Survey of Statistical Disclosure Control Technique

Jony N[1], AnithaM[2],

Department of Computer Science and Engineering [1, 2], *Adithya Institute of technology, Coimbatore*[1, 2]

Email: nesajony@gmail.com[1], anithash0911@gmail.com [2]

Abstract- In distributed scenarios the concept of double-phase micro aggregation has been proposed which is an improvement of classical micro aggregation. It is mainly used for protecting the privacy without fully trusted parties. The proposed architecture allows the remote collecting and sharing of biomedical data. The concerned private data should be kept back in secret and conveyed proficiently between numerous storing and grouped entities lacking any fail. The dual phase aggregation turns the essentials for privacy conservation in biomedical data in the scattered framework of transportable healthiness. Besides, the aggregation attains glowing and similarly to classical micro aggregation in terms of data loss, disclosure risk, and association safeguarding, while avoiding the limitations of a consolidated method.

Index Terms- Distributed environments, Information loss, Disclosure risk.

1. INTRODUCTION

Most organizations involve the handler to switch as of admittance mode to a broadcasting module when mining data which trust such mutual accountabilities as data collection and data superiority control should be done on the same page as the data entry. Example of how to aggregate data over days, months and years and enterprise with a few clicks. All facts may at any level be export to any database by clicking the [Copy all] button. Information may unpalatable route be combined above any managerial or environmental or other sight that the user chooses to operate at the same time as running with the data.

2. RECENT TRENDS

2.1 *Web – Healthiness*

The benefits of online health societies to accumulate, systems must be technologically advanced that are manageable, hospitable, easy to traverse and use, and able to help associates recognize information quality and interact with other participants in meaningful ways. The effective design of such schemes will be assisted by collaborations among clinicians, informed designers, and patients. Health professionals and patients can help explain the physical and emotional stages that entities go over after they are analyzed with a specific illness.

2.2 *Training*

Gathering and investigating data related to education archive will acquire, process, document, and disseminate data collected by administrations. Data files, documentation, and reports are downloadable from the website in public-use format. The website Features an online Files Analysis System (DAS) that allows users to conduct analyses on selected datasets within the Archive.

2.3 Database & data warehousing

Material stores are databases that are used for exposure and data investigation. Big data however, requires changed data warehouses than the straight pattern ones used in the past 10-20 years. There are numerous open source data warehouses accessible for dissimilar purposes..

2.4 Multidimensional database

A multidimensional database is optimized for data online analytical processing (OLAP) applications and for data warehousing. They are often created with the input from relational databases. Multidimensional databases can be charity for queries around business operations or trends. Multidimensional database organization systems (MDDBMS) can process the data in a database a high speed and can generate answers quickly.

2.4 Data aggregation

Data aggregation is the process of transforming scattered data from numerous sources into a single new one. The objective of data aggregation can be to combine sources together as such that the output is smaller than the input. This helps processing massive amounts of data in batch jobs and in real time applications. This reduces the network traffic and increases the performance while in progress.Data aggregation is any process in which information is expressed in a summary form for purposes such as reporting or analysis. Ineffective data aggregation is currently a major component that limits query performance. And, with up to 90 percent of all reports containing aggregate information, it becomes clear why proactively implementing an aggregation solution can generate significant performance benefits, opening up the opportunity for companies to enhance their organizations' analysis and reporting abilities. Aggregate Data in a Comprehensive, Out-of-the-Box Environment Informatics solution for B2B data exchange fully automate key steps of the data aggregation process, freeing your IT team to focus on your core competencies. Taking those steps further, the Informatics solution for B2B files aggregation increases efficiency, accelerates delivery times, and dramatically reduces costs with a broad range of fully integrated capabilities that include:

2.4.1 *Data collection*

To gather data from core and exterior sources using managed file transfer, this leverages secure communication protocols as S/FTP, AS1, AS2, HTTP/S, and PGP Data and format validation to confirm the integrity of data's structure and syntax.

2.4.2 Data transformation

To convert and translate from any external or internal file and message format to a canonical format (i.e., XML).

2.4.3 Data normalization

To cleanse and match data and handle all exceptions to ensure high-quality data.

2.4.4 Data enrichment

To access additional sources and systems to extract and append additional information necessary to create a complete data set.

2.4.5 Data mapping

To plot the format and structure of data between its source and target systems according to certain transformation rules and business logic.

