Enhanced Privacy Preserving Accesscontrol in Incremental Datausing Microaggregation

Ravi Sharma
Department of Computer Engineering
Indira College of Engineering and Management, Pune
raviusharma73@gmail.com

Manisha Bharati
Department of Computer Engineering
Indira College of Engineering and Management,Pune
manisha.bharati@indiraicem.ac.in

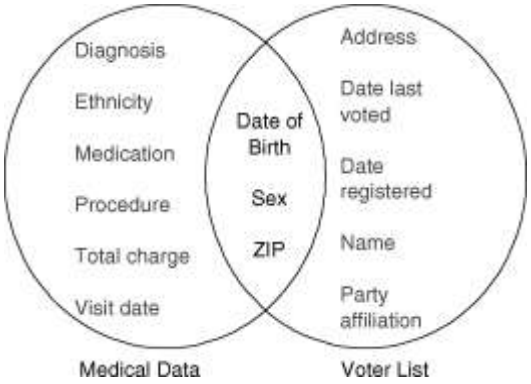
Abstract-In microdata releases, main task is to protect the privacy of data subjects. Microaggregation technique use to disclose the limitation at protecting the privacy of microdata. This technique is an alternative to generalization and suppression, which use to generate k-anonymous data sets. In this dataset, identity of each subject is hidden within a group of k subjects. Microaggregation perturbs the data and additional masking allows refining data utility in many ways, like increasing data granularity, to avoid discretization of numerical data, to reduce the impact of outliers. If the variability of the private data values in a group of k subjects is too small, k-anonymity does not provide protection against attribute disclosure. In this work Role based access control is assumed. The access control policies define selection predicates to roles. Then use the concept of imprecision bound for each permission to define a threshold on the amount of imprecision that can be tolerated. So the proposed approach reduces the imprecision for each selection predicate. Anonymization is carried out only for the static relational table in the existing papers. Privacy preserving access control mechanism is applied to the incremental data.

Keywords: Generalization, k-Anonymity, Microaggregation, Microdata, Masking, Suppression.

*****

I. INTRODUCTION

Data anonymization is a form of information refinement whose intent is privacy protection. It is the process of either converting or removing personally recognizable information, so that the individuals whom the data describe remain unknown. Creating an anonymized data set that is suitable for public release is basically a matter of finding a good stability between exposer risk and information loss. To provide highest utility for user, system must have to release the original data set, which suffers the maximum disclosure risk for the subjects in the data set. k-Anonymity is the oldest among the syntactic privacy models. Models in this type address thetrade-off between confidentiality and utility by requiring the anonymized data set to follow a specific arrangement that is known to limit the risk of disclosure. Yet, the method to be used to generate such an anonymized data set is not specified bythe privacy model and must be selected to maximize data utility (because satisfying the model already ensures privacy). The main methodology to obtain an anonymized data set sustaining k-anonymity or any of its refinements is established on generalization and suppression. The objective of generalization-based methods is to find the minimal generalization that fulfills the requirements of the essential privacy model. These algorithms can be adapted to the abovementioned k-anonymity refinements: it is simply a matter of introducing the additional limitations of the target privacy model when testing whether a specific generalization is viable.



Therefore, the objective is to limit the disclosure risk to an acceptable level while maximizing the utility. This is achieved by anonymizing the data before release. The first step of anonymization is to remove explicit identiers. However, this is not enough, as an adversary may already know the quasi-identier values of some individuals in the table. This knowledge can be either from personal knowledge (e.g., knowing a particular individual in person), or from other publicly-available databases (e.g., a voter registration list) that include both explicit identiers and quasi-identiers. This is also known as background knowledge attack.A common anonymization approach is generalization, which replaces quasi-identier values with values that are less-specific but semantically consistent. As a result, more records will have the same set of QI values.This leads to the definition of an equivalence class. An equivalence class of an anonymized table and of the

sensitive attribute is defined as a set of records that have the same values for all the QIs. Another is suppression, which suppresses the value of an attribute if that value causes the overall k-anonymity or any other privacy measure to fail. But the suppression is minimized using the maximum suppression count or percentage.

