Microdata Protection Method Through Microaggregation: A Median Based Approach

Md Enamul Kabir and Hua Wang

Department of Mathematics and Computing

University of Southern Queensland

Toowoomba, QLD 4350, Australia

Emails: {kabir, wang}@usq.edu.au

ABSTRACT

Microaggregation for Statistical Disclosure Control (SDC) is a family of methods to protect microdata from individual identification. SDC seeks to protect microdata in such a way that can be published and mined without providing any private information that can be linked to specific individuals. The aim of SDC is to modify the original microdata in such a way that the modified data and the original data are similar. Microaggregation works by partitioning the microdata into groups, also called clusters of at least k records and then replacing the records in each group with the centroid of the group. In this work we introduce a new microaggregation method, where the centroid is considered as median. The new method guarantees that the microaggregated data and the original data are similar by using statistical test. Another contribution of this work is that we propose a distance metric, called absolute deviation from median (ADM) to evaluate the amount of mutual information among records in microdata. We showed that ADM is always less than the most commonly used measure of distortion called sum of squares of errors (SSE) for any dataset. Thus

ADM causes least information loss and can be used as a measure of information loss for a microaggregated microdata set.

Keywords: Privacy; Microaggregation; Microdata protection; *k*-anonymity; Disclosure control;

1. INTRODUCTION

In recent years, the phenomenal advance technological developments in information technology enable government agencies and corporations to accumulate an enormous amount of personal data for analytical purposes. These agencies and organizations often need to release individual records (microdata) for research and other public benefit purposes. This propagation has to be in accordance with laws and regulations to avoid the propagation of confidential information. In other words, microdata should be published in such a way that preserves the privacy of the individuals. To protect personal data from individual identification, SDC is often applied before the data are released for analysis (Domingo-Ferrer & Torra, 2005, Willenborg & Waal, 2001). The purpose of microdata SDC is to alter the original microdata in such a way that the statistical analysis from the original data and the modified data are similar and the disclosure risk of identification is low. As SDC requires to suppress or alter the original data, the quality of data and the analysis results can be damaged. Hence, SDC methods must find a balance between data utility and personal confidentiality.

Microaggregation is a family of SDC methods for protecting microdata sets that have been extensively studied recently (Domingo-Ferrer & Mateo-Sanz, 2002, Domingo-Ferrer & Torra, 2002, 2003 & 2005, Han, Cen, Yu, & Yu, 2008, Hansen &

Mukherjee, 2003). The basic idea of microaggregation is to partition a dataset into mutually exclusive groups (called clusters) of at least *k* records prior to publication, and then publish the centroid over each group instead of individual records. The resulting anonymized dataset satisfies *k*-anonymity (Sweeny, 2002), requiring each record in a dataset to be identical to at least (*k*-1) other records in the same dataset. As releasing microdata about individuals poses a privacy threat due to the privacy-related attributes, called quasi-identifiers, both *k*-anonymity and microaggregation only consider the quasi-identifiers. Microaggregation is traditionally restricted to numeric attributes in order to calculate the centroid of records, but also been extended to handle categorical and ordinal attributes (Domingo-Ferrer & Torra, 2002 & 2005, Torra, 2004). In this paper we proposed a microaggregated method that also applicable for the numeric attributes.

2. MOTIVATION

As stated before, the rationale behind microaggregation is to divide the dataset into some groups, where each group contains at least *k* records. For each variable, the average value over each group is computed and is used to replace each of the original averaged values. Groups are formed using a criterion of maximum similarity. Once the procedure has been completed, the resulting dataset can be published. Now a natural question arise, what centroid value should be used instead of individual records in each group as the common centre values that describe a set of values are mean, median and mode. The simplest answer is to use that value which apparently grantee that the modified data and the original data are similar by using statistical test. Previously mean was used as a centroid value but it does not grantee the similar modified data and the original data. In this work we used median as the centre value

as it shows by sign test that the modification has no effect and produces similar modified and original data. There are also some advantages of using median as the centre values. Firstly, median is the appropriate measure for skewed distribution. If the records in each group follow skewed distribution, median should be used as the measure of central tendency. Mean is the appropriate measure for symmetric distribution, however for symmetric distribution mean and median are equal and thus there is no difference of using mean or median as the centre value. Secondly, mean is affected by extreme values, which mean if a group contains any extreme values the total information loss will be increased. But median is not at all affected by extreme values. Fox example, if a dataset consists of five values, 24, 21, 28, 25 and 95, then the mean of these values is 38.6 (affected by extreme value 95) and the median of these values is 25, not affected by extreme value 95. Lastly it is computationally more convenient of using median to measure the distortion of the original data. The distortion is measured as the difference between the original values and the modified values, but sum of these differences is zero if mean is used as the centre value. Thus sum of squares of differences is normally used to measure the distortion, if mean is used as a centre value that is computationally difficult. But this sum of differences is not zero if median is used as a canter value. Thus the sum of absolute differences can be used to measure the distortion that is computationally less difficult. Thus in this paper we proposed a median based microaggregation method for SDC. Using median as the centre value produces similar original but not the same dataset, so there are still chances of being loss of information. Thus the effectiveness of a microaggregation method is measured by calculating its information loss. A lower information loss implies that the anonymized dataset is less distorted from the original dataset, and thus provides better data quality for analysis.

