

A NOVEL APPROACH FOR PUBLISHING DATA WITH PRIVACY

**Maramreddy Nagendra Babu¹, K Chandra Sekhar Reddy²,
Prof.S.V.Achutha Rao³**

¹M.Tech Scholar (CSE), ²Asst. Professor, (CSE), ³Professor & Head, (CSE)

Vikas College of Engineering and Technology, Nunna, Vijayawada, A.P., (India)

ABSTRACT

In this scenario we consider as the collaborative data publishing problem for Anonymizing horizontal partitioned data at multiple data providers. Here we consider the new type "insider attack" by colluding data provider who may use their own data records a subset of the overall data in addition to the external background knowledge to the infer the data record contributed by other data providers. In this addresses this new threat and makes the several contributions. Firstly, we introduced a notion of m-privacy, constraint against any group of up to m colluding data providers and secondly we present heuristic algorithms exploiting an equivalence group monotonicity of privacy constraints and adaptive ordering technique for an efficiently checking m-Privacy given the sets of record. Finally, in that we present the data provider checking strategies to ensure high utility and m-privacy of anonymized the data with the efficiency. The experiments on real-life dataset suggests that our approach achieves better or comparable utility and efficiency than existing and the baseline algorithms while providing m-privacy guarantee.

I. INTRODUCTION

There is the increasing need for the sharing data that contain personal information from the distributed databases. For ex in the healthcare domain, the national agenda is to develop a Nationwide Health hospital Network (NHIN) to share information among hospitals and the others provider and support appropriate use of the health information beyond direct patient care with the privacy protection. The Privacy preserving data analysis and the data publishing, have receive considerable attention in recent years as promising approaches for sharing data distributed among multiple data providers or data owners, two main setting are used for anonymization. One approach is for every provider to the anonymized the data independently which results in potential loss of integrated data utility. The more desirable approach is collaborative data publishing as if they would come from one source protocol to do computation. Problem setting: we consider the collaborative data publishing setting with the horizontal partitioned data across multiple data providers, each contributing a subset of records T_i , as a special case, a data provider could be the data owner itself who is contributing it is self-records.

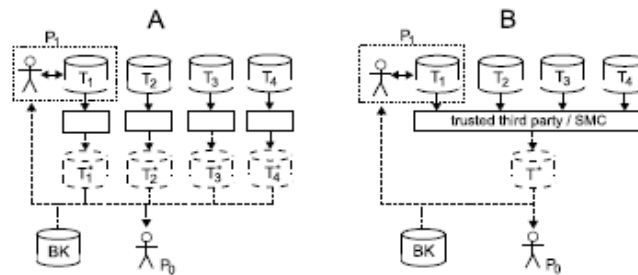


Fig. 1. Distributed data publishing settings.

This is the very Common scenario in the social networking and recommendation systems. Our goal is to publish a 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 information they can use in attack, we identify three main categories of attack scenarios. While in that two addressed in existing work next will receive the small attention and will be focus of the next paper. Attacks by external data recipient using anonymized data: a data recipient e.g. P_0 could be an attacker and attempts to infer additional information about the records using the published data (T^*) and some background knowledge such as the publicly available an external data. More literature on the privacy preserving data publishing in the single provider setting considers only such attacks. Many of them adopt a weak or consider only such attacks. Many of them adopt a weak or relaxed adversarial or Bayes-optimal by assuming prevent identity disclosure attacks include.

II. PRIVACY DEFINATION

We firstly formally describe our problem setting. Later we present our m-privacy definition with to a given privacy constraint to present inference attacks by m-adversary followed by it is properties.

Let $T = \{t_1, t_2, \dots\}$ be a set of records horizontally distributed among n data providers $P = \{p_1, p_2, \dots, p_n\}$, such that $T_i \subseteq T$ is the set of records the provided by the p_i , we assume A_s is the sensitive attribute with the domain D_s . whether the records contain multiple the sensitive attributed. Our goal is to publish an anonymized table T^* while preventing any m-adversary from inferring A_s for any single record.

M-Privacy

To protect data from external recipients with the certain back ground knowledge BK, we assume the given privacy requirement of C, defined by the conjunction of the privacy constraints $C = C_1 \wedge C_2$, where C_1 is k-anonymity with $k=3$ and C_2 is l-diversity with $l=2$ both anonymity tables T_a^* may be compromised by an m-adversary such as P_1

We now formally define a notion of m-privacy with respect to a privacy constraint C, to protect the anonymized data against m-adversaries in addition to the external data recent. The notion explicitly models the inherent data knowledge of an m-adversary, a data records they are the jointly contribute, and requires owned by an m-adversary still satisfies C.