2.4.6 Data extraction

To select and mine relevant data using specific parameters.

The rest of this paper is organized as follows: Section II introduces the background of SDC. Section III then describes the proposed algorithm which is a two-phase method. Finally, Section 4 presents conclusions.

3 LITERATURE SURVEY

Micro aggregation is a family of methods for statistical disclosure control (SDC) of micro data (records on individuals and/or companies), that is, for masking micro data so that they can be released while preserving the privacy of the underlying individuals. The principle of micro aggregation is to aggregate original database records into small groups prior to publication .Each group should contain at least *k* records to prevent disclosure of individual information, where *k* is a constant value preset by the data protector. Recently,Micro aggregation has been shown to be useful to achieve *k*-anonymity, in addition to it being a good masking method. Optimal micro aggregation (with minimum within-groups variability loss) can be computed in polynomial time for univariate data. Unfortunately, for multivariate data it is an NP-hard problem.

Several heuristic approaches to micro aggregation have been proposed in the literature. Heuristics yielding groups with fixed size *k* tends to be more efficient whereas data oriented heuristics yielding variable group size tends to result in lower information loss. This paper presents new data-oriented heuristics which improve on the trade-off between computational complexity and information loss and are thus usable for large datasets.

A class of perturbative SDC methods for micro data. Given an original set of micro data whose respondents (i.e., contributors) must have their privacy preserved, micro aggregation yields a protected data set consisting of aggregate information(e.g., mean values) computed on small groups of records in the original dataset. Since this protected dataset contains only aggregate data, its release is less likely to violate respondent privacy. For the released dataset to stay analytically useful, the information loss caused by micro aggregation must be minimized: a way to approach this minimization is for records within each group to be as homogeneous as possible.

Multivariate micro aggregation (for several attributes) with maximum within groups record homogeneity is NP-hard, so heuristics are normally used. There is a dichotomy between fixed-size heuristics yielding groups with a fixed number of records and data-oriented heuristics yielding groups whose size varies depending on the distribution of the original records. Even if the latter heuristics can in principle achieve lower information loss than fixed-size micro aggregation, they are often dismissed for large datasets due to complexity reasons. For example, the μ-Argus SDC package only features fixed-size micro aggregation. Our contribution in this paper is an approach to turn some fixed-size heuristics for multivariate micro aggregation of numerical data into data-oriented heuristics with little additional computation.The resulting new heuristics improves the trade-off between information loss and computational complexity.

One approach to facilitate health research and alleviate some of the problems documented above is to de-identify data beforehand or at the earliest opportunity. Many research ethics boards will waive the consent requirement if the data collected or disclosed is deemed to be de-identified. A commonly used de-identification criterion is k-anonymity, and many k-anonymity algorithms have been developed. This criterion stipulates that each record in a dataset is similar to at least another k-1 records on the potentially identifying variables. For example, if k-5 and the potentially identifying variables are age and gender, then a k-anonym zed dataset has at least 5 records for each value combination of age and gender.

A new k-anonymity algorithm, Optimal Lattice Anonymization (OLA), which produces a globally optimal de-identification solution suitable for health datasets. We demonstrate on six datasets that OLA results in less information loss and has faster performance compared to current de-identification algorithms.

o Quasi-identifiers

The variables that are going to be de-identified in a dataset are called the quasi-identifiers. Examples of common quasi-identifiers are dates (such as birth, death, admission,
Discharge, visit, and specimen collection), locations (such as postal codes, hospital names, and regions), race, ethnicity, languages spoken, aboriginal status, and gender.

o Equivalence Classes

All the records that have the same values on the quasi-identifiers are called an equivalence class. For example, all the records in a dataset about 17-year-old males admitted on Jan 1, 2008 are an equivalence class. Equivalence class sizes potentially change during de-identification. For example, there may be 3 records for 17-year-old
Males admitted on Jan 1, 2008. When the age is recoded to a five year interval, then there may be 8 records for males between 16 and 20 years old admitted on Jan 1, 2008.

o De-identification Optimality Criterion

A de-identification algorithm balances the probability of re-identification with the amount of distortion to the data (the information loss). For all k-anonymity algorithms, disclosure risk is defined by the k value, which stipulates a maximum probability of re-identification.There is no generally accepted information loss metrics.

