International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 23 Issue 8 –NOVEMBER 2016.

# ANONYMIZING TREE STRUCTURE WITH PRIVACY PRESERVING DATA

GALLA PEDDAPUNNAIAH [#1] and YANGALADASU KIRAN[*2]

[#] *M.Tech (CSE), M.V.R COLLEGE OF ENGINEERING & TECHNOLOGY, A.P., India.*

[*] *Assistant Professor, Dept. of Computer Science & Engineering, M.V.R COLLEGE OF ENGINEERING & TECHNOLOGY, A.P., India.*

***Abstract*— Today comprehensively arranged frameworks represents a danger to individual security and authoritative confidentiality. So data anonymization strategies have been proposed so as to permit handling of individual information without compromising user's privacy. Data anonymization is a type of information sanitization whose intent is privacy protection. It is the process of either encrypting or removing personally identifiable information from data sets, so that the people whom the data describe remain anonymous. Data anonymization techniques have been proposed in order to allow processing of personal data without compromising user's privacy. the data management community is facing a big challenge to protect personal information of individuals from attackers who try to disclose the information. So data anonymization strategies have been proposed so as to permit handling of individual information without compromising user's privacy. Data anonymization is a type of information sanitization whose intent is privacy protection. It is the process of either encrypting or removing personally identifiable information from data sets, so that the people whom the data describe remain anonymous. We are presenting k(m;n)-anonymity privacy guarantee which addresses background knowledge of both value and structure using improved and automatic greedy algorithm. (k (m,n) - obscurity ensure) A tree database D is considered k (m,n) - unknown if any assailant who has foundation information of m hub names and n auxiliary relations between them (ancestor descendant), is not ready to utilize this learning to distinguish not as much as k records in D. A tree dataset D can be transformed to a dataset D0 which complies to k (m,n) - anonymity, by a series of transformations. The key idea is to replace rare values with a common generalized value and to remove ancestor descendant relations when they might lead to privacy breaches.**

***Index Terms*— Anonymity, Privacy, Tree data, k m –anonymization.**

## I. INTRODUCTION

The k-anonymity privacy for publishing micro data requires that each equivalence class contains at least k records. Many authors have studied that k-anonymity cannot prevent attribute disclosure. The technique of l-diversity has been introduced to address this; ldiversity requires that each equivalence class must have at least well-represented values for every sensitive attribute. In this paper, we show that l-diversity has many limitations. In particular, it is not necessary or sufficient to prevent attribute disclosure. Motivated by these limitations, we propose a new method to detect privacy which is called as closeness. We first present the base model t-closeness, which includes the distribution of sensitive attributes in any of the equivalence classes is near to the distribution of the attribute in the overall table ( i.e., the difference between the two given distributions should be no more than threshold value t). tcloseness that gives higher utility. We present our methode for designing a distance measure between given two probability distributions and give two distance measures. Here we discuss the methode for implementing closeness as a privacy concern and illustrate its advantages through examples and experiments. The concept of k-anonymity tries to express on the private table PT to be released, one of the main necessity that has been followed by the statistical community Agencies releasing the data, and according to which the released data should be equivalent related to no less than a certain number of respondents. The set of attributes involved in the private table, also externally obtainable and therefore exploitable for linking, is called quasi-identifier . The requirement just expressed is then translated in the k-anonymity requirement, which states that every tuple released cannot be related to fewer than k respondents. While k-Anonymity forces one to derive an attribute value even if all but one of the records in a cluster have the identical value, the above clustering-based anonymization technique allows us to pick a cluster center whose value along this attribute dimension is the identical as the common value, thus enabling us to release more information without losing privacy. K-anonymity is one of anonymization approaches proposed by Samarati and Sweeney[6] that each record in dataset cannot be distinguished with at least another (k-1) records under the projection of quasi-identifiers of dataset after a series of anonymity operations (e.g. replace specific value with general value). K-anonymity assures that the probability of uniquely representing an individual in released dataset will not great than 1/k. For example in table 1, we learn about Miss Yoga has diabetes by linking census data table with patient data table by Birthday, Sex and ZipCode attributes even removing identifier. What if it cannot uniquely determine a record?

Thus attacker has no ability to identify sensitive information with full confidence. How to make patient table in Table 1 meet 2-anonymity? One of practical ways is that replacing data with year for Birthday attribute and using * replace the last two character of ZipCode attribute. K-anonymity has been extensively studied in recent years [7,8,9,10]. After 2-anonymity, it cannot infer that Miss Yoga has diabetes, or maybe she has cancer. Because in patient data table, there are two records that can be linked to one record in census data table about Miss Yoga. We can see that k-anonymity has an effective impact on this scenario.