3. RELATED WORK

This work is related to several topics in the area of microaggregation in SDC. Usually, in microaggregation methods mean is used as a centriod and different authors' proposed different methods in order to minimize the information loss.

k- anonymity (Samarati, 2001 & Sweeny, 2002) provides sufficient protection of personal confidentiality of microdata, while to ensure the quality of the anonymized dataset, an effective microaggregation method should incur information loss as minimum as possible. It is a natural expectation that information loss can be reduced by placing similar records in the same groups. In data mining environment, clustering is an effective method of grouping similar records together and many microaggregation methods derive from traditional clustering algorithms. For instance, Domingo-Ferrer and Mateo-Sanz (2002) proposed univariate and multivariate *k*-Warld algorithms that extend the agglomerative hierarchical clustering method of Ward et al. (1963), Domingo-Ferrer and Torra (2002 & 2003) proposed a microaggregation method based on the fuzzy *c*-means algorithm (Bezdek, 1981), and Laszlo and Mukherjee (2005) extended the standard minimum spanning tree partitioning algorithm for microaggregation (Zahn, 1971). All of these microaggregation methods build all groups gradually but simultaneously. There are some other methods for microaggregation that have been proposed in the literature that build one cluster at a time. Notable examples include Maximum Distance (Solanas, 2008), Diameter-based Fixed-Size microaggregation and centroid based Fixed-size microaggregation (Laszlo & Mukherjee, 2005), MDAV (Domingo-Ferrer & Mateo-Sanz, 2002, Domingo-Ferrer & Torra, 2005), MHM (Domingo-Ferrer et al.,

2006) and the Two Fixed Reference Points method (Chang et al., 2007). Most recently, Lin et al. (2010) proposed a density-based microaggregation method that forms records by the descending order of their densities, and then fine-tunes these groups in reverse order. All the works stated above proposed different microaggregation methods to form the groups, where within groups the records are homogeneous but between groups the records are heterogeneous and sum of squares of errors (SSE) are used the measure the information loss. As median is used as a measure of location to represent each group, in this paper we proposed sum of absolute deviations from median (ADM) to measure the information loss that is always less than the SSE. That means by using ADM as a measure of information loss always produce less information loss than the SSE. Thus the proposed median based microaggregation method has the following features:

- It divides the whole microdata set into a number of mutually exclusive and exhaustive groups prior to publication and then publishes the median over each group instead of individual records.
- It guarantees that the modification has no effect and the modified data and the original data are similar by using statistical test.
- As microaggregated data causes information loss, it uses sum of absolute deviations from median (ADM) as a measure of distortion that is always less than the so called distortion measure sum of squares of errors (SSE).

The reminder of this paper is organized as follows. We present a brief description of our proposed microaggregation method in Section 4. In Section 5, we present proposed distortion metric to measure the homogeneity of the records in a group.

Important properties of the proposed and metric are discussed in Section 6. Finally, concluding remarks are included in Section 7.

4. THE PROPOSED APPROACH

Microdata protection through microaggregation has been intensively studied in recent years. Many techniques and methods have been proposed to deal with this problem. In this section we first describe some basic concept of microaggregation and the proposed approach of microaggregation.

When we microaggregate data we should keep in mind in two goals, data utility and preserving privacy of individuals. For preserving the data utility we should introduce as little noise as possible into the data and for preserving privacy data should be sufficiently modified in such a way that it is difficult for an adversary to re identify the corresponding individuals.

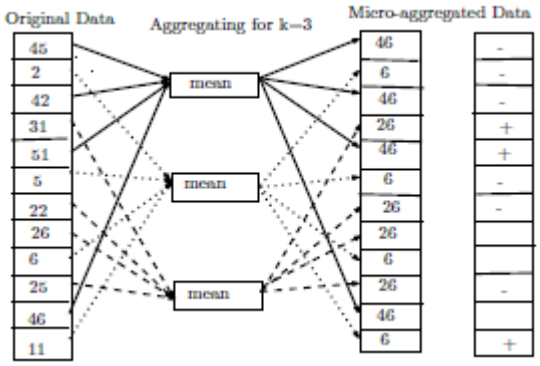

FIGURE 1 Example of Microaggregation using mean

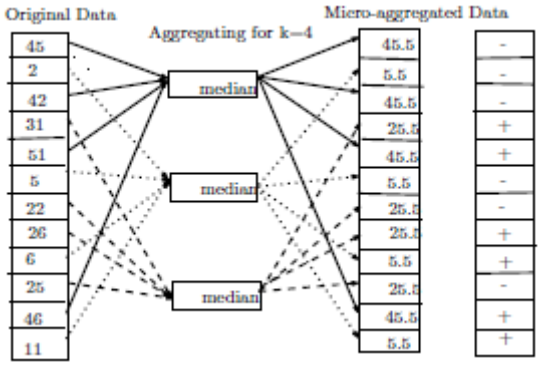

FIGURE 2 Example of Microaggregation using median

Figure 1 and Figure 2 show examples of microaggregated data where in Figure 1, the centroid is replaced by mean and in Figure 2; the centroid is replace by median. Both the figures show that 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.

