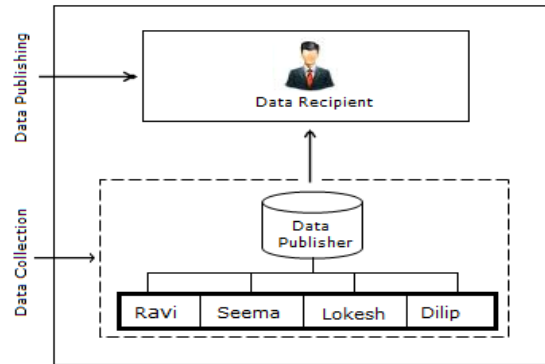


Fig. 1. Data Publisher and data collection.

Two models of data publishers:

1. Trusted model
2. Untrusted model

1. Trusted model: the data publisher is trust worthy and record owners are willing to provide their personal information to the data publisher; however, the trust is not transitive to the data recipient. we assume the trusted model of data publishers and consider privacy issues in the data-publishing phase in this survey.

2. Untrusted model: The data publisher is not trusted and may attempt to identify sensitive information from record owners. Various cryptographic solutions anonymous communications and statistical were proposed to collect records anonymously from their owners without revealing the owners' identity.

PRIVACY AND ATTACK MODELS

Table A. privacy and attack model

Privacy Model	Attack Model			
	Record Linkage	Attribute Linkage	Table Linkage	Probabilistic Attack
k-Anonymity	√			
Multi k-Anonymity	√			
l-Diversity	√	√		
Confidence Bounding		√		
(a,k)Anonymity		√		
(X,Y)Privacy	√	√		
(k,e)Anonymity	√	√		√
(e,m)Anonymity		√	√	
Personalized Privacy		√	√	√
t-Closeness			√	√
Distributional Privacy			√	√

OPERATIONS FOR ANONYMIZATION

Generalization and Suppression

- Generalization
 - Replace the original value by a semantically consistent but less specific value
- Suppression
 - Data not released at all
 - Can be Cell-Level or (more commonly) Tuple-Level

Table 1. Sample of Data

	Non-Sensitive Data			Sensitive Data	
	Zip	Age	Nationality	Name	Condition
1	13053	28	Indian	Ravi	Heart Disease
2	13067	29	Indian	Lokesh	Heart Disease
3	13053	35	Indian	Seema	Viral Infection
4	13067	36	Indian	Dilip	Cancer

Table 2. Data with generalization and suppression

	Non-Sensitive Data			Sensitive Data	
	Zip	Age	Nationality	Name	Condition
1	13053	<40	*	Kumar	Heart Disease
2	13067	<40	*	Bob	Heart Disease
3	13053	<40	*	Ivan	Viral Infection
4	13067	<40	*	Umeko	Cancer



Anatomization and Permutation

Anatomization [8]. Anatomization does not modify the quasi-identifier or the sensitive attribute unlike generalization and suppression, but dissociates the relationship among the two. exactly, the technique release the data on QID and the data on the sensitive attribute in two separate tables: a quasi-identifier table (QIT) contains the QID attributes, a sensitive table (ST) contains the sensitive attributes, and both QIT and ST have one common attribute, Group ID. All records in the same group will have the same value on Grouped in both tables, and therefore are linked to the sensitive values in the group in the exact same way. If a group has distinct sensitive values and each distinct value occurs exactly once in the collection, then the likelihood of linking a documentation to a sensitive value by Group ID is 1. The attribute linkage attack can be distorted by increasing.

EXISTING SYSTEM

Column generalization used in Existing anonymization algorithms, e.g., Mondrian. Existing anonymization algorithms can be applied on the sub table containing only attributes in one column to ensure the anonymity[9] requirement. On sliced data existing data analysis (e.g., query answering) methods can be easily used. Existing privacy measures for membership disclosure protection include presence and differential privacy[10].

PROPOSED SYSTEM

We present a new technique which called slicing, which partition the facts together vertically and horizontally. Another important advantage of slicing is that it can handle high-dimensional data .We proof that slicing preserves best data utility than generalization and can be used for membership disclosure protection. We show how slicing data can be used for attribute disclosure protection and develop an effective algorithm for computing the sliced data that obey the l-diversity must. Our workload experiment confirm that slicing preserves is more effective than bucketization in workloads

involving the sensitive attribute and better utility than generalization.

