www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 4 Issue 12 Dec 2015, Page No. 15161-15164

Strict Privacy With Enhanced Utility Preservation By N, T Closeness Through Microaggregation

Ms.S.Nathiya , Mr. U.Gowrisankar , Ms. P.Jayapriya

[#1] *M.E Scholar, Department of CSE, College, V.S.B.Engineering College, Karur,*
[*2] *Assistant Professor, Department of CSE, V.S.B.Engineering College, Karur,*
[*3] *M.E Scholar, Department of CSE, College, V.S.B.Engineering College, Karur,*
nathiya.s.cse @gmail.com
gowriker@rediffmail.com
jayapriya311 @gmail.com

Abstract:- ***Micro aggregation is a technique for disclosure limitation aimed at protecting the privacy of data subjects in micro data releases, like releasing medical data, census data etc… It has been subjected to generalization and suppression to generate k-anonymous data sets, where the identity of each subject is hidden within a group of k subjects. Unlike generalization, micro aggregation perturbs the data and this additional masking freedom allows improving data utility in several. Existing algorithms like k-anonymity and t-closeness is based on generalization and suppression k-anonymity does not deal with attribute disclosure and hence the work focuses on closeness. This paper proposes and shows how to use micro aggregation to generate n,t close data sets. The advantages of micro aggregation with n,t-closeness are analyzed, and the ultimate aim of the project is to make comparative analysis and evaluate micro aggregation algorithms for t-closeness and n,t closeness . There are many real-life situations in which personal data is stored: (i) Electronic commerce results in the automated collection of large amounts of consumer data. These data, which are gathered by many companies, are shared with subsidiaries and partners. (ii) Health care is a very sensitive sector with strict regulations. In the U.S., the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) requires the strict regulation of protected health information for use in medical research. In most western countries, the situation is similar.***

Keywords:- k-anonymity , HIPAA , t-closeness

1.Introduction

Generating an anonymized data set that is suitable for public release is essentially a matter of finding a good equilibrium between disclosure risk and information loss. Releasing the original data set provides the highest utility to data users but greatest disclosure risk for the subjects in the data set. On the contrary, releasing random data incurs no risk of disclosure but provides no utility. K-Anonymity, in particular, seeks to make record re-identification unfeasible by hiding each subject within a group of k subjects. To this end, k-anonymity requires each record in the anonymized data set to be indistinguishable from another k -1 records as far as the quasi-identifier attributes are concerned. Online browsing methods use a representing concept-based user profiles. The weights of the vector elements, which could be positive or negative, represent the interestingness (or uninteresting nests) of the user on the concepts. Micro-Aggregation is Statistical Disclosure Control (SDC), also known as Statistical Disclosure Limitation (SDL), seeks to transform data in such a way that they can be publicly released whilst pre-serving data utility and statistical confidentiality, where the latter means avoiding disclosure of information that can be linked to specific individual or corporate respondent entities. When we micro-aggregate data we have to keep two goals in mind: (i) Preserving data utility. To do so, we should introduce as little noise as possible into the data i.e. we should aggregate similar elements instead of divergent ones. In the example given in Figure 1 for a security parameter k = 3, groups of three elements are built and aggregated. (ii)Protecting the privacy of the respondents. Data have to be modified to make re-identification difficult i.e. by increasing the number of aggregated elements, we in- crease data privacy. In the example given in Figure 1, after aggregating the chosen elements, it is impossible to distinguish them, so that the probability of linking any respondent is inversely proportional to the number of aggregated elements.

2.Related work

This part talks about the researches that have done previously about K. LeFevre, D. J. DeWitt, and R. Ramakrishnan[2] suggest to micro-aggregation is a clustering problem with minimum size constraints on the resulting clusters or groups; the number of groups is unconstrained and the within-group homogeneity should be maximized. In the context of privacy in statistical databases, micro-aggregation is a well-known approach to obtaining anonymized versions of confidential microdata. To propose a new heuristic method for multivariate microaggregation called V-MDAV. V-MDAV stands for Variable-size Maximum Distance to Average Vector. It improves on the well-known MDAV method in terms of lower SSE while maintaining an equivalent computational cost.