II. MOTIVATION

The main issue with t-closeness is its specificity of application. It is totally dependent upon the dataset in question because of its dependence on k-anonymity for complete execution. This dependence causes high overload in deployment of any anonymization algorithm involving t-closeness.The main emphasis of this was on reducing the effective time in deploying anonymization while also reducing the dependence on specific datasets for every step of the process starting from the definition of Domain Generalization Hierarchy to Sensitive Attribute selection.

III. LITERATURE SURVEY

[1] Jordi Soria-Comas, Josep Domingo-Ferrer, Fellow, IEEE, David S anchez, and Sergio Mart nez, tCloseness through Microaggregation: Strict Privacy with Enhanced Utility Preservation, IEEE transactions on knowledge and data engineering, vol. 27, no. 11, november 2015.

In this work, author have proposed and evaluated the use of microaggregation as a technique to attain k-anonymous t-closeness. The a priori remunerations of microaggregation vs generalization recoding and local suppression have been discussed. Global recoding may recode additional thanneeded, whereas local recoding complicates data analysis by mixing together values corresponding to dissimilar levels of generalization. Also, recoding creates a greater loss of granularity of the data, is more affected by outliers, and changes algebraic values to ranges. Regarding local suppression, it complicates data analysis with missing values and is not obvious to combine with recoding in order todecrease the volume of generalization. Microaggregation is free from all the above downsides. They proposed and evaluated three different microaggregation based algorithms to generate k-anonymous t-close data sets. The first one is a simple merging step that can be run after any microaggregationalgorithm. The other two algorithms, k-anonymity-first and t-closeness-first, take the t-closeness requirement into account at the moment of cluster formation during microaggregation. The t-closeness first algorithm considers t-closeness earliest and provides the best results: smallest average cluster size, smallest SSE for a given level of t-closeness, and shortest run time [1].

[2] Jianneng Cao, Panagiotis Karras, PanosKalnis, Kian-Lee Tan propose a SABRE: a Sensitive Attribute Bucketization and Redistribution framework for t-closeness, May 2009.

This paper proposed SABRE, a different framework for distribution-aware microdataanonymization based on the t-closeness principle. In this work author explain the need of microdata privacy and cover the gap with SABRE, a SA Bucketization and REdistribution framework for t-closeness. SABREfirst greedily divide a table into buckets of parallel SA values and then redistributes the tuples of each bucket into dynamically determined ECs. They explained that this approach is expedited by a property of the Earth Movers Distance (EMD) that employ as a measure of distribution closeness: If the tuples in an EC are picked proportionally to the sizes of the buckets they hail from, then the EMD of that EC is strongly upper-bounded using localized upper bounds derived for each bucket. Their work shown that if the t-closeness constraint is properly followed during partitioning, then it is obeyed by the derived ECs too. They improve two instantiations of SABRE and extend it to a streaming situation. Extensive experimental evaluation demonstrates that SABRE achieves information quality superior to schemes that merely applied algorithms tailored for other prototypes to t-closeness, and can be much quicker as well. They determine as, SABRE provides the best known resolution of the tradeoff between privacy, information quality, and computational efficiency with a t-closeness guarantee in mind [2].

[3] JosepDomingoFerrer, Hybrid microdata using microaggregation 10 April 2010.

Researcher has presented a new hybrid data group method whose goal is to produce hybrid microdata sets that can be free with low disclosure risk and acceptable data utility. The method combines microaggregation and any synthetic data generator. Depending on a single integer parameter k, it can yield data which are very adjacent to the original data (or even the original data themselves if k = 1) or wholly synthetic data (when k is equal to the number of records in the data set). Thus, the parameterization of the method is simpler and more intuitive for users than in the hybrid data generation alternatives proposed so far. For the particular case of numerical microdata, shown that the hybrid data set obtained preserves the mean vector and the covariancematrix of the original data set. Furthermore, if a worthy microaggregation heuristic yielding a small intracluster variance is used: Approximate preservation of third-order central moments has been verified; Approximate preservation of fourth-order central moments has been empiricallyexposed; For subdomains (i.e. data subsets), approximate preservation of means, variances, covariances, and third-order and fourth-order central moments has been

empirically presented; this feature was not open by the current hybrid or synthetic data generation methods in the literature. Last but not least, compared to plain multivariate microaggregation, the new method offers better data utility for confidential attributes (due to variance and covariance preservation) and at the same time it achieves a lower disclosure threat [3].