Now it is necessary to check which figure shows similar original data and microaggregated data by using statistical test. The sign test can be used to test the hypothesis that there is no difference between the distributions of original data and the microaggregated data. Usually, sign test is used to test the effectiveness of an experiment. For example, in weight reducing program weight may be taken before the experiment start and after the experiment completed. Thus it is possible to get pair values of each individual and sign test is used to test whether the program is effective or not. In this situation, we have original values as well as corresponding modified values. Thus we get pair values of each record, so sign test can be used whether the modification has any effect or not. Both the figures consist of three groups and each group has four elements. First group consists of the elements 45, 42, 51 and 46, the second group consists of the elements 2, 5, 6 and 11 and the third group consists of the elements 31, 22, 26 and 25 where in Figure 1 these values are replaced by their

corresponding group mean and in Figure 2 these values are replaced by their corresponding group median. We would now like to test whether the original data and the micrioaggregated data are similar. Set up a null hypothesis H_0: the microaggregation method has no effect and the alternative hypothesis is H_a: the microaggregation method has an effect. Take the difference of microaggregated data from original data, give a '+' sign if the difference is positive and give a '-' sign if the difference is negative. We omit pairs for which there is no difference and count the number of positive differences (X).

If we use median as centroid value then total pairs is, n = 12 (as no tie) and the number of positive sign is, X = 6. This is exactly what we would expect if there is no difference. Thus we can't reject H_0, no evidence to support the hypothesis that the microaggregation method has an effect. That means the modification has no effect and both the microaggregated data and the modified data are similar. So, it can be concluded that by using median as centroid value, always give guarantee of producing similar original and modified data. This is true for any dataset as median is the middle most observations in a set of values.

On the contrary, if we use mean as centroid value then total pairs is, n = 12-3 = 9 (as three tie) and the number of positive sign is, X = 3. This is not exactly what we would expect if there is no difference. That means we can't say anything unless getting p-value as the acceptance or rejection of H_0 depends on p-value. So there is no guarantee that the microaggregated data and the original data are similar. For some cases this may be true but this is not universally true for any particular dataset. So, it

can be concluded that by using mean as centroid value, does not give any guarantee of producing similar original and modified data for any dataset.

As discussed, microaggregation method using median provides sufficient evidence that the modified data are similar to the original data, in this paper we propose to use median as the centroid point of each group. So before publish, microdata should be partitioned into some groups such that within groups the records are more close to each other and between groups the records are more distant to each other, and then publish the median over each group instead of individual records.

5. PROPOSED DISTORTION METRIC

Consider a microdata set T with p numeric attributes and n records, where each record is represented as a vector in a p-dimensional space. For a given positive integer $k \leq n$, a microaggregation method partitions T into g groups where each group contains at least k records (to satisfy k-anonymity), and then replaces the records in each group with the median of the group. Let n_i denote the number of records in the ith group, and $x_{ij}, 1 \leq j \leq n_i$, denote the jth record in the ith group. Then, $n_i \geq k$ for $i = 1$ to g, and $\sum_{i=1}^{g} n_i = n$. The centroid of the ith group, denoted by m_i, is calculated as the middle most (median) vector of all the records in the ith group. By using median, microaggregated dataset produces similar as the original dataset but not the same data and so there is still chance of being information loss. Information loss is used to quantify the amount of information of a dataset that is lost after applying a microaggregation method. To reduce the information loss it is necessary to form the groups using a criterion of maximum similarity. That means the records in each group

are more close to each other. To measure whether the records in each group are close to each other, in this paper we use sum of absolute deviations from median (ADM) of each group and is defined as

$$ADM = \sum_{i=1}^{g}\sum_{l=1}^{p}\sum_{j=1}^{n_i} | (x_{ilj} - m_{il}) | \qquad (1)$$

where, x_{ilj} is the jth record of lth attribute in the ith group and m_{il} is the median of lth attribute in ith group. As we replace each record by their corresponding group median, the distortion is measured by the difference between individual record and its median. The lower the distance, median is close to its original value and higher the distance median is far from its true value. We are only measuring the distance as it is not interest to us whether the distance is positive or negative. Thus we take the absolute difference and the ADM is used to measure the information loss due to using median based microaggregation method. On the other hand, ADM could also be used to measure the homogeneity of the groups. The lower the ADM, the records of the group are more homogeneous to each other. Previously, the most common measure of information loss proposed by Domingo-Ferrer and Mateo