A Survey on Novel Privacy Technique to Secure Sensitivity of Micro Data

V. Yamini Priya[1] R. Prabhu[2] R. Krishnakumar[3] M. Brindha Devi[4]
[1,2,3,4]Assistant Professor
[1,2,3,4]Department of Computer Science & Engineering
[1,2,3,4]VSBCETC, Coimbatore, India

Abstract— The collaborative information publishing difficulty for anonymizing horizontally partitioned data at numerous data providers. Data providers can attempt to infer information about data coming from other providers during the anonymization. A distributed databases there is an ever-increasing require for distribution information that hold individual information. The offered scheme presented collaborative data publishing problem for anonymizing horizontally partitioned information at numerous information providers. So m-privacy, which guarantees that the anonymized data satisfies a given privacy constraint against any group of up to m colluding data providers. There will be M number of users so the anonymization algorithm for each individual user cannot be provided individually. To overcome this, a novel technique called overlapping slicing which provides better data utility using l diversity requirement for high dimensional data is used.

Key words: Data Anonymization, Overlapping Slicing, Data Privacy, Security, Integrity

I. INTRODUCTION

A privacy conserving business enterprise of micro data has been studied extensively in recent years. Micro data contains records each of which contains information about an individual entity, such as a person, a household, or an organization. Several micro data anonymization techniques have been proposed. For anonymization the attributes are of three categories [2]- 1) identifiers used to identify an individual. E.g.: username, 2) Quasi Identifiers are public to an individual and when they are linked to the other databases E.g.: gender, birth date 3) Sensitive attributes which must be protected and kept in privacy E.g.: disease, phone number. Privacy preserving data analysis and data publishing has received considerable attention in recent years as promising approaches for sharing data while preserving individual privacy. In a non-interactive representation, a information supplier (e.g., hospital) publishes a "sanitary" version of the statistics, At the same time providing effectiveness for data users (researchers), and privacy security for the individuals represented in the information (patients).

As soon as data are gathered from numerous data providers or data owners, two most important settings are used for anonymization. One approach is for every supplier to anonymize the information autonomously which outcome in potential loss of integrated data effectiveness. An additional accepted approach is collaborative information distributes, which anonymizes information starting all suppliers as if they would come beginning of one source, using either a trusted third-party (TTP) or Secure Multi-party Computation (SMC) protocols. The address the problem of privacy preserving information mining particularly, then consider a situation in which two parties owning secret databases wish to run a data mining algorithm on the combination of their databases, without enlightening any superfluous information. Our work is motivated by the need to both protect privileged information and enable its use for research or other purposes

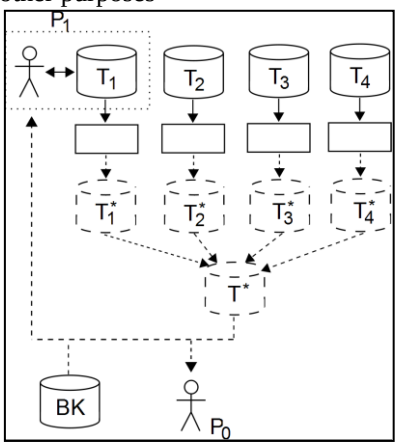

Fig. 1: Anonymize-And-Aggregate

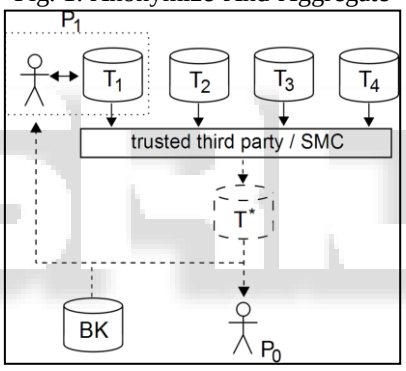

Fig. 2: Distributed Data Publishing Settings for Four Providers

When data are gathered from multiple data providers, it should be anonymized before publishing it. Each provider to anonymize the data independently Fig. 2(a), which results in potential loss of integrated data utility. An additional popular approach is collaborative data publishing which anonymizes data from all providers as if they would come from one source Fig. 2(b), using either a trusted third-party (TTP) or Secure Multi-party Computation (SMC) protocols.

