An Efficient Collaborative Data Publishing by M-Privacy Technique

Gokulavani.M
Department of CSE,
PPG Institute of Technology, Coimbatore,
Tamilnadu, India.

Santhamani.V
Department of CSE,
PPG Institute of Technology, Coimbatore,
Tamilnadu, India.

Gayathri.A
Department of CSE,
PPG Institute of Technology,
Coimbatore,
Tamilnadu, India.

S.R.Ramya
Department of CSE,
PPG Institute of Technology,
Coimbatore,
Tamilnadu, India.

ABSTRACT
For privacy preserving data publishing many anonymization techniques such as generalization and bucketization have been designed. Next, a novel technique is presented, called slicing to have a clear separation between quasi-identifying attributes. It partitions the data both horizontally and vertically and can be used to prevent membership disclosure protection. For anonymizing horizontally partitioned data at multiple data providers, the collaborative data publishing problem is considered. A new type of "insider attack" is being introduced by colluding data providers who may use their own data records with the external background knowledge to infer the data records contributed by other data providers. m-privacy is introduced which guarantees that the anonymized data satisfies a given privacy constraint against any group of up to m colluding data providers. A data provider-aware anonymization algorithm is presented with m-privacy checking strategies to ensure high utility and efficiency. This approach achieves better utility and efficiency in real-life datasets.

Keywords
m-privacy, quasi-identifying attributes, m-adversary, t-closeness, l-diversity, k-anonymity.

I. INTRODUCTION
Data mining, otherwise known as knowledge discovery, helps to extract knowledge from database. There are various possibilities of databases getting attacked by an external recipient by using the background knowledge about the users. Many privacy issue occurs in data mining as all the information are stored in a single database and hence there is a high risk of privacy issue. Recently distributed database is being followed as it reduces the time of extracting knowledge from database and later on the collaborative data publishing was introduced.

A main problem that arises in mass collection of data is confidentiality. Privacy has become very essential due to law (e.g., for medical databases). There is an increasing need for sharing data that contain personal information from distributed databases. For example, in social networking sites, government databases, etc.,Privacy preserving data analysis and data publishing have received considerable attention in recent year. When the data are distributed among multiple data providers or data owners, for anonymization two main approaches are used, one approach is to anonymize the data independently for each provider. Another approach is collaborative data publishing.

Problem Settings:
The problem is that data recipient including the data providers will not be able to compromise the privacy of the individual records provided by other parties. Considering different types of malicious users and information they can use in attacks, three main categories of attack are identified.

External Data Recipient Attacks. A data recipient be an attacker and attempts to infer additional information about the records using the published data and some background knowledge. Many literature provides a notion to protect against specific types of attacks by assuming limited background knowledge. For example, k-anonymity, l-diversity and t-closeness.

Data Providers Attack using Intermediate Results. The data providers can attempt to infer additional information about data coming from other providers by analyzing the data received during the anonymization. A trusted third party (TTP) or Secure Multi-Party Computation (SMC) protocols can be used to guarantee there is no disclosure of intermediate information during the anonymization. However, either TTP or SMC do not protect against data providers to infer additional information about other records using the anonymized data and their own data.

Data Providers Attack Using Anonymized Data and Their Own Data. Each provider has additional data knowledge of their own records, compared to the attack in the first scenario. When multiple data providers collude with each other it gets worse.
A new type of "insider attack" by data providers is addressed. An m-adversary of m colluding data providers or data owners is defined, who have access to their own data records as well as publicly available background knowledge and attempts to infer data records contributed by other data providers. Since each provider holds a subset of the overall data, this inherent data knowledge has to be explicitly modelled and checked.

RELATED WORK
In recent years privacy preserving data analysis and publishing has received great attention. Previously it was focused on a single data provider and considered the external data recipient as an attacker. Many techniques such as k-anonymity, l-diversity and t-closeness assumes limited background knowledge of the attacker. Each data holder knows its own records in case of distributed setting. Many different anonymization algorithms have been introduced so far

including Datafly, Incognito and Mondrian. For high efficiency and extensibility, the Mondrian algorithm is considered as a baseline. For vertical partitioning the data, k-anonymity is used, here the data is collected from individual data owners. Next, l-diversity to ensure the anonymity for data providers. Secure Multi-party Computation techniques for anonymizing distributed data. In the collaborative data publishing setting the data providers are considered as potential attackers and knowledge of the data providers as well as collusion between them for any weak privacy.

II. m-PRIVACY DEFINITION

In order to prevent attacks by m-adversary, m-privacy along with a given privacy constraint was introduced. A set of records T which is horizontally distributed among n data providers P is considered here. Our goal is to publish an anonymized table T.

