MICROAGGREGATION FOR THE PROTECTION OF MICRODATA

G.Ramya (ME), A.Selvaraj (ME), N.Anandh (ME)

Department of Computer Science and Engineering,
Muthayammal Engineering College, INDIA
ramya14@gmail.com, engineerselvaraj@yahoo.com, anandhme1983@gmail.com

Abstract- Microaggregation is a family of methods for statistical disclosure control (SDC) of microdata (records on individuals and/or companies), that is, for masking microdata so that they can be released while preserving the privacy of the underlying individuals. Microaggregation is a technique that protects the privacy of individuals by aggregating similar microdata records and producing microaggregated data sets satisfying the property of k-anonymity. Microaggregation is used for the protection of privacy in distributed scenarios without fully trusted parties. Microaggregation properly fits the needs for privacy preservation of individuals. 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. The new data-oriented heuristics which improve on the trade-off between computational complexity and information loss and are thus usable for large data sets.

Keywords: **Microaggregation, Statistical disclosure content, k-anonymity, Privacy**

1 INTRODUCTION

Due to the recent advances in information and communication technologies the gathering, storage and sharing of data are becoming simpler and faster than ever. Protecting individual privacy is paramount for many institutions, namely statistical agencies, healthcare centers, Internet companies, manufacturers, etc. Many efforts have been devoted to develop techniques that assure a given degree of privacy to the people, whose data are collected and shared. Currently, these efforts come from very diverse fields of knowledge, namely cryptography, statistics, artificial intelligence, etc. Among all these disciplines, statistical disclosure control (SDC) was the first to consider the problem; initially on tabular data, and later on microdata (i.e., data from individuals).

It mainly focuses on the protection of individual privacy by means of protecting their microdata. By doing so, we aim at avoiding the re-identification of individuals through their released microdata. To achieve this goal microdata sets have to the be properly modified prior to their publication. The degree of modification varies between two extremes: (i) encrypting the microdata and, (ii) leaving the microdata intact.

In the first extreme, the protection is perfect (i.e., only the owner of a decryption key can see the data), however, the utility of the data is almost nonexistent because the encrypted microdata can be hardly studied or analyzed by others than the owner of the decryption key. Although there exist fully homomorphic encryption schemes that allow computations on encrypted data, they are still not fully practical and require a significant computation

effort, that is not affordable to most services yet. In the other extreme, the microdata are totally useful (i.e., all their information remains intact), however, the privacy of the individuals is endangered because sensitive private data can be seen without limitation. SDC methods aim at distorting the original microdata sets to protect individuals privacy and avoid their re-identification while maintaining some of the statistical properties of the data and minimizing the information loss as much as possible. The challenge in SDC is to tune modification so that both privacy and information loss are acceptable: both the risk of disclosing private confidential information and the loss of data utility should be kept below reasonable thresholds preset by the data protector.

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.

A. Basics of Microaggregation

Microaggregation is a family of perturbative SDC methods originally designed for continuous numerical data and recently extended for categorical data. The microaggregation can be operationally defined in terms of two steps:

Partition The set of original records is partitioned into several groups in such a way that records in the same group are *similar* to each other and so that the number of records in each group is at least k. A partition meeting this requirement on minimal group size is called a k-partition.

Aggregation An aggregation operator (for example, the mean for numerical data) is used to compute a centroid for each group. Then, each record in a group is replaced by the group centroid.

There exist two main types of heuristics:

• *Fixed-size microaggregation* These heuristics yield k-partitions where all groups have size k, except perhaps one group which has size between k and $2k$-1

• *Data-oriented microaggregation* These heuristics yield k-partitions where all groups have sizes varying between k and $2k-1$. The adaptation of standard clustering techniques (e.g., k-means) for microaggregation is not trivial: the challenge is how to enforce cardinality constraints on groups without substantially increasing sum of squares.

B. Basic Definitions and Concepts

Some basic definitions related to the field of statistical disclosure control that are used through the paper:

Microdata In opposition to macrodata, that refers to large aggregates of information generally represented in tables, microdata refers to individual data such as the social security number (SSN), age, ethnicity, height, income, etc. that are represented with records.

Microdata set. A microdata set is the union of microdata records sharing the same attributes. Thus, a microdata set is understood as a two-dimensional matrix in which rows represent individual data and columns represent specific attributes.