[4]JosepDomingoFerrer , Jordi SoriaComas, From t-closeness to differential privacy and vice versa in data 4 anonymization, 11 November 2014.

This paper has highlighted and exploited several links between k-anonymity, t-closeness and e-differential privacy. These models are more related than believed so far in thecase of data set anonymization. On the one hand author have introduced the concept of stochastic t-closeness, which, in its place of being based on the empirical distribution like classic t-closeness, is based on the distribution induced by a stochastic function that alters the confidential attributes. Theyhave shown that k-anonymity for the quasi-identifiers joint with e-differential privacy for the trusted attributes yields stochastic t-closeness, with a function of e, the size of the data set and the size of the similarity classes. This result shows that differential privacy is robust than t-closeness as a privacy notion. From a practical point of view, it provides a way of generating an anonymized data set that fulfills both (stochastic) t-closeness and differential privacy. On the other hand, they have demonstrated that the k-anonymity family of models is great enough to achieve differential privacyin the context of data set anonymization, provided that a few reasonable assumptions on the intruder's side knowledge hold. They have shown that closeness implies differential privacy. Apart from partitioning into equivalence classes, a prior bucketization of the values of the confidential attributeis required. The optimal size of the buckets and the optimal size of equivalence classes have been determined [4].

[5] Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian, tCloseness: Privacy Beyond kAnonymity and Diversity, 2007 IEEE.

In this paper the stated as, While k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. The notion of diversity attempts to solve this problem by requiring that each equivalence class has at least well-represented values for each sensitive attribute. They have shown that diversity has a number of limitations and have proposed a novel privacy notion called t-closeness, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). One key novelty of approach is that they separate the information gain an observer can get from a released data table into two parts: that about all population in the released data and that about specific individuals. This enables to limit only the second kind of information gain. They use the Earth Mover Distance measure for t-closeness requirement; this has the advantage of taking into consideration the semantic closeness of attribute values[5].

IV. PROPOSED SYSTEM

Traditionally, research in the database community in the area of data security can be broadly classified into two-access control research and data privacy research. The idea of access control is to authorize a user to access only a subset of the data. This authorization is enforced by explicitly rewriting queries to limit access to the authorized subset. The main limitation of traditional access control mechanism in supporting data privacy is that it is "black and white" [10]. That is, the access control mechanism offers only two choices: release no aggregate information thereby preserving privacy at the expense of utility, or release accurate aggregates thus risking privacy breaches for utility. Thus, a hybrid system is needed that combines a set of authorization predicates restricting access per user to a subset of data and privacy preserving mechanism.

In the proposed system, a relational table, containing sensitive information, is taken. This table is anonymized. The database contains incremental data, with the administrator having the permission to add data into the table. The table has to be anonymized each time data is added into the database. Role based access control is being used here. The concept of role-based access control (RBAC) began with multi-user and multi-application online systems. The central notion of RBAC is that permissions are associated with roles and users are assigned to appropriate roles. This greatly simplifies management of permissions. Roles are created for the various job functions in an organization and users are assigned roles based on their responsibilities and qualifications. Users can be easily reassigned from one role to another [3]. The access control policies define selection predicates available to roles while the privacy requirement is to satisfy the k-anonymity or ℓ-diversity [6].

Another constraint that needs to be satisfied by the privacy protection mechanism is the imprecision bound for each selection predicate. Query imprecision is defined as the difference between the number of tuples returned by a query evaluated on an anonymized relation R* and the number of tuples for the same query on the original relation R [9]. The imprecision bound for each permission define a threshold on the amount of imprecision that can be tolerated. If imprecision bound is not satisfied, then unnecessary false alarms are generated due to high rate of false positives. The imprecision bound is preset by the administrator and this

 ISSN: 2321-8169
342 – 347

information is not shared with the users because knowing the imprecision bound can result in violating the privacy requirement. The imprecision bound will be different for the different roles that exist within the organization. So, in a nutshell, it can be said that the privacy preserving module anonymizes the data to meet the privacy requirement along with the imprecision bound for each permission.