## II. LITERATURE REVIEW

[1]R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai, and L. Xiong , "Publishing set-valued data via differential privacy" This states set-valued data provides enormous opportunities for various data mining techniques. This mentioned the problem of preparing set-valued data for data mining tasks under the rigorous differential privacy model. All existing data producing methods for set-valued data are based on partition based privacy models, for example k-anonymity, which are unsafe to privacy attacks based on background knowledge. In contrast, differential privacy provides strong privacy guarantees individualistic of an adversary's background knowledge and computational power. Existing data publishing approaches for differential privacy, however, are not sufficient in terms of both utility and scalability in the context of set-valued data due to its high dimensionality. It indicate that set-valued data could be efficiently released under differential privacy with guaranteed beneficial with the help of context-free taxonomy trees. We propose a probabilistic top-down partitioning algorithm to produce a differentially private release, which scales linearly with the input data size. It also indicates the applicability of our idea to the context of relational data. We prove that our result is (ϱ, δ)-applicable for the class of counting queries, the foundation of many data mining tasks.

[2] R. J. Bayardo and R. Agrawal, "Data Privacy through Optimal k-Anonymization." Data de-identification reconciles the demand for release of data for research purposes and it demands individuals privacy. This paper proposes and evaluates an optimization algorithm for the powerful procedure of de-identification known as k-anonymization. A k-anonymized dataset has the property that each record is indistinguishable from at least other k - 1. More simple restrictions of optimized k-anonymity are NP-hard, leading to significant computational challenges. It present a new approach to exploring the space of possible anonymizations that tames the combinatorics of the problem, and it develop data-management strategies to reduce reliance on expensive operations like as sorting. Through experiments on real census data, the resulting algorithm can find optimal kanonymizations under two illustrative cost measures and a wide range of k. The algorithm can produce good anonymizations in circumstances where the input data or input parameters restrict finding an optimal solution in reasonable time. This algorithm to explore the effects of various coding approaches and problem variations on anonymization quality and performance. This result signifying optimal k-anonymization of a non-trivial dataset under a general model of the problem.

[3]J. Cheng, A. W.-c. Fu, and J. Liu , " K-isomorphism : privacy preserving network publication against structural attacks." states Serious concerns on privacy protection in social networks have been increased in recent years; however, research in this area is still in its begining. The problem is demanding due to the diversity and complexity of graph data, on which an adversary can help many types of background knowledge to conduct an attack. Our investigations show that k-isomorphism, or anonymization by forming k pairwise isomorphic subgraphs, is both sufficient and necessary for the protection. The problem is shown to be NP-hard. We devise a number of techniques to enhance the anonymization efficiency while retaining the data utility.

[4] G. Cormode, "Personal privacy vs population privacy: learning to attack anonymization." states that Over the last decade great strides have been made in expanding techniques to compute functions privately. In particular, Differential Privacy gives strong promises about closure that can be drawn about an individual. In this paper, we consider the capability of an attacker to use data meeting privacy definitions to build an accurate classifier. Even under Differential Privacy, such classifiers can be used to deduce "private" attributes accurately in realistic data.

## III. RELATED WORK

Anonymity for relational data has received considerable attention due to the need of several organizations to publish data (often called microdata) without revealing the identity of individual records. Even if the identifying attributes (e.g., name) are removed, an attacker may be able to associate records with specific persons using combinations of other attributes (e.g., hzip, sex, birth date i), called quasi-identifiers (QI). A table is k-anonymized if each record is indistinguishable from at least k − 1 other records with respect to the QI set [18, 19]. Records with identical QI values form an anonymized group. Two techniques to preserve privacy are generalization and suppression [19]. Generalization replaces their actual QI values with more general ones (e.g., replaces the city with the state); typically, there is a generalization hierarchy (e.g., city→state→country). Suppression excludes some QI attributes or entire records (known as outliers) from the microdata. The privacy preserving transformation of the microdata is referred to as recoding. Two models exist: in global recoding, a particular detailed value must be mapped to the same generalized value in all records. Local recoding, on the other hand, allows the same detailed value to be mapped to different generalized values in each anonymized group. The recoding process can also be classified into single-dimensional, where the mapping is performed for each