Collaborative Data Publishing

We proposed new collaborative data publishing till used single data publisher For example, two credit card companies want to integrate their customer data for developing a fraud detection system or for publishing to a bank. However, the credit card companies do not want to indiscriminately disclose their data to each other or to the bank for reasons such as privacy protection and business competitiveness.

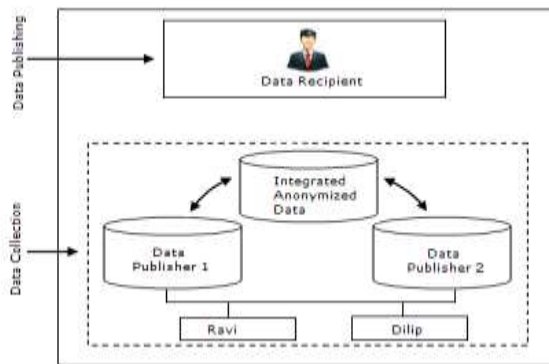


Fig. 2. Collaborative data publisher and data collection.

Figure 2 depicts this scenario, called collaborative data publishing, where several data publishers own different sets of attributes on the same set of records and want to publish the integrated data on all attributes. Say, publisher 1 owns {R ID, Job,Sex, Age}, and publisher 2 owns {R ID,Salary, Disease}, where R ID, such as the SSN, is the documentation identifier common by all facts publishers. They want to publish an integrated k-anonymous table on all attributes.

This scheme ensures the equality of two values encrypted in a different order on the same set of keys, that is, $EncKey1(EncKey2(RID)) = EncKey2(EncKey1(RID))$

PROPOSED ALGORITHM

Proposed algorithm consists of three phases:

- attribute partitioning,
- column generalization,
- and tuples partitioning

Algorithm: Partition_Tup (tuples[], l)

1. Query = {tuples};
2. while Query is not blank
3. remove the first container from
4. Query Divide B container into two containers as in Mondrian.
5. if check diversity (T, Query U {B1,B2} U SB, l)
6. Query = Query U {B1,B2}.
7. else SB = SB U {B}.
8. return SB.

Fig3: Algorithm of Partiton_tuples

RESULT

We conduct extensive workload experiments on Table 3. Table 3 have following attributes are Age, Sex, Zip Code and Disease.

Table 3. original table

Age	Sex	Zipcode	Disease
24	M	452001	Dyspepsia
24	F	452001	Bron
26	M	452051	Dyspepsia
34	f	452091	Flu
44	M	452061	Bron
60	F	452001	Bron
64	M	452061	Flu
66	F	452001	Dyspepsia

Apply slice algorithm on above Table3 and result is shown in Table4. Slice algorithm which partitions the data both horizontally and vertically. We demonstrate that slicing preserves better facts utility than generalization and can be used for membership disclosure protection.

Table4: Slicing Table

(Age,Sex)	(Zipcode,Disease)
(24,M)	(452001,flu)
(26,M)	(452051,dysp)
(34,F)	(452091,flu)
(24,F)	(452001,bron)
(64,M)	(452061,bron)
(66,F)	(452001,dysp)
(60,F)	(452001,bron)
(44,M)	(452061,flu)

we experiments on two different data publisher confirm that slicing preserves is more effective than bucketization in workloads involving the sensitive attribute and good utility than generalization. Our experiments also show that slicing can be used to prevent membership disclosure. We taken the value of $L=3, k=3$ and $C=.5$ for our experiment.

Table5: Anonymous mash up data

Shared		Data provider X		Data provider Y	
UID	Class	Sensitive	Gender	Job	Age
1	Y	S1	M	Non Tech.	[20-50]
2	N	S2	M	Professional	[20-50]
3	Y	S1	M	Non Tech.	[20-50]
4	N	S2	M	Professional	[1-20]
5	N	S2	M	Non Tech.	[20-50]
6	Y	S2	M	Non Tech.	[20-50]
7	N	S2	M	Professional	[1-20]
8	N	S2	F	Professional	[20-50]
9	N	S2	F	Professional	[20-50]
10	Y	S2	F	Tech.	[50-90]
11	Y	S2	F	Tech.	[50-90]

A 3-anonymous table by generalizing QID = {Zipcode, Age, Disease} is show in figure 4. Since each group contains at smallest amount 3 records, the chart is 3-anonymous.