V. SYSTEM ARCHITECTURE

Access control mechanisms are used to ensure that sensitive information is available to authorized users only. When Privacy Protection Mechanism (PPM) is not used, authorized users can misuse the sensitive information and the privacy of the consumer is compromised. Privacy requirement is satisfied by PPM which uses suppression and generalization approaches to anonymize the relational data. K-anonymity or l-diversity is used to anonymize and satisfy privacy requirement. However, privacy is obtained by the precision of the authorized information. The anonymity technique can be used with an access control mechanism to ensure both security and privacy of the sensitive information. In this paper Role based access control is assumed. The access control policies define selection predicates to roles. Then we use the concept of imprecision bound for each permission to define a threshold on the amount of imprecision that can be tolerated. So the proposed approach reduces the imprecision for each selection predicate. Anonymization is carried out only for the static relational table in the existing papers. In this paper privacy preserving access control mechanism is applied to the incremental data.

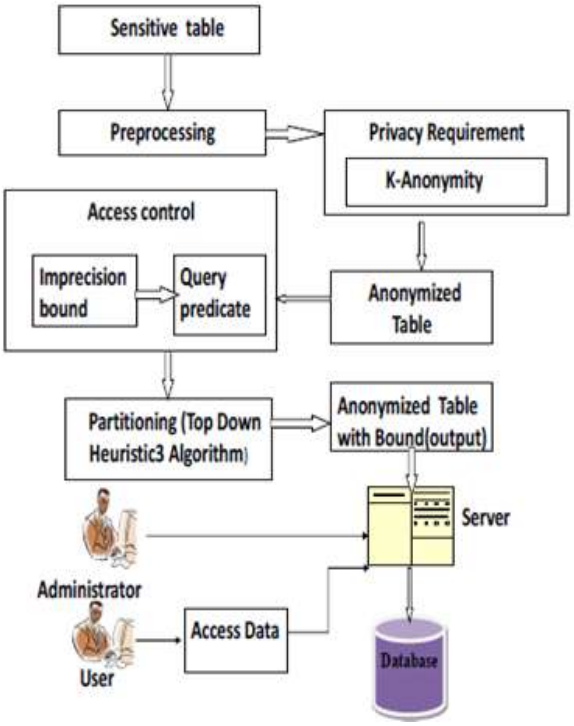


Fig 1: System Architecture

VI. SYSTEM MODULES

Preprocessing- Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user. In this module, Dataset is taken from file which is downloaded from web. For example UIMechanism. Dataset contains raw data. It will not be in proper format. We cannot use that raw data for anonymization. So we have to make it into proper format. Dataset contains data in the text file is converted in to the table form for further processing. Forming table from the selected Dataset is known as pre-processing. After that anonymization technique is applied for the Dataset.

Cluster Formation -A cluster is a subset of objects which are similar. Clustering is a process of partitioning a set of data into a set of meaningful sub-classes, called clusters. Help users understand the natural grouping or structure in a data set. Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.
After preprocessing we can get a table with lots of data. From that table we do the anonymization process. In that we have to select related attribute to form cluster. Clustering is the process of partitioning. By using this clustering concept we can easily anonymize data. If we didn't select related attribute, it will be very tough to search and find data.

Anonymization -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 Datasets, so that the people whom the data describe remain anonymous. By clustering process we can partition the large database.
Now we have to anonymize the data. Anonymization is the process of converting a text in a range of value or into a non-readable symbol form. After clustering process completed we need to anonymize data using algorithms. Anonymization process will improve efficiency of data. Because of anonymization we can also save time. We can use methods like Generalization and Suppression to anonymize data.

Collecting AnonymizedData -Because of clustering we cannot get all anonymized data together. We will get cluster by cluster after anonymization. Now we have to gather all data after anonymization. So we can store the total anonymized data in the server. Then we can release this anonymized dataset for further use. So we can efficiently use the dataset with privacy.