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 anonym zed versions of confidential micro data. Optimally solving micro aggregation on multivariate data sets is known to be difficult. Therefore, heuristic methods are used in practice. This paper presents a new heuristic approach to multivariate micro aggregation, which provides variable-sized groups (and thus higher within-group homogeneity) with a computational cost similar to the one of fixed-size micro aggregation heuristics.

Micro aggregation is a problem appearing in statistical disclosure control (SDC), where it is used to cluster a set of records in groups of at least k records, with k being a user-definable parameter. The collection of groups is called a k-partition of the data set. The micro aggregated data set is built by replacing each original record by the centroid of the group it belongs to. The micro aggregated data set can be released without jeopardizing the privacy of the individuals which form the original data set: records within a group are indistinguishable in the released data set. The higher the within-group homogeneity in the original data set, the lower the information loss incurred when replacing records in a group by their centric therefore within-group homogeneity is inversely related to information loss caused by micro aggregation.

A simple and efficient implementation of Lloyd's k-means clustering algorithm, which we call the filtering algorithm. This algorithm is easy to implement, requiring a kd-tree as the only major data structure. We establish the practical efficiency of the filtering algorithm in two ways. First, we present a data-sensitive analysis of the algorithm's running time, which shows that the algorithm runs faster as the separation between clusters increases. Second, we present a number of empirical studies both on synthetically generated data and on real data sets from applications in color quantization, data compression, and image segmentation.

Clustering based on k-means is closely related to a number of other clustering and location problems. These include the Euclidean k-medians in which the objective is to minimize the sum of distances to the nearest center and the geometric k-center problem in which the objective is to minimize the maximum distance from every point to its closest centre. There are no efficient solutions known to any of these problems and some formulations are NP-hard. An asymptotically efficient approximation for the k-means Clustering problem has been presented by Matousek but the large constant factors suggest that it is not a good candidate for practical implementation. One of the most popular heuristics for solving the k-means problem is based on a simple iterative scheme for finding a locally minimal solution. This algorithm is often called the k-means algorithm.

An encryption method is presented with the novel property that publicly revealing An encryption key does not thereby reveal the corresponding decryption Key. This method provides an implementation of a public-key cryptosystem an elegant concept invented by Diffie and Hellman. This has two important consequences:

1. Couriers or other secure means are not needed to transmit keys, since a message can be enciphered using an encryption key publicly revealed by the intended recipient. Only he can decipher the message, since only he knows the corresponding decryption key.

2. A message can be signed using a privately held decryption key. Anyone can verify this signature using the corresponding publicly revealed encryption

key. Signatures cannot be forged, and a signer cannot later deny the validity of his signature. This has obvious applications in electronic mail and electronic funds transfer systems. a public-key cryptosystem can ensure privacy and enable signatures.

All classical encryption methods suffer from the key distribution problem.The problem is that before a private communication can begin, another private transaction is necessary to distribute corresponding encryption and decryption keys to the sender and receiver, respectively. Typically a private courier is used to carry a key from the sender to the receiver. Such a practice is not feasible if an electronic mail system is to be rapid and inexpensive. A public-key cryptosystem needs no private couriers; the keys can be distributed over the insecure communications channel.

Micro aggregation is a Statistical Disclosure Control (SDC) technique that aims at protecting the privacy of individual respondents before their data are released .Optimally micro aggregating multivariate data sets is known to be an NP-hard problem. Thus, using heuristics has been suggested as a possible strategy to tackle it. Specifically, Genetic Algorithms have been shown to be serious candidates that can find good solutions on small data sets. However, due to the very nature of these algorithms and the coding of the micro aggregation problem, GA can hardly cope with large data sets. In order to apply them to large data sets, the latter have to be previously partitioned into smaller disjoint subsets that
the GA can handle.

With the aim to protect from re-identification of individual respondents, micro data sets are properly modified prior to their publication. The degree of modification can vary between two extremes:

(i) Encrypting the micro data.
(ii) leaving the micro data intact.