III PROCESS FLOW

The data providers store their database separately and then are collaboratively distributed among the data providers. Previously, the external recipients where the attackers who would attack the database to obtain information about the data, but now the data providers themselves act as an attacker with the external background knowledge. In order to avoid this certain constraints for the data providers are provided.

For example, if four data providers publish their data collaboratively only they can view their respective data. If the other data providers need to view the details of another data provider they have to satisfy the constraints provided by the respective data provider.

A. m-Privacy

A given privacy requirement C, is assumed to protect data from external recipients with some background knowledge. C(T) is said to be true, if table T satisfies C. Any of the existing privacy principles can be used as a component constraint. In table T, the privacy constraint C is defined as C = C1 ^ C2, where C1 is k-anonymity and C2 is l-diversity.

To protect the anonymized data against m-adversaries in addition to the external data recipients, a notion m-privacy is defined with respect to privacy constraint C.

Definition 2.1: (m-PRIVACY) A set of records T_I is contributed by n data data providers, a set of records T, an anonymization mechanism A, an m-adversary is a coalition of m providers. Sanitized records T*=A(T) satisfy m-privacy, i.e. are m-private, with respect to a privacy constraint C.

m-Privacy and Weak Privacy Constraints. A set of records of Table T satisfying C will only guarantee 0-privacy with respect to C, i.e, C is not guaranteed to hold for each equivalence group after excluding records belonging to any malicious data provider when a weak privacy constraint C does not consider instance level background knowledge, such as k-anonymity, l-diversity and t-closeness. Hence, breaching of records provided by others can be performed by each data provider.

Strengths and weaknesses of C will be inherited as m-privacy is defined with respect to a privacy constraint C. For example, if C is defined by k-anonymity, then ensuring m-privacy with respect to C will not protect against homogeneity attack or deFinetti attack. If C protects against the privacy attack by any external data recipient then m-privacy with respect to C will protect against the same privacy attack issued by any m-adversary. m-Privacy constraint is orthogonal to the privacy constraint C being used.

m-Privacy and Differential Privacy.

Privacy is guaranteed even if an attacker knows all records except the victim record and differential privacy, does not assume specific background knowledge. Thus, any records satisfying differential privacy also satisfies (n-1) privacy, i.e. maximum level of m-privacy, when any (n-1) providers can collude. m-privacy offers a practical trade off between preventing m-adversary attacks with bounded power m and the ability to publish generalized but truthful data records while m-privacy with respect to any weak privacy notion does not guarantee unconditional privacy.

B. Monotonicity of Privacy Constraints

For privacy constraints generalization based monotonicity has been defined.

Definition 2.2: (GENERALIZATION MONOTONICITY OF A PRIVACY CONSTRAINT)
If for any set of anonymized record T satisfying C, all its further generalizations satisfy C as well, then a privacy constraint C is generalization monotonic.

Generalization monotonicity makes an assumption about the original records T, that they have been already anonymized and uses them for further generalizations. Record-based definition of monotonicity is introduced to facilitate the analysis and design of efficient algorithms for checking m-privacy.

Definition: (EQUIVALENCE GROUP MONOTONICITY OF A PRIVACY CONSTRAINT)
If any set of anonymized records T satisfies C then a privacy constraint C is said to be EG monotonic.

EG monotonicity is more restrictive than generalization monotonicity. If a constraint is generalization monotonic, it need not be EG monotonic, but vice-versa exist. Examples of EG monotonic constraints are k-anonymity and l-diversity that requires l distinct values of sensitive attribute in an equivalence group, which are also generalization monotonic. Examples of generalization monotonic constraints that are not EG monotonic at the same time are entropy l-diversity and t-closeness. Entropy l-diversity will not be EG monotonic if a record is added that will change the distribution of sensitive values significantly.

III. VERIFICATION OF m-PRIVACY

To check a set of records that satisfies m-privacy creates a potential computational challenge due to the combinatorial number of m-adversaries that need to be checked. Here the problem is analyzed by modelling the checking space. For efficiently checking m-privacy for a set of records with respect to an EG monotonic privacy constraint C, heuristic algorithms is presented with effective pruning strategies and adaptive ordering techniques.

Heuristic Algorithms

The key idea is to efficiently search through the adversary space with effective pruning such that not all m-adversaries need to be checked. By two different pruning strategies, an adversary ordering technique and a set of search strategies that enable fast pruning above is achieved.

Pruning Strategies.