Definition m-PRIVACY given n the data providers, the sets of record T , and an anonymization mechanism A , an m-adversary I ($m \leq n - 1$) is a coalition of m the providers, which jointly contributes the set of records T_I . Sanitized records $T^* = A(T)$ the satisfy m-privacy, i.e. are m-private, with the respect to the privacy constraint of C, only if, $\forall I \subseteq P, |I| = m, \forall T' \subseteq T : T' \not\subseteq T_I, C(A(T')) = \text{true}$

For all $m \leq n-1$ if T^* is m -private then it is also $(m-1)$ -private if T^* is not m -private, then it is also not $(m+1)$ -private

III. VERIFICATION OF M-PRIVACY

The checking whether a set of the records satisfies the m -privacy creates the potential computational challenge due to a combinatorial number of the m -adversaries that the need to be checked. In this scenario, we firstly analyzed the problem by modeling a checking space. After we present heuristic algorithms with the effective pruning strategies and adaptive ordering techniques for the efficiently checking m -privacy for the set of records w.r.t. the EG monotonic privacy constraint in C .

The Adversary Space Enumeration: Given the set of nG data provider, an entire space of the adversaries (m varying from 0 to $nG-1$) can be the represented using the lattice. Every node at layer m represents the m -adversary of the particular combination of the m providers. A number of all possible m -adversaries is equal to nGm . Every node have parent representing their pruning strategies are possible thanks to the EG monotonicity of the m -privacy. If a coalition is unable breach the privacy, then all its sub coalitions will not unable to do and hence do unable be checked downward pruning. On the other way, if the coalition is able to breach privacy, then all it is the super coalitions will be able to do so and hence do not wish to be checked upward pruning. In adaptive Ordering of Adversaries In order to facilitate a above pruning in the both directions, we adaptively order a coalitions based on their attack power.

IV. ANONYNIZATION FOR M-PRIVACY

After defining the m -privacy verification algorithm, we can now use it in the anonymization of the horizontally distributed dataset to achieve a m -privacy. In that we will discuss the baseline algorithm, and after our approach that utilizes the data provider aware algorithm with the adaptive m -privacy checking strategies to ensure the high utility and m -privacy for the anonymized data. After all we have shown that the m -privacy with respect the generalization monotonic constraint is generalization monotonic most existing generalization based anonymization algorithms can be modified to achieve the m -privacy every time a set of records is tested for a privacy constraint C , we check m -privacy w.r.t C instead. As a baseline algorithm to achieve the m -privacy, we adapted a multidimensional Mondrian algorithm designed for the k -anonymity.

However the algorithm has three main innovative features as follow:-

- 1) It takes into the account the data provider as the additional dimension for the splitting.
- 2) It uses the privacy fitness score as the general scoring metric for selecting the split point;
- 3) It adapts it is an m -privacy verification strategy for the efficient verification. Pseudo code for the provider aware anonymization the algorithms are presented in an algorithm.

The provider-aware algorithm

Data: A set of records $T = \bigcup_{j=1}^n T_j$ provided by $\{p_1, p_2, \dots, p_n\}$, a set of QI attribute

$A_i (i = 1, \dots, q), m, a \text{ privacy constraint } C$

Result: Anonymized T^* that satisfies m -privacy w.r.t C . Begin

1. $\Pi = \text{get_splitting_point_for_attributes}(A_i)$
2. $\Pi = \Pi \cup \text{get_splitting_point_for_providers}(A_0)$

3. $\Pi' = \{a_i \in \Pi, i \in \{0, 1, 000, q\}\}$:
4. $\text{are_both_split_subpartitions_m_private}(T, a_i)$
5. if Π' is \emptyset then $T^* \cup \text{generalize_all_QIs}(T)$
6. return T^*
7. $A_j = \text{choose_splitting_attribute}(T, C, \Pi')$
8. $(T'_r, T'_l) = \text{split}(T, A_j)$
9. Run recursively for T'_l and T'_r .

Provider Aware Partitioning The algorithm first generates all possible the splitting point π , for QI attributes and the data. In the addition to a Multidimensional QI a domain space, there we consider a data provider or the data source of the each record as the additional attribute.

V. EXISTING SYSTEM

Here is an increasing need for sharing data that contain personal information from distributed database. In that we are using the heuristic algorithm for sending the data and also store the data records. Then we introduce the notion of m-privacy which guarantees that the anonymized data satisfies a given privacy constraint against any group of up to m colluding data providers. We are also present the data provider aware Anonymization algorithm with adaptive ordering technique for efficiently checking m privacy to given set of record. Our goal 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 the parties.

VI. PROPOSED SYSTEM

While the m-privacy w.r.t. any weak privacy notions do not guarantee the unconditional privacy, it offers a practical trade-off between preventing m-adversary attacks with the bounded power and the ability to publish the generalized but truthful data records. In this we will focus on checking and achieving m-privacy w.r.t. weak privacy constraints. Generalization monotonicity assumes that the original records T has been already anonymized and the uses them for further generalization. In this scenario we also introduce more general, record-based definition of the monotonicity in the order to facilitate analysis and design of the efficient algorithms for checking the m-privacy.

VII. EXPERIMENTAL RESULT

7.1 Experimental Results

There are two set of experimental results on this to compare and evaluate the different m-privacy verification algorithm given the set of record another is evaluate and compare the proposed anonymization algorithm for the given dataset with baseline algorithm in the term of both utility and efficiency.