Table 1

The dataset. "Company name" is an identifier to be suppressed before publishing the dataset

Company name	Surface (m^2)	No. employees	Turnover (Euros)	Net profit (Euros)
A&A Ltd	790	55	3,212,334	313,250
B&B SpA	710	44	2,283,340	299,876
C&C Inc	730	32	1,989,233	200,213
D&D BV	810	17	984,983	143,211
E&E SL	950	3	194,232	51,233
F&F GmbH	510	25	119,332	20,333
G&G AG	400	45	3,012,444	501,233
H&H SA	330	50	4,233,312	777,882
I&I LLC	510	5	159,999	60,388
J&J Co	760	52	5,333,442	1,001,233
K&K Sarl	50	12	645,223	333,010

Identifiers. Those attributes in a microdata set that point out to a unique individual are called identifiers. In our ex- ample the social security number (SSN) is an identifier. (This attributes are deleted before releasing microdata for public use).

Quasi-identifiers. Those attributes containing information about an individual that, when taken individually, do not identify him/her. The combination of quasi-identifiers might lead to the identification of a unique individual. Examples of this kind of attribute could be zip code, weight and height (e.g., a very tall person in a small village could be easily identified).

Confidential outcome attributes. Those attributes that have sensitive information like religion, salary, health condition, etc.

Non-confidential outcome attributes Those are attributes which contain non-sensitive information on the respondent. The attributes of this kind cannot be neglected when protecting a dataset, because they can also be key attributes. For instance, job and town of residence may reasonably be considered non-confidential outcome attributes, but their combination can be a quasi-identifier because everyone knows who is the doctor in a small village.

Table 2

The anonymous version of the dataset after optimal microaggregation of key attributes

Surface (m^2)	No. employees	Turnover (Euros)	Net profit (Euros)
747.5	46	3,212,334	313,250
747.5	46	2,283,340	299,876
747.5	46	1,989,233	200,213
756.67	8	984,983	143,211
756.67	8	194,232	51,233
322.5	33	119,332	20,333
322.5	33	3,012,444	501,233
322.5	33	4,233,312	777,882
756.67	8	159,999	60,388
747.5	46	5,333,442	1,001,233
322.5	33	645,223	333,010

II. PROPOSED SYSTEM

A multivariate dataset consisting of n records and p numerical attributes can be represented as n points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in R^p. The new data-oriented heuristics proposed to operate in three steps:

1. Find a path T traversing all n points of the dataset. Let the π_T be the permutation of $\{1, \ldots, n\}$ expressing the order in which the points are traversed by T.

2. Use the ordering π_T to feed the p-variate records to a multivariate version of the Hansen–Mukherjee algorithm, called MHM algorithm. Formally, pass the ordered tuple $(\mathbf{x}_{\pi_T(1)}, \ldots, \mathbf{x}_{\pi_T(n)})$ to MHM, which outputs a data-oriented k-partition for that ordered dataset.

3. Microaggregate $(\mathbf{x}_{\pi_T(1)}, \ldots, \mathbf{x}_{\pi_T(n)})$ using the k-partition output by MHM

The MHM algorithm

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be an ordered dataset with n records where each record $\mathbf{x}_i$ contains the values of p

attributes. Let k be an integer group size such that $1 \leq k < n$. Then, a graph $\mathbf{G}_{n,k}$ is constructed as follows:

1. For each value $\mathbf{x}_i$ in $\mathbf{X}$, create a node with label i. Create also an additional node with label 0.

2. For each pair of graph nodes (i, j) such that $i + k \leq j < i + 2k$, create a directed arc (i, j) from node i to node j.

3. Map each arc (i, j) to the group of values $C_{(i,j)} = \{\mathbf{x}_h : i < h \leq j\}$. Let the length $L_{(i,j)}$ of the arc be the within group sum of squares for $C_{(i,j)}$, that is,

$$L_{(i,j)} = \sum_{h=i+1}^{j} (\mathbf{x}_h - \bar{\mathbf{x}}_{(i,j)})'(\mathbf{x}_h - \bar{\mathbf{x}}_{(i,j)})$$

where $\mathbf{x}_{(i,j)}$ is a p-dimensional record computed as the centroid (average) of records in $C_{(i,j)}$.

A. Nearest point next (NPN)

The path is constructed as follows:
1.Compute the centroid (average record) **x** of all points in the dataset.

2. Compute the most distant point **r** from **x** and take **r** as the first point in the path.

3. The second point is the closest one to the first (among the remaining points), the third point is the closest one to the second (among the remaining points different from the first and the second) and so on until all n points have been added to the path.