 ISSN: 2321-8169
342 – 347

Privacy Protection Mechanism When organization shares data, they must do so in a way that fully protects individual privacy. The privacy protection mechanism is used to guarantee the confidentiality or privacy of the data by using anonymization. The anonymization techniques such as generalization and suppression are used to ensure the kanonymity. But generalization and suppression are decrease the quality or utility of data. To overcome this, Top Down Selection Mondrian algorithm is used to achieve the k-anonymity.

Access Control: Access Control Mechanisms are used to care for the delicate information from unauthorized users. The privacy of persons may still be matter to identity disclosure by authorized users. Privacy Protection Mechanism is used to preserve the privacy of data defendants. It can be done by either suppression or generalization or both. Anonymization due to PPM introduces some changes that affect the accuracy of the data required. It introduces imprecision to the data. The
aim of this work is to anonymize the data considering the imprecision bound for each of the query predicates. Role-based Access Control (RBAC) allows defining permissions on objects based on roles in an organization. This work considered the imprecision added to each permission in the anonymized micro data.

VII. ALGORITHM

Top-Down Heuristic Algorithm:

Input: T,K,Q AND BQi where T for total tuples, K=cluster, Q=query, B=bound for query i.
Output: P the output partitions
STEPS:
Initialize candidate partitions CP < T
For all CP do the following:
1) Find the queries that overlap in that partition.
2) Select the queries with least IB and IB 0
3) Select the query with smallest bound.
4) Create query cut.
5) if(skewed partition) then feasible cut is found andadd to CP.
Else
Reject the cut.
_ Return (P).

Imprecision Query: Imprecision is defined as the variance between the amount of tuples returned by a query calculated on an anonymized relation T* and the total number of tuples for the same query on the main relation T.

Imprecision bound: The query imprecision bound, represented by BQi , is the total imprecision acceptable for a query predicate Qi and is set by the access control administrator. The query imprecision slack for a query is well-defined as the difference between the query imprecision bound and the real query imprecision.

Query cut: It is defined as splitting the partition beside the query intervals. Overlap semantics include all tuples in all partitions that overlap the query region. This will add false positives to the original query result. The imprecision under any query evaluation scheme is reduced if the amount of
tuples in the partitions that overlap the query region can be minimized.
The query given by the user is rewritten according to the authorized query predicates assigned for the role. Then that query is evaluated over the tuples space. Each of the queries is considered as a hyper-rectangle. Each of the partitions are measured as hyper-rectangles. The rectangles equivalent to the partitions and the query region overlap if the partition has tuples satisfying the query predicate. Select the partitions that have low imprecision cost for the query. Imprecision cost refers to the number tuples that are present in the partition but not in the query. Imprecision for a query is obtained by the sum of imprecision cost of all the partitions.