7.2 Experimental Setup

We used combined training and test sets of the adult dataset. Records with the missing attribute values have been removed. The occupation has been chosen as a sensitive attribute. As this attribute have 14 distinct values.

Data are distributed follows the uniform or exponential distribution. We observe similar result for both of them and only report those for an exponential distribution in a paper. The privacy C is defined by k -anonymity and l -diversity. C is EG monotonic. The impact of the weight parameter to overall performance was experimentally investigated and values of α for a most efficient runs have been chosen as defaults. All experiment and algorithm parameter, and their default values are listed in table

Name	Description	Verification	Anonymization
A	Weight parameter	0.3	0.8
M	Power of m-privacy	5	3
N	Total number of data providers	-	10
n_G	Number of data providers contributing to the group	15	-
$ T $	Total number of records	-	45,222
$ T_G $	Number of records in a group	{150,750}	-
K	Parameter of k-anonymity	50	30
L	Parameter of l-diversity	4	4

Experiment Parameter and Default Values

All experiment has been performed on Sun microsystems sun fire V880 with 8 CPU's 16 GB of RAM and running Solaris 5.10.

M-privacy verification

The object of a first set of experiments is to evaluate the efficiency of the different algorithms for m-privacy verification given the set of records T_G , With respect to the previously defined privacy constraint C .

Attack Power: We first evaluated and compared the two algorithms with varying m for two algorithms significantly outperform the baseline algorithm. In contrast the baseline algorithm preserves the available number of providers in the each subgroup, which incurs the high cost for m-privacy verification. As expected, both algorithms show a peak cost when $m \sim n/2$.

VIII. RELATED WORK

The Privacy preserving data analysis and the publishing has received the considerable attention in the recent years. The most work has focused on the single data provider setting and considered a data recipient as the attacker. The large body of literature assumes limited background knowledge of a attacker and defines the privacy using relaxed adversarial notion by the considering specific types of the attacks. Representative principles include k -anonymity l -diversity, and t -closeness. Few recent works have the modelled an instance level background knowledge as corruption and studied perturbation techniques under these weak privacy notions. In the distributed setting we studied, for each data holder knows its own records, a corruption of records is the inherent element in our attack model and is further complicated by a collusive power of a data provider. On the other hand, differential privacy is the unconditional privacy guarantee for the statistical data release or the data computations. While providing the desirable unconditional privacy guarantee and non-interactive data release with the differential privacy remains the open problem. More different anonymization algorithms have

been introduced so far the including Data fly Incognito and Mondrian. In our research we considered a Mondrian algorithm as the baseline because its efficiency and extensibility. There are some work focused on the anonymization of the distributed data studied distributed anonymization for the vertically partitioned data using the k-anonymity. Zhong et al studied classification on data collected from individual data owners every record is contributed by one data owner while maintaining the k-anonymity. Jurczyk et al. proposed the notion called l'-site-diversity to ensure the anonymity for the data providers in the addition to the privacy of a data subjects. Mironov et al. studied SMC techniques to achieved differential privacy. Mohammed et al. proposed SMC techniques for the anonymizing distributed data using a notion of LKCprivacy to address high dimensional data. Our work is a first that considers the data providers as a potential attackers in the collaborative data publishing setting and the explicitly models an inherent instance knowledge of the data providers as well as Potential collusion between them for any weak privacy.

IX. CONCLUSION




In this scenario, we consider as a new type of potential attackers in the collaborative data publishing the coalition of data providers known as m-adversary. To prevent the privacy disclosure by any m-adversary we showed that the guaranteeing m-privacy is enough. We mention heuristic algorithms exploiting equivalence group monotonicity of the privacy constraints and the adaptive ordering techniques for the efficiently checking m-privacy. We introduced also the provider aware anonymization algorithm with the adaptive m-privacy checking strategies to ensure the high utility and the m-privacy of anonymized data. Our experiments confirmed that our approach achieves the better or comparable utility than existing algorithms while ensuring the m-privacy efficiently. There are more remaining research questions. Defining the proper privacy fitness score for the different privacy constraints is one of them. It also remains the question to the address and model the data knowledge of the data providers when the data are distributed in the vertical or ad-hoc fashion. It would be also the interesting to verify if our methods can be the adapted to other kinds of data such as a set valued data.

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 Velopments," *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 microdatarelease," *IEEE T. Knowl. Data En.*, vol. 13, no. 6, pp. 1010–1027, 2001.

AUTHORS PROFILE

	Maramreddy Nagendra Babu , pursuing M.Tech(CSE) Vikas College of Engineering and Technology, Nunna, Vijayawada. Affiliated to JNTU, Kakinada, A.P., India
	K Chandrasekhar Reddy , working as an Asst. Professor at Vikas College of Engineering and Technology, Nunna, Vijayawada, Affiliated to JNTU, Kakinada, A.P., India
	Prof S.V.Achutha Rao , is working as a HOD of CSE at Vikas College of Engineering and Technology, Nunna, Vijayawada, Affiliated to JNTU, Kakinada, A.P., India