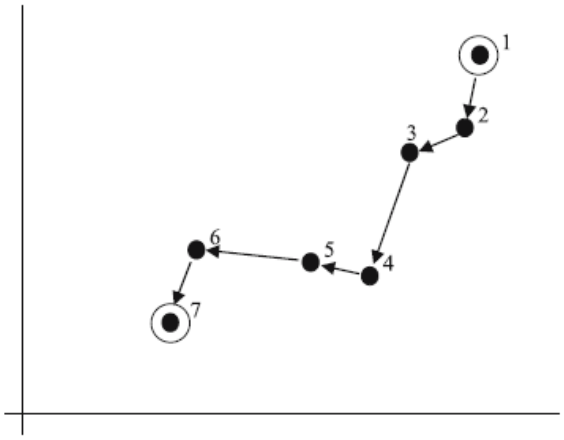

Fig. 1 Path construction on bivariate microdata using NPN

III RESULTS

MHM runs in $O(k2n)$ time, the complexity of the heuristics is dominated by the path construction. The order of magnitude of their complexity is the same as the one of the most efficient fixed-size microaggregation heuristics in the literature. If n is very large, a good strategy is to use blocking attributes to split the n records into several smaller sub data sets which can be microaggregated independently. One or several categorical attributes (like province, activity sector, etc.) are used as blocking attributes, and the typical block sizes are a few thousands of records.

A. Empirical comparative analysis

MD–MHM, MDAV–MHM and CBFS–MHM can dramatically reduce information loss (measured as within-groups sum of squares *SSE*) with respect to the fixed-size microaggregation heuristics Maximum Distance, MDAV and CBFS, respectively.

Finally, we give some experimental results on how close to optimality are the partitions obtained using the new heuristics and their fixed-size counterparts.

Table 3

Complexities of the combinations of the four path constructions of MHM

Heuristic	Complexity
CBF–MHM	$O(\frac{n^2}{2})$
MD–MHM	$O(\frac{n^3}{12k})$
MDAV–MHM	$O(\frac{n^2}{2k})$
CBFS–MHM	$O(\frac{n^2}{2k})$

IV CONCLUSION

Microaggregation is an SDC technique used world-wide to preserve the privacy of respondents contributing to microdata sets. The level of privacy required is controlled by a parameter *k* (minimum group size). Once *k* has been chosen, the data protector (and the data users) is interested in minimizing information loss.

Hansen–Mukherjee's algorithm for optimal univariate microaggregation can be used to enhance the heuristics of multivariate microaggregation, so as to reduce information loss.

For very homogeneous datasets (without obvious clusters), the MHM-enhanced heuristics display a performance as good as their fixed-sized counterparts regarding information loss and nearly as good regarding speed. For mildly skewed or clustered datasets, MDAV– MHM and CBFS–MHM do pretty well in exploiting their data-orientedness to reduce information loss with respect to fixed-size alternatives. For heavily skewed or clustered datasets, NPN–MHM is the best heuristic method.

REFERENCES

[1] Agusti Solanas and Antoni Mart´ınez-Ballest´, "V-MDAV: A Variable Group Size with multivariate microaggregation" Rovira i Virgili University. Av.Pa¨ısos Catalans 26. 43007 Tarragona. Catalonia

[2] Chris Skinner, "Statistical Disclosure Control for Survey Data" Univ. of Southampton

[3] J. Domingo-Ferrer et al., "Statistical Disclosure Risk & Statistical Disclosure Control"

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

[5] Josep Domingo-Ferrer,Antoni Martínez-Ballesté, Josep Maria Mateo-Sanz,Francesc Sebé, "Multivariate microaggregation with variable group size"

[6] Latanya Sweeney, k-Anonymity: "A Model for protecting privacy", School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

[7] 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, pp113–124

[8] Matthias Templ, "Data Access and Personal Privacy: Appropriate Methods of Disclosure Control", Department of Methodology, Statistics Austria, Guglgasse 13, 1110 Vienna, Austria

[9] Oleg Chertov and Anastasiya Pilipyuk, "Statistical Disclosure Control Methods for Microdata", National Technical University "Kyiv Polytechnic Institute", Kyiv, Ukraine

[10] Defays, D., Anwar, N.: Micro-aggregation: a generic method.In: Proceedings of the 2nd International Symposium on Statistical Confidentiality, pp. 69–78. Eurostat, Luxemburg (1995)

[11] Hansen, S.L., Mukherjee, S.:Apolynomial algorithm for optimal univariate microaggregation. IEEE Trans. Knowl. Data Eng. **15**(4), 1043–1044 (2003)

[12] Mateo-Sanz, J.M., Domingo-Ferrer, J.: Heuristic techniques for multivariate microaggregation. In: COMPSTAT'2000, Utrecht. CBS-Statistics, Netherlands (2000)