VIII. RESULTS

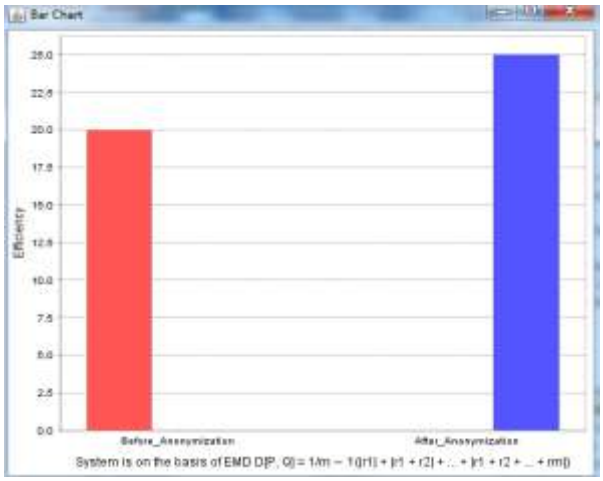


Fig 2: Comparison between Before and after anonymization using EMD

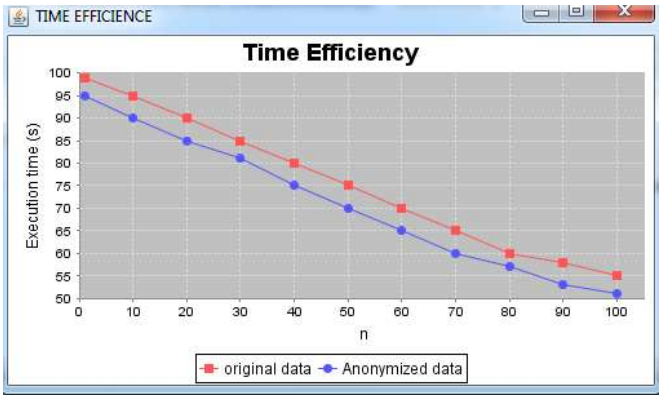


Fig 3: Time efficiency between original data and anonymized data

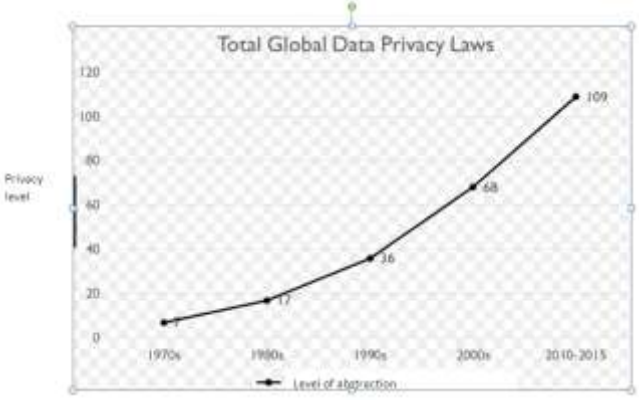


Fig 4:Graph between privacy and algorithm

IX. CONCLUSION

The proposed work uses microaggregation as a method to attain k-anonymous t-closeness. This work proposed and evaluates different microaggregation based algorithms to generate k- anonymous t-close data sets. This system is proposed to provide strictest privacy with additional masking and reducing the impact of outliers and avoiding discretization of numerical data.

ACKNOWLEDGMENT

I thankful to the researchers and publishers for makingtheir resources available and teachers for their guidance. I would like to thank the authorities of SavitribaiPhulePune University and members of International Journal on Recent and Innovation Trends in Computing and CommunicationJournal. I expressgreat many thanks to reviewer for their comments and thecollege authorities for providing the required infrastructure andsupport.

REFERENCES

[1] Jordi SoriaComas, JosepDomingoFerrer, Fellow, IEEE, David S anchez, and Sergio Mart nez, tCloseness through Microaggregation: Strict Privacy with Enhanced Utility Preservation, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 11, NOVEMBER 2015.

[2] Jianneng Cao, Panagiotis Karras, PanosKalnis, KianLee Tan propose a SABRE: a Sensitive Attribute Bucketization and Redistribution framework for tcloseness, May 2009.

[3] JosepDomingoFerrer, Hybrid microdata using microaggregation 10 April 2010.

[4] JosepDomingoFerrer , Jordi SoriaComas, From tcloseness to differential privacy and vice versa in data 4 anonymization, 11 November 2014.

[5] Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian, tCloseness: Privacy Beyond kAnonymity an Diversity, 2007 IEEE.

[6] D. Defays and P. Nanopoulos. Panels of enterprises and confidentiality: the small aggregates method. In Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys, pp. 195204, Ottawa, 1992.Statistics Canada.

[7] J. Domingo-Ferrer and J. M. Mateo-Sanz.Practical data-oriented microaggregation for statistical disclosure control. IEEE Transactions on Knowledge and Data Engineering, 14(1):189-201, 2002.

[8] J. Domingo-Ferrer, D. S.anchez and G. Rufian-Torrell.Anonymization of nominal data based on semantic marginality. Information Sciences, 242:3548, 2013.

[9] J. Domingo-Ferrer and V. Torra.Ordinal, continuous and heterogeneous k-anonymity through microaggregation.Data Min. Knowl.Discov., 11(2):195212, 2005.