The above scheme is tailored with a variation that is proposed in N. Li, T. Li, and S. enkatasubramanian [3]

suggest This new property requires that there be at least *p* different values for each confidential attribute within the records sharing a combination of key attributes. Like *k*-anonymity, the algorithm originally proposed to achieve this property was based on generalisations and suppressions; when data sets are numerical this has several data utility problems, namely turning numerical key attributes into categorical, injecting new categories, injecting missing data, and so on. In this article, we recall the foundational concepts of micro-aggregation, *k*-anonymity and *p*-sensitive *k*

J. Soria-Comas and J. Domingo-Ferrer suggest k-Anonymity[4] is a privacy property used to limit the risk of re-identification in a micro data set. A data set satisfying k-anonymity consists of groups of k records which are indistinguishable as far as their quasi-identifier attributes are concerned. Hence, the probability of re-identifying a record within a group is 1/k. We introduce the probabilistic k-anonymity property, which relaxes the indistinguishability requirement of k-anonymity and only requires that the probability of re-identification be the same as in k-anonymity. Two computational heuristics to achieve probabilistic k-anonymity based on data swapping are proposed: MDAV microaggregation on the quasi-identifiers plus swapping, and individual microaggregation on individual confidential attributes plus swapping. We report experimental results, where we compare the utility of original, k-anonymous and probabilistically k-anonymous data.

J. Soria-Comas and J. Domingo-Ferrer suggest T-Closeness is another extension of k-anonymity which also tries to solve the attribute disclosure problem. A data set is said to satisfy T-closeness if, for each group of records sharing a combination of quasi-identifier attribute values, the distance between the distribution of each confidential attribute in the group and the distribution of the same confidential attribute in the whole data set is no more than a threshold t. This property clearly solves the attribute disclosure vulnerability, although the original T-closeness paper did not propose a computational procedure to achieve this property and did not mention the large utility loss that this property is likely to inflict on the original data.

3. Existing System

Service provider computing is a new computing paradigm that is built on virtualization, parallel and distributed computing, utility computing, and service-oriented architecture. Although the great benefits brought by computing paradigm are exciting for IT companies, academic researchers, and potential users, security problems in service provider computing become serious obstacles which, without being appropriately addressed, will prevent service provider computing extensive applications and usage in the future. To achieve flexible and fine-grained access control, a number of schemes have been proposed more recently.

Unfortunately, these schemes are only applicable to systems in which data owners and the service providers are within the same trusted domain. Since data owners and service providers are usually not in the same trusted domain in service provider computing, a new access control scheme employing attributed encryption. The notion of attributed encryption was first introduced as a new method for fuzzy identity-based encryption. The primary drawback of the scheme in is that its threshold semantics lacks expressibility. Several efforts followed in the literature to try to solve the expressibility problem.

4. Proposed System

We have proposed and evaluated the use of micro aggregation as a method to attain k-anonymous t-closeness. The a priori benefits of microaggregation vs generalization/ recoding and local suppression have been discussed. Global recoding may recode more than needed, whereas local recoding complicates data analysis by mixing together values corresponding to different levels of generalization. Also, recoding produces a greater loss of granularity of the data, is more affected by outliers, and changes numerical values to ranges. Regarding local
suppression, it complicates data analysis with missing values and is not obvious to combine with recoding in order to decrease the amount of generalization. Microaggregation is free from all the above downsides.

We have 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 microaggregation algorithm. 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 micro aggregation. 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 (because the actual micro aggregation level is computed beforehand

4.1 Admin module

Initially, the micro aggregation algorithm is run on the quasi-identifier attributes of the original data set; this step produces a k-anonymous data set. Then, clusters of

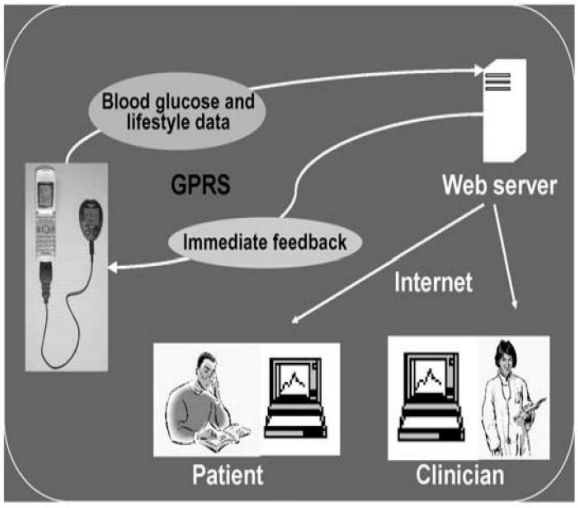
Micro aggregated records are merged until t-closeness is satisfied. Selecting the cluster whose confidential attribute distribution is most different from the confidential attribute distribution in the entire data set (that is, the cluster farthest from satisfying t-closeness); and ii) merging it with the cluster closest to it in terms of quasiidentifiers.