The above problem is a specific illustration of secure multi-party computation and as such, can be solved using known nonspecific protocols. On the other hand, information mining algorithms are usually complex and, furthermore, the input usually consists of massive data sets. The generic protocols in such a case are of no practical use and therefore more efficient protocols are required. This focus on the problem of decision tree learning with the popular ID3 algorithm. Our protocol is considerably more efficient than generic solutions and demands both very few rounds of communication and reasonable bandwidth.

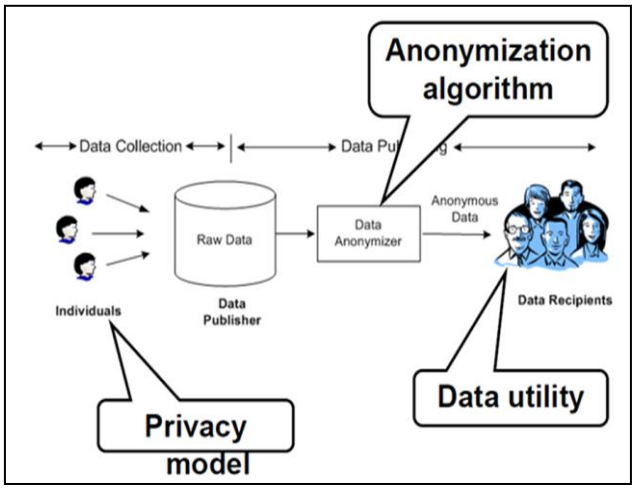

Fig. 3: Privacy Preserving in Data Mining

II. RELATED WORK

An easy privacy preserving reformulation [1] of a linear program whose equal opportunity limitation matrix is partitioned off into teams of rows. Every cluster of matrix rows and its equivalent mitt aspect vector are in hand by a definite non-public unit that's unwilling to contribute to or build community its row cluster or mitt aspect vector. By multiplying every incamera command constraint cluster by Associate in Nursing suitably generated and in camera command accidental matrix, the initial linear program is reworked interested in the same one that doesn't make public any of the in camera command information or build it open. The answer vector of the reworked safe and sound linear plan is in public generated and is obtainable to any or all entities.

Privacy preserving organization and data processing, whereby the info to be confidential or well-mined is in hand by completely different entities that are un eager to reveal the info they hold or build it public, has unfold to the sector of improvement and specifically applied math. In an exceedingly variety of short comings within the privacy preserving applied math journalism are known.

In an exceedingly methodology for handling privately command vertical partitions of a applied math restraint matrix and privacy vector is projected that's supported personal random transformations of the equivalent downside variables. The BIRCH algorithmic program [2] may be a documented algorithmic program for agglomeration for successfully computing clusters in an exceedingly massive information position. Because the information is often disseminated more than many sites, agglomeration over distributed information is a crucial downside. The info may be circulated in horizontal, vertical or every which way partitioned off databases. But, thanks to privacy problems no party might split its information to alternative parties. The matter is however the parties will cluster the distributed information while not breaching privacy of others information. The solutions in every which way partitioned off information location usually work for each flat and vertically partitioned off databases. It provides a method for firmly running BIRCH algorithmic program over every which way partitioned off information. Introduce protected protocols for distance metrics and provides a system for exploitation these metrics in firmly computing clusters in excess of every which way partitioned off any kind of information.

The Privacy conserving [3] data processing has been a well-liked analysis space for quite a decade attributable to its immense spectrum of applications. The aim of privacy conserving data processing researchers is to build up data processing techniques that might be functional on databases while not violating the privacy of people. This work propose ways for constructing the un similarity matrix of objects from completely different sites in an exceedingly privacy conserving manner which might be used for privacy conserving agglomeration similarly as information joins, record linkage and alternative operations that need pair-wise comparison of individual non-public information objects horizontally distributed to multiple sites. ID3 algorithmic program [4] describes, Privacy and security issues will stop sharing of knowledge, and derailing data processing comes. Introduce a generalized privacy conserving variant of the ID3 algorithmic program for vertically partitioned off information distributed over 2 or additional parties.