attribute individually, and multi-dimensional, which maps the Cartesian product of multiple attributes. Our work is based on global recoding and can be roughly considered as single-dimensional (although this is not entirely accurate), since in our problem all items take values from the same domain. [13] proved that optimal k-anonymity for multidimensional QI is NP-hard, under both the generalization and suppression models. For the latter, they proposed an approximate algorithm that minimizes the number of suppressed values; the approximation bound is O(k · logk). [2] improved this bound to O(k), while [17] further reduced it to O(log k). Several approaches limit the search space by considering only global recoding. [4] proposed an optimal algorithm for single-dimensional global recoding with respect to the Classification Metric (CM ) and Discernibility Metric (DM ), which we discuss in Section 3.3. Incognito [9] takes a dynamic programming approach and finds an optimal solution for any metric by considering all possible generalizations, but only for global, full-domain recoding. Full-domain means that all values in a dimension must be mapped to the same level of hierarchy. For example, in the country→continent→world hierarchy, if Italy is mapped to Europe, then Thailand must be mapped to Asia, even if the generalization of Thailand is not necessary to guarantee anonymity. A different approach is taken in [16], where the authors propose to use natural domain generalization hierarchies (as opposed to user-defined ones) to reduce information loss. Our optimal algorithm is inspired by Incognito; however, we do not perform full-domain recoding, because, given that we have only one domain, this would lead to unacceptable information loss due to unnecessary generalization. As we discuss in the next section, our solution space is essentially different due to the avoidance of full-domain recoding. The computational cost of Incognito (and that of our optimal algorithm) grows exponentially, so it cannot be used for more than 20 dimensions. In our problem, every item can be considered as a dimension. Typically, we have thousands of items, therefore we develop fast greedy heuristics (based on the same generalization model), which are scalable to the number of items in the set domain.

## IV. PROPOSED SYSTEM

The anonymization methodology does not just sum up qualities that partake in uncommon thing blends additionally rearranges the structure of the records. We concentrate on the anonymization of tree-organized individual records where qualities are connected through basic connections. The proposed anonymization strategies address datasets like D. The first information possessed by the distributer may be in an alternate structure, e.g., a multirelational plan. Proposed the use of disassociation in set-esteemed information, where an exchange could be part in two or more parts furthermore anonymize exchange information. We propose a security ensure that secures the personality of the people who are connected with tree records from aggressors by augmenting the km -anonymity guarantee [6] to address auxiliary information. K m -namelessness ensures that any assailant, who knows up to m components of a record, won't have the capacity to distinguish not as much as k records in the distributed information. We characterize k (m,n) - obscurity as: Definition 1: (k (m,n) - obscurity ensure) A tree database D is considered k (m,n) - unknown if any assailant who has foundation information of m hub names and n auxiliary relations between them (ancestor descendant), is not ready to utilize this learning to distinguish not as much as k records in D. A tree dataset D can be transformed to a dataset D0 which complies to k (m,n) - anonymity, by a series of transformations.The key idea is to replace rare values with a common generalized value and to remove ancestordescendant relations when they might lead to privacy breaches.

## V. CONCLUSION & FUTURE WORK

Our analysis techniques allow trace publishers to compute an upper bound for the risk of host de-anonymization in the context of adversaries assumed capable of collecting a given class of external information. In the future we hope to use these techniques to formally evaluate partial prefix preservation alternatives which can maximize utility relative to a desired level of trace privacy. To deal with bigger and more expressive datasets, we plan to work with the Greedy Cut Search Algorithm GCS, which we assume would follow the most promising paths and can significantly reduce the search space and computational time.

## REFERENCES

[1] Australian Privacy Act. www.austlii.edu.au/au/legis/cth/consol act/pa1988108.
[2] Canadian Privacy Act. laws-lois.justice.gc.ca/eng/acts/P-21/.
[3] Data Protection Act 1998, UK. www.legislation.gov.uk/ukpga/1998/29/contents.
[4] GR Law. www.dpa.gr/portal/page? pageid=33,43560& dad=portal.
[5] M. Nergiz, C. Clifton, and A. Nergiz. Multirelational k-anonymity. IEEE TKDE, pages 104–1117, 2009.
[6] M. Terrovitis, N. Mamoulis, and P. Kalnis. Privacy- preserving Anonymization of Set-valued Data. PVLDB, 1(1), 2008.
[7] M. Terrovitis, N. Mamoulis, and P. Kalnis. Local and global recoding methods for anonymizing set-valued data. The VLDB Journal, 2010.
[8] R. Chaytor and K. Wang. Small-domain randomization: Same privacy more utility. In VLDB, 2010.
[9] R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai, and L. Xiong. Publishing set-valued data via differential privacy. PVLDB,4(11):1087–1098, 2011.
[10] J. Cheng, A.W.-c. Fu, and J. Liu. K-isomorphism: privacy preserving network publication against structural attacks. In SIGMOD, 2010.
[11] G. Cormode , Personal privacy vs population privacy: learning to attack anonymization.

**GALLA PEDDAPUNNAIAH** is a student of M.V.R College of Engineering and Technology, PARITALA . She is presently pursuing her M.Tech degree from JNTU,Kakinada.

**MrYANGALADASUKIRAN** is presently working as Assistant professor in CSE department,M.V.RCollege of Engineering and Technology, PARITALA.