In the first extreme, the protection is perfect however, the utility of the data is almost nonexistent because the encrypted micro data can be hardly studied or analysed. In the other extreme, the micro data are extremely useful (i.e. all their information remains intact), however, the privacy of the respondents is endangered. SDC methods for micro data protection aim at distorting the original data set to protect respondents from re-identification whilst maintaining as much as possible, some of the statistical properties of the data and minimising the information loss. The goal is to find the right balance between data utility and respondents privacy micro data sets are organised in records that refer to individual respondents. Each record has several attributes in a micro data set X can be classified in three categories as follows:

1) Identifiers:
 Attributes in X that unambiguously identify the respondent. For example, passport numbers, full names, etc. the attribute "social security number" is an identifier.
2) Key attributes:
 If properly combined can be linked with external information sources to re-identify some of the respondents to whom some of the records refer. For example, address, age, gender, etc.
3) Confidential outcome attributes:
 It containing sensitive information on the respondent, namely salary, religion, political affiliation, health condition, etc.

A formal protection model named k-anonymity and a set of accompanying policies for deployment. A release provides k-anonymity protection if the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appears in the release. This paper also examines re-identification attacks that can be realized on releases that adhere to k anonymity unless accompanying policies are respected. The k-anonymity protection model is important because it forms the basis on which the real-world systems known as Data fly, m-Argus and k-Similar provide guarantees of privacy protection.

Micro aggregation is a technique used by statistical agencies to limit disclosure of sensitive micro data. Noting that no polynomial algorithms are known to micro aggregate optimally, Domingo-Ferrier and Mateo-Saenz have presented heuristic micro aggregation methods. This paper is the first to present an efficient polynomial algorithm for optimal univariate micro aggregation. Optimal partitions are shown to correspond to shortest paths in a network .It is at least as efficient as published heuristic methods and can be used on very large data sets. While our algorithm focuses on uni variate data, it can be used on multivariate data when the data vectors are projected to a single Axis.

It formulate the micro aggregation problem as a shortest-path problem which construct a graph and show that optimal micro aggregation corresponds to a shortest path in this graph in a natural way. Each arc of the graph corresponds to a possible group that may be part of an optimal partition. Each arc is labelled by the error that would result if that group were to be included in the partition.

A clustering algorithm for partitioning a minimum spanning tree with a constraint on minimum group size. The problem is motivated by micro aggregation, a disclosure limitation technique in which similar records are aggregated into groups containing a minimum of k records. Heuristic clustering methods are needed since the minimum information loss micro aggregation problem is NP-hard. Our MST partitioning algorithm for micro aggregation is sufficiently efficient to be practical for large data sets and yields results that are comparable to the best available heuristic methods for micro aggregation. For data that contain pronounced clustering effects, results in significantly lower information loss. Algorithm is general enough to accommodate different measures of information loss and can be used for other clustering applications that have a constraint on minimum group size.

To protect the anonymity of the entities (called respondents) to which information refers, data holders often remove or encrypt explicit identifiers such as names, addresses, and phone numbers. De-identifying data, however, provides no guarantee of anonymity .Released information often contains other data, such as race, birth date, sex, and ZIP code that can be linked to publicly available information to re identify respondents and inferring information that was not intended for disclosure. In this paper we address the problem of releasing micro data while safeguarding the anonymity of the respondents to which the data refer. The approach is based on the definition of k-anonymity.

How k-anonymity can be provided without compromising the integrity (or truthfulness) of the information released by using generalization and suppression techniques. We introduce the concept of minimal generalization that Captures the property of the release process not to distort the data more than needed to achieve k-anonymity, and present an algorithm for the computation of such a generalization. We also discuss possible preference policies to choose among different minimal generalizations.

4 CONCLUSION

An architecture that allows the private gathering and sharing of biomedical data in the context of m-health .We have introduced to concept of double-phase micro aggregation to limit the information accessible by intermediate entities (such as the SAS). It preserves the correlations of the original data set. Then, we can conclude that the distributed double-phase micro aggregation proposed can be applied in a distributed environment to protect the privacy of individuals with the same effects of classical micro aggregation . Further research might include the analysis of the influence of time in the series of data collected using our model.

REFERENCES

[1] Brand R. (2002) ‘Micro data protection through noise addition’, Lecture Notes in Computer Sci., vol. 2316, pp. 97–116.

[2] C. D. Brown (2000) ‘Body mass index and prevalence of hypertension and dyslipidemia ‘,Obesity Res., vol. 8, no. 9, pp. 605–619.

[3] T. Dalenius and S. P. Reiss (1982) ‘Data-swapping: A technique for disclosure control’, Statistical Planning and Inference, vol. 6, no. 1, pp. 73–85.