At the side of the algorithmic program, it provides a complete proof of security that offers a good sure on the knowledge discovered. Whereas this has been in deep trouble horizontally partitioned off information. It gift Associate in Nursing algorithmic program for vertically partitioned off data: a little of every instance is gift at each web site; however no web site contains complete info for any instance. This downside has been self-addressed, however the answer is proscribed to the case wherever each party have the category attribute.

Data providers can attempt to infer information about data coming from other providers during the anonymization. In this Collaborative data publishing can be considered as a multi-party computation problem, in which multiple providers wish to compute an anonymized view of their data without disclosing any private and sensitive information. There will be M number of users so the anonymization algorithm for each individual user cannot be provided individually.

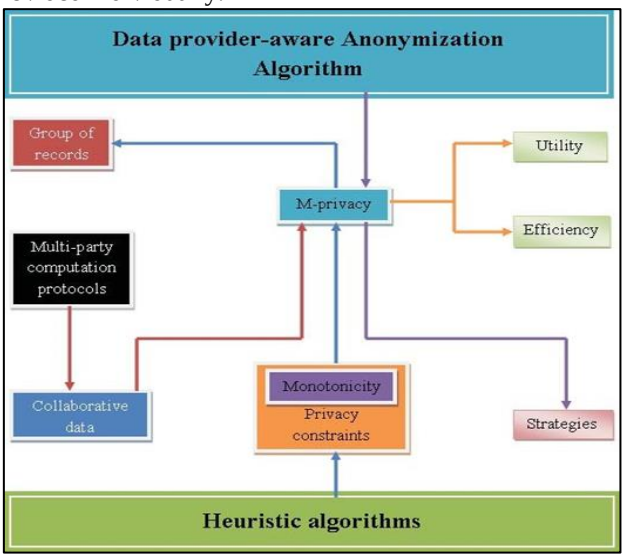

Fig. 4: Architecture Diagram of m-Partition Privacy Scheme to Anonymizing Set-Valued Data

The phases involved in this scheme are

– Anonymization for M-Privacy

- K-Anonymity in Set Valued Data
- Partition based Anonymization

A. Anonymization for M-Privacy

The baseline algorithmic utilizes information an information provider-aware algorithmic program with adaptation verification methods to confirm high utility and m-privacy for anonymized records. The SMC implements the m privacy anonymization during a distributed surroundings whereas protective security. For a privacy constraint C that's generalization monotonic m-privacy with relevancy is generalization monotonic. Most existing generalization-based anonymization algorithms square measure changed to ensure m-privacy with relevancy.

The Adoption is easy whenever a collection of report is experienced for confidentiality fulfillment check m-privacy with relevancy. C. the Binary house Partitioning (BSP) recursively chooses Associate in nursing characteristic to separate information points in flat area house till information can't be split any more while not breaching m-privacy with relevancy. The options of BSP takes under consideration the info supplier as an extra measurement for rending uses privacy strength score as a general marking metric for choosing the split purpose. It adapts its m-privacy checking strategy for economical confirmation.

B. K-Anonymity in Set Valued Data

The K- Anonymity is set valued data privacy model consider Let I = {I1, I2... I|I|} be the set of items from which elements of the sets are drawn and Let D = {t1, t2... t|D|} be a transactional database over I where each transaction ti within D is a non-empty subset of I. The equality class in transactional database D consists of a multi set of transactions. An equivalence class for D is the set of all transactions with identical sets of items S.

The k anonymity in set esteemed information transactional record D is k-anonymous if every transaction in D occurs at slightest k times, or equivalently the size of each equivalence class in D is at least k. The Transactional database is k- anonymous if each transaction is identical to at least k − 1 others. The states that given any m or fewer items chosen from any transaction there are at least k−1 other transactions containing same set of m items.

The km anonymity simply protects individuals' privacy as soon as adversary knows m or smaller amount items whereas k anonymity, with the nonexistence of parameter m, requires no boundary on numeral of objects adversary can be familiar with smaller the m in km-anonymity and weaker privacy km-anonymity provides When m = Mmax. Mmax is the maximum length of transaction.