4.2 Search module

In this section that the values of the confidential attribute(s) can be ranked, that is, be ordered in some way. For numerical or categorical ordinal attributes, ranking is straight forward. Even for categorical nominal attributes, the ranking assumption is less restrictive than it appears, because the same distance metrics that are used to microaggregate this type of attributes can be used to rank them.

EMD distance with respect to microaggregation. To minimize EMD between the distributions of the confidential attribute within a cluster and in the entire data set, the values of the confidential attribute in the cluster must be as spread

as possible over the entire data set. Consider the case of a cluster with k records. The following proposition gives a lower bound of EMD for such a cluster.



Identify the Patient Situation and any Modification

4.3 Filtering Module

In a first battery of tests we used as evaluation data the Census data set, which is usual to test privacy protection methods and contains 1,080 records with numerical attributes. Because k-anonymity and t-closeness pursue different Goals. we defined two data sets according to the correlation between the values of quasi-identifier and confidential attributes.

5. Conclusion

This paper's contributions are threefold: 1) the introduction of new cost models and insights that explain and quantify the advantages of deploying trusted hardware for data processing, 2) the design and development of Trusted DB, a trusted hardware based relational database with full data confidentiality and no limitations on query expressiveness, and 3) detailed query optimization techniques in a trusted hardware-based query execution model. This work's inherent thesis is that, at scale, in outsourced contexts, computation inside secure hardware processors is orders of magnitude cheaper than equivalent cryptography performed on provider's unsecured server hardware, despite the overall greater acquisition cost of secure hardware. We thus propose to make trusted hardware a first-class citizen in the secure data management arena. Moreover, we hope that cost-centric insights and architectural paradigms will fundamentally change the way systems and algorithms are designed.

6. Future Scope

Classic work on database security has examined using a trusted filter in front of an untrusted relational database.The aim was to keep the performance advantage of a relational database system and use a minimal filter to enforce access control.

References

M. .M. Halldo´rsson and B. Chandra, N. Li, T. Li, and S. Venkatasubramanian. t-closeness: privacy beyond k-anonymity and l-diversity. In Proceedings of the 23rdIEEE International Conference on Data Engineering (ICDE 2007), pp.106–115. IEEE, **2007.**

N. Solanas, F. Seb´e, and J. Domingo-Ferrer. Micraggregationbasedheuristics for p-sensitive k-anonymity: one step beyond.In Proceedings of the 2008 International Workshop on Privacy and Anonymity in Information Society (PAIS 2008), pp. 61–69, New York,NY, USA**, 2008.** ACM.

J. Soria-Comas and J. Domingo-Ferrer. Probabilistic k-anonymity through microaggregation and data swapping. In Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2012), pp. 1–8. IEEE**, 2012.**

J. Soria-Comas and J. Domingo-Ferrer. Differential privacy via t-closeness in data publishing. In Proceedings of the 11th Annual International Conference on Privacy, Security and Trust (PST2013),pp. 27–35 2014

J. Soria-Comas, J. Domingo-Ferrer, D. S´anchez and S.Mart´ınez. Enhancing data utility in differential privacy viamicroaggregation-based k-anonymity. VLDB Journal 23(5):771–794**, 2014.**

J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogeneous. k-anonymity through microaggregation. Data Min.Knowl. Discov.,2005
C. Dwork. Differential privacy. In Proc. of the 33rd Intl. Colloquium on Automata, Languages and Programming (ICALP **2006)**,

Solanas, A. Mart´ınez-Ballest´e. V-MDAV: Variable group size multivariate microaggregation. In Proceeding of the International Conference on Computational Statistics **2006**

T. Sweeney. k-anonymity: a model for protecting privacy. InternationalJournal of Uncertainty, Fuzziness and Knowledge-Based Systems,10(5):557–570, **2002**.

Authors Detail:

Ms.S.Nathiya, Completed her BE CSE) in V.S.B. Engineering College under Anna University and now currently doing ME (CSE) in same institute. She had participated in several workshops and international conferences and presented papers.

Mr.U.Gowrisankar, Completed M.E (CSE) degree from Anna University, Now He is currently working as an Assistant Professor in Computer Science and Engineering in V.S.B Engineering College, Karur, Tamil Nadu, and India. His research interests include Data Mining, Networking. He had participated in several workshops and international conferences and published papers.

DOI: 10.18535/Ijecs/v4i12.13

Ms. P.Jayapriya, Completed her Diploma (IT) in Thiru Ramakrishna Nallamai Polytechnic college and also she had received BE degree (CSE) in Sasurie College of Engineering under Anna University and now currently doing ME (CSE) in V.S.B. Engineering College under Anna University and her research interests include Data Mining, Big Data analysis of distributed system.