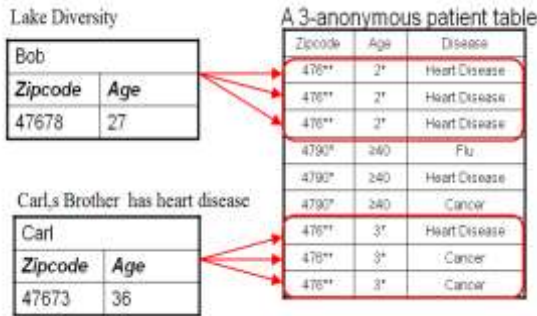


Fig-4 -anonymous patient table

CONCLUSION

We developed an effective algorithm for computing the sliced table that satisfy l-diversity. Our algorithm partition attribute keen on columns, apply column simplification, and partitions tuples into container. Attribute that are highly-related are in the same column. a new data anonymization technique called slicing to improve the current state of the art.

We prove that slicing can be effectively used for preventing attribute discovery, based on the isolation condition of l-diversity Our results confirm that slicing preserves much better facts utility than simplification. In workloads involving the receptive attribute, slicing is also more effective than bucketization. In some categorization experiment, slicing show better performance than using the



original data (which may over fit the model). Our experiments also show the limitations of bucketization in membership disclosure protection and slicing remedies these limitations.

REFERENCES

- [1] M. Atallah, A. Elmagarmid, M. Ibrahim, E. Bertino, and V. Verykios, Disclosure limitation of sensitive rules, in Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange, ser. KDEX 99. Washington, DC, USA: IEEE Computer Society, pp. 45-52 1999
- [2] Mahmoud Hussein, Ashraf El-Sisi, Nabil Ismail Fast Cryptographic Privacy Preserving Association Rules Mining on Distributed Homogenous Data Base, Knowledge-Based Intelligent Information and Engineering Systems, Lecture Notes in Computer Science, Volume 5178/2008, pp. 607-616 2008.
- [3] Agrawal D. Aggarwal C. C. On the Design and Quantification of Privacy-Preserving Data Mining Algorithms.ACM PODS Conference, 2002.
- [4] Machanavajjhala A., Gehrke J., Kifer D., and Venkita subramaniam M.: l-Diversity: Privacy Beyond k-Anonymity. ICDE, 2006.
- [5] Pinkas B.: Cryptographic Techniques for Privacy-Preserving Data Mining.ACM SIGKDD Explorations, 4(2), 2002.
- [6] Blum A., Dwork C., McSherry F., Nissim K.: Practical Privacy: The SuLQ Framework.ACM PODS Conference, 2005. Intelligent Information and Engineering Systems, Lecture Notes in Computer Science, Volume 5178/2008, pp. 607-616 2008.
- [7] Agrawal, D. and C.C. Aggarwal, 2001. On the design and quantification of privacy preserving data mining algorithms. Proceedings of the 20th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, May 21-23, ACM, Santa Barbara, California, USA., pp: 247-255. DOI: 10.1145/375551.375602
- [8] Bhuvana, J. and T. Devi, 2011. Performance of secure multiparty computation for preserving privacy in collaborative data mining. Int. J. Res. Rev. Comput. Sci., 2: 463-469
- [9] R. Agrawal and R. Srikant, "Privacy-preserving datamining," In Proceedings of the ACM SIGMOD Conference on Management of Data, ACM Press, pp. 439-450
- [10]D. Agrawal and C. Aggarwal, "On the design and quantification of privacy preserving data

mining algorithms,” In Proceedings of the 20th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Santa Barbara, CA, pp. 247–255, May 21–23

Author Bibliography

	<p>Shrishti pawar Research Scholar, Gyan Ganga Institute of Technology And Science, Jabalpur, M.P, India.</p>
	<p>Hare Ram Shah Associate Professor, Gyan Ganga Institute of Technology And Science Jabalpur M.P, India. Department of Computer Science & Engg.</p>