C. Partition based Anonymization

The Partition based method of anonymization recursively straightening out set valued data hooked on clusters where data in each separation split a generalized demonstration. In Mondrian anonymization algorithm generalization hierarchy has to be used in deciding which transactions are similar be grouped together.

The partition based method of anonymization algorithm establish by generalizing every transactions to starting place in the hierarchy. The initial point at all times produces a trivial anonymization by means of one partition, as long as there are at least k transactions in the database. All transactions share same representation ("ALL") after being generalized to the root The Pass the initial partition to the anonymize routine which splits the current partition into sub-partitions recursively invokes anonymize on all resulting sub partitions. The partitioning process terminates when no further split is possible.

III. PROPOSED SYSTEM

In this proposed an overlapping slicing method for handling high into more than one column is used. It provides better data utility using l diversity [1] requirement for high dimensional data. It use an efficient algorithm called chi matrix for attribute correlation, to protect privacy by breaking the association of uncorrelated attributes and preserve data utility by preserving the association between highly correlated attributes. In this technique discharge no correlations among attributes there by, are overlapping slicing preserves improved information effectiveness in workloads involving the responsive attributes.

A. Algorithm

Algorithm of "Overlapping Slicing" is presented below:

1) Load Dataset;
2) Attribute Partition and Column
3) Process Tuple Partition and Buckets
4) Slicing
5) Undergo Column Generalization
6) Do Matching Buckets
7) End;

1) Load Dataset

The Fig5 is original table is the micro data which is to be anonymized before publishing it.

Age	Sex	Zipcode	Disease
22	M	47906	Dyspepsia
22	F	47906	Flu
33	F	47905	Flu
52	F	47905	Bronchitis
54	M	47302	Flu
60	M	47302	Dyspepsia
60	M	47304	Dyspepsia
64	F	47304	Gastritis

Fig. 5: Load Dataset Table

2) Attribute Partition & Column

A tuple partition consists of several subsets of T , such that each tuple belongs to exactly one subset of tuples called a bucket. Specifically, b buckets B1, B2,.. .Bn.

Age	Sex	Zipcode	Disease
[20-52]	*	4790*	dyspepsia
[20-52]	*	4790*	flu
[20-52]	*	4790*	flu
[20-52]	*	4790*	bronchitis
[54-64]	*	4730*	flu
[54-64]	*	4730*	dyspepsia
[54-64]	*	4730*	dyspepsia
[54-64]	*	4730*	gastritis

Fig.6: Attribute partition Table

3) Tuple Partition and Buckets

Microdata table T and a column Ci= {fAi1; Ai2; . . .;Aij} where Ai1;Ai2; . . .;Aij are attributes,

Age	Sex	Zipcode	Disease
22:2,33:1,52:1	M:1,F:3	47905:2,47906:2	dysp.
22:2,33:1,52:1	M:1,F:3	47905:2,47906:2	flu
22:2,33:1,52:1	M:1,F:3	47905:2,47906:2	flu
22:2,33:1,52:1	M:1,F:3	47905:2,47906:2	bron.
54:1,60:2,64:1	M:3,F:1	47302:2,47304:2	flu
54:1,60:2,64:1	M:3,F:1	47302:2,47304:2	dysp.
54:1,60:2,64:1	M:3,F:1	47302:2,47304:2	dysp.
54:1,60:2,64:1	M:3,F:1	47302:2,47304:2	gast.

Fig.7: Tuple Partition Table

4) Overlapping Slicing

Overlapping slicing, which duplicates an attribute in more than one column and releases more attribute correlations in the micro data.

(Age,Sex,disease)	(Zipcode,Disease)
(22,M,flu)	(47906,flu)
(22,F,dysp.)	(47906,dysp.)
(33,F,bron.)	(47905,bron.)
(52,F,flu)	(47906,flu)
(54,M,gast.)	(47304,gast.)
(60,M,flu)	(47302,flu)
(60,M,dysp.)	(47302,dysp.)
(64,F,dysp.)	(47304,dysp.)