[4] J. Domingo-Ferrer, F. Sebé, and J. Castellà,(2004) ‘On the security of noise addition for privacy in statistical databases’, Lecture Notes in Computer Sci., vol. 3050, pp. 149–161.

[5] J. Domingo-Ferrer (2006) ‘Efficient multi variate data-oriented micro aggregation’, Int. J. Very Large Databases, vol. 15, no. 4, pp. 355–369.

[6] Domingo-Ferrer, J., Mateo-Sanz, J.M (2002) ‘Practical data-oriented micro aggregation for statistical disclosure control’, IEEE Trans. Knowl. Data Eng. 14(1), 189–201.

[7] K. Emam (2009) ’Globally optimal k-anonymity for de-identification of health data’, J. Amer. Med. Inform. Assoc., vol. 16, no. 5, pp. 670–682.

[8] T. ElGamal, (1985) ‘A public-key cryptosystem and a signature scheme based on discrete logarithms’, IEEE Trans. Inf. Theory, vol. 31, no. 4, pp. 469–472, Jul.

[9] B. Greenberg (1987) ‘Rank Swapping for Masking Ordinal Micro data’, Tech. report. U.S. Bureau of the Census , unpublished

[10] Hansen, S.L., Mukherjee, S (2003) ‘A polynomial algorithm for optimal univariate micro aggregation’, IEEE Trans. Knowl. Data Eng. 15(4), 1043–1044.

[11] M. Naehrig, K. Lauter, and V. Vaikuntanathan, ’Can homomorphic encryption be practical?’, in Proc. 3rd ACM Workshop on Cloud Computing Security Workshop (CCSW’11), New York, NY, USA, pp. 113–124.

[12] G. J. Matthews and O. Harel (2011) ‘Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy’, Statist. Surveys, vol. 5, pp. 1–29.

[13] D. Pagliuca and G. Seri (1998) ‘Some Results of Individual Ranking Method on the System of Enterprise Accounts Annual Survey’, Esprit SDC Project, Deliverable MI3/D2.

[14] R. Rivest, A. Shamir, and L. Adleman (1978) ‘A method for obtaining digital signatures and public-key cryptosystems’, Commun. ACM, vol. 21, no. 2, pp. 120–126.

[15] P. Samarati,(2001) ‘Protecting respondents’ identities in micro data release’, IEEE Trans. Knowl. Data Eng., vol. 13, no. 6, pp. 1010–1027, Nov.Dec.

[16] Solanas and A. Martínez-Ballesté (2006) V-MDAV: Variable group size multivariate micro aggregation,” in Proc. COMPSTAT, pp.917–925.

[17] Solanas, A. Martínez-Ballesté, and Ú. González-Nicolas (2010)’A variable-MDAV-based partitioning strategy to continuous multivariate micro aggregation with genetic algorithms’, in Proc. Int. Joint Conf.Neural Networks (IJCNN), pp. 1–7.

[18] L. Sweeney (2002) ‘k-anonymity: A model for protecting privacy’, Int. J.Uncertainty, Fuzziness and Knowledge-Based Syst., vol. 10, no. 5, pp.557–570.

[19] L. Willenborg and T. DeWaal (1996)‘Statistical Disclosure Control in Practice’,in Lecture Notes in Statistics. New York, NY, USA: Springer-Verlag, vol. 111.

[20] L.Willenborg and T. DeWaal (2001) ‘Elements of Statistical Disclosure Control’, in Lecture Notes in Statistics. New York, NY, USA: Springer-Verlag, vol. 155.

[21] J. Domingo-Ferrer and J.M. Mateo-Sanz, “Practical Data-Oriented Microaggregation for Statistical Disclosure Control,” IEEE Trans. Knowledge and Data Eng., vol. 14, no. 1, pp. 189-201, Jan./Feb. 2002.

[22] J. Domingo-Ferrer and V. Torra, “A Quantitative Comparison of Disclosure Control Methods for Microdata,” Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies, P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, eds., pp. 111-133, Amsterdam: North-Holland, 2001.

[23] S.L. Hansen and S. Mukherjee, “A Polynomial Algorithm for Optimal Univariate Microaggregation,” IEEE Trans. Knowledge and Data Eng., vol. 15, no. 4, pp. 1043-1044, July/Aug. 2003.

[24] A.K. Jain, M.N. Murty, and P.J. Flynn, “Data Clustering: A Review,” ACM Computing Surveys, vol. 31, no. 3, 1999.