Downward Pruning: If a coalition is not able to breach privacy, then the sub coalitions will not be able to breach the privacy

and hence they need not to be checked. Upward Pruning: If a coalition is able to breach privacy, then the super-coalitions will be able to do breach the privacy and hence do not need to be checked. The upward pruning allows the algorithm to terminate immediately as the m-adversary will be able to breach privacy, if a sub-coalition of an m-adversary is able to breach privacy.

Adaptive Ordering of Adversaries.
The coalitions are adaptively ordered based on their attack powers to facilitate the above pruning in both directions. Super-coalitions of m-adversaries with limited attack powers are preferred to check first in case of downward pruning. Downward pruning chance is increased when pruning strategies for m-privacy check breach privacy. Upward pruning chance is increased when sub-coalitions of m-adversaries with significant attack powers are preferred to check first as they are more likely to breach privacy.

Definition: (PRIVACY FITNESS SCORE) Privacy fitness C for a set of records T is a level of the fulfilment of the privacy constraint C. C(T) is true when a privacy fitness score is a function f of privacy fitness with values greater or equal to 1.

The privacy fitness score of the records jointly contributed by its members is used to measure the attack power of a coalition. Breaching the privacy for the remaining records in a group after removing their own records are done when the privacy fitness score is higher.

The super-coalitions of m-adversaries and the sub-coalitions of m-adversaries are generated in the order of ascending fitness scores (ascending attack powers) and descending fitness scores (descending attack powers) to maximize the benefit of both pruning strategies.

To enable fast pruning all heuristic algorithms use adaptive ordering of adversaries.

The Top-Down Algorithm. The top-down algorithm uses downward pruning, starting from (n_G - 1)-adversaries and moves down until a violation by an m-adversary is detected or all m-adversaries are pruned or checked.

The Bottom-Up Algorithm. The bottom-up algorithm uses upward pruning, starting from 0-adversary and it moves up until a violation by any adversary is detected (early-stop) or all m-adversaries are checked.

The Binary Algorithm. The binary algorithm, checks coalitions between (n_G - 1)-adversaries and m-adversaries and takes advantage of both upward and downward pruning's. The goal of each iteration is to search for a pair of Isub and Isuper, so that Isub is a direct sub-coalition of Isuper and Isuper that breaches privacy while Isub does not breach privacy. Then Isub and all its sub-coalitions are pruned, Isuper and all its super-coalitions are pruned as well.

Adaptive Selection of Algorithms.
The characteristics of a given group of providers is decides the algorithm that is to be used. To select the most suitable verification algorithm the privacy fitness score, which quantifies the level of privacy fulfilment of records, may be used. m-privacy will be satisfied when the fitness score of attacked records is higher. Hence a top-down algorithm with downward pruning will significantly reduce the number of adversary checks.

Time Complexity
Here, the time complexity is considered for the m-privacy verification algorithms. An assumption is made such that, each check of C takes a constant time since the algorithms involve multiple checks of privacy constraint C used to define m-privacy for various combinations of records. To adapt m-privacy verification strategy to domain settings is difficult to achieve, on average, a low runtime.

IV. ANONYMIZATION FOR m-PRIVACY

Now the m-privacy verification is used in anonymization of a horizontally distributed dataset, so that m-privacy is acheived. A baseline algorithm and provider-aware algorithm with adaptive m-privacy checking strategies are used to ensure high utility and m-privacy for anonymized data.

Most existing generalization-based anonymization algorithms can be modified to achieve m-privacy every time a set of records is tested for a privacy constraint C, as m-privacy with respect to a generalization monotonic constraint is generalization monotonic, m-privacy is checked with respect to C. The multidimensional Mondrian algorithm designed for k-anonymity is adapted for a baseline algorithm to achieve m-privacy. A main limitation is that groups of records may result in over-generalization in order to satisfy m-privacy.

To overcome this, a simple and general algorithm is introduced based on the Binary Space Partitioning (BSP). In the multidimensional domain space it recursively chooses an attribute to split data points until the data cannot be split further while satisfying m-privacy with respect to C. The algorithm has three novel features, first feature is that it takes the data provider as an additional dimension for splitting, the second feature is that in order to select the split point, it uses the privacy fitness score for general scoring metric, the third feature is that it adapts m-privacy verification strategy for efficient verification.

Provider-Aware Partitioning.
For Quasi Identifying(QI) attributes and data providers the algorithm first generates all possible splitting points. The data provider or data source of each record is considered as an additional attribute of each record(A0) in case of multidimensional QI domain space. For instance, each data record T contributed by data provider P1 will have t[A0] = P1. By decreasing the number of providers in each partition using A0 and hence it increases the chance that more sub-partitions will be m-private and it will be feasible for further splits. This leads to more splits resulting a more precise view of the data.

A total order on the providers can be imposed, by sorting the providers alphabetically or based on the number of records they provide, then the potential split point is found and partition the records into two approximately equal groups by using the splitting point.

Adaptive m-privacy verification.
Records that satisfy m-privacy are added to a candidate set when m-privacy is verified for all possible splitting points. Our algorithm adaptively selects an m-privacy verification strategy using the fitness score of the partitions to minimize the time. The partitions are large and likely m-private in the early stage of the anonymization algorithm. For fast verification a top-down algorithm may be used. As the algorithm continues, the partitions become smaller, the downward pruning will be less efficient.

To allow upward pruning binary algorithm or others may be used instead.

Splitting Point Selection Based on Privacy Fitness Score.
For a non-empty candidate set, the privacy fitness score is used and choose the best splitting point. More splitting takes place if the resulting partitions have higher fitness scores and if they satisfy m-privacy with respect to the privacy constraint.

V. CONCLUSIONS

In this paper, a new type of potential attackers in collaborative data publishing – a coalition of data providers, is considered, called m-adversary. Guaranteeing m-privacy is enough, to prevent privacy disclosure by any m-adversary.

For efficiently checking m-privacy the heuristic algorithms is presented for exploiting equivalence group monotonicity of privacy constraints and adaptive ordering techniques. To ensure high utility and m-privacy of anonymized data a provider-aware anonymization algorithm is introduced with adaptive m-privacy checking strategies. While ensuring m-privacy efficiently our approach achieves better or comparable utility than existing algorithms. Many research questions remain such as a proper privacy fitness score for different privacy constraints, to model the data providers when data are distributed in a vertical fashion.

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] N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. Lee, "Centralized and 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.

[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] W. Jiang and C. Clifton, "A secure distributed framework for achieving k-anonymity," VLDB J., vol. 15, no. 4, pp. 316–333, 2006.

[7] O. Goldreich, Foundations of Cryptography: Volume 2, Basic Applications. Cambridge University Press, 2004.

[8] Y. Lindell and B. Pinkas, "Secure multiparty computation for privacypreserving data mining," The Journal of Privacy and Confidentiality, vol. 1, no. 1, pp. 59–98, 2009.

[9] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in ICDE, 2006, p. 24.

[10] P. Samarati, "Protecting respondents' identities in microdata release," IEEE T. Knowl. Data En., vol. 13, no. 6, pp. 1010–1027, 2001.

[11] L. Sweeney, "k-anonymity: a model for protecting privacy," Int. J. Uncertain. Fuzz., vol. 10, no. 5, pp. 557–570, 2002.

[12] N. Li and T. Li, "t-closeness: Privacy beyond k-anonymity and ldiversity," in In Proc. of IEEE 23rd Intl. Conf. on Data Engineering (ICDE), 2007.

[13] R. Burke, B. Mobasher, R. Zabicki, and R. Bhaumik, "Identifying attack
models for secure recommendation," in In Beyond Personalization: A Workshop on the Next Generation of Recommender Systems, 2005.

[14] D. Kifer, "Attacks on privacy and definetti's theorem," in Proc. of the 35th SIGMOD Intl. Conf. on Management of Data, 2009, pp. 127–138.

[15] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in Proc. of the 2011 Intl. Conf. on Management of Data, 2011, pp. 193–204.

[16] K. Lefevre, D. J. Dewitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in ICDE, 2006.

[17] S. Goryczka, L. Xiong, and B. C. M. Fung, "m-privacy for collaborative
data publishing," Emory University, Tech. Rep., 2011.

[18] X. Xiao and Y. Tao, "Anatomy: simple and effective privacy preservation," in Proc. of the 32nd Intl. Conf. on Very Large Data Bases, 2006, pp. 139–150.

[19] G. Cormode, D. Srivastava, N. Li, and T. Li, "Minimizing minimality and maximizing utility: analyzing method-based attacks on anonymized data," Proc. VLDB Endow., vol. 3, Sept. 2010.

[20] Y. Tao, X. Xiao, J. Li, and D. Zhang, "On anti-corruption privacy preserving publication," in Proc. of the 2008 IEEE 24th Intl. Conf. On Data Engineering, 2008, pp. 725–734.

[21] 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.

[22] 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.

[23] 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.

[24] S. Zhong, Z. Yang, and R. N. Wright, "Privacy-enhancing kanonymization of customer data," in Proc. of the 24th ACM SIGMODSIGACT- SIGART Symposium on Principles of Database Systems, 2005, pp. 139–147.