Fig. 8: Overlapping Slicing Table

IV. FUTURE WORK & CONCLUSION

Thus from implementation it is prove that Overlapping Slicing overcomes the limitations of existing techniques of m privacy and providing better utility while protecting against privacy threats. Overlapping slicing to prevent attribute disclosure and membership disclosure. . The common method proposed by this work is that previous to anonymizing the data, one be capable of examine the data characteristics and make use of these characteristics in data anonymization. As future work is plan to design more effective tuple grouping algorithms. The trade-off between column generalization and tuple partitioning is the subject of future work. The blueprint of tuple consortium algorithms is absent to opportunity employment.

REFERENCES

[1] Olvi L. Mangasarian, "Privacy-Preserving Horizontally Partitioned Linear Programs" 2003.
[2] P. Krishna Prasad and C. Pandu Rangan "Privacy Preserving BIRCH Algorithm for Clustering over Arbitrarily Partitioned Databases".
[3] Ali Inan, Yücel Saygın, Erkay Sava, Ayça Azgın Hinto lu, Albert Levi "Privacy Preserving Clustering on Horizontally Partitioned Data".
[4] Jaideep Vaidya and Chris Clifton "Privacy-Preserving Decision Trees over Vertically Partitioned Data".
[5] Zhiqiang Yang and Rebecca N. Wright, "Privacy-Preserving Computation of Bayesian Networks on Vertically Partitioned Data" IEEE Transactions On Knowledge And Data Engineering, VOL. 18, NO. 9, SEPTEMBER 2006.
[6] Khuong Vu and Rong Zheng Jie Gao "Efficient Algorithms for K-Anonymous Location Privacy in Participatory Sensing" 2012 Proceedings IEEE INFOCOM.
[7] Sebastian Schrittwieser, Peter Kieseberg, Isao Echizen, Sven Wohlgemuth, Noboru Sonehara, and Edgar Weippl "An Algorithm for k- anonymity-based Fingerprinting".
[8] 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.
[9] W. Jiang and C. Clifton, "A secure distributed framework for achieving k-anonymity," VLDB J., vol. 15, no. 4, pp. 316–333, 2006.
[10] L. Sweeney, "k-Anonymity: a model for protecting privacy," Int. J. Uncertain. Fuzziness Knowl.-Based Syst., vol. 10, no. 5, pp. 557–570, 2002.
[11] C. Aggarwal, "On k-Anonymity and the Curse of Dimensionality," Proc. Int"l Conf. Very Large Data Bases (VLDB), pp. 901-909, 2005.
[12] Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical Privacy: The SULQ Framework," Proc. ACM Symp. Principles of Database Systems (PODS), pp. 128-138, 2005.
[13] B.-C. Chen, K. LeFevre, and R. Ramakrishnan, "Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge," Proc. Int"l Conf. Very Large Data Bases (VLDB), pp. 770-781, 2007.
[14] Dinur and K. Nissim, "Revealing Information while Preserving Privacy," Proc. ACM Symp. Principles of Database Systems (PODS), pp. 202-210, 2003.
[15] C. Dwork, "Differential Privacy: A Survey of Results," Proc. Fifth Int"l Conf. Theory and Applications of Models of Computation (TAMC), pp. 1-19, 2008.
[16] J.H. Friedman, J.L. Bentley, and R.A. Finkel, "An Algorithm forFinding Best Matches in Logarithmic Expected Time," ACM Trans. Math. Software, vol. 3, no. 3, pp. 209-226, 1977.
[17] D. Kifer and J. Gehrke, "Injecting Utility into Anonymized Data Sets," Proc. ACM SIGMOD Int"l Conf. Management of Data (SIGMOD), pp. 217-228, 2006.
[18] N. Koudas, D. Srivastava, T. Yu, and Q. Zhang, "Aggregate Query Answering on Anonymized Tables," Proc. IEEE 23rd Int"l Conf. Data Eng. (ICDE), pp. 116-125, 2007.
[19] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional k-Anonymity," Proc. Int"l Conf. Data Eng. (ICDE), p. 25, 2006.
[20] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and „-Diversity," Proc. IEEE 23rd Int"l Conf. Data Eng. (ICDE), pp. 106-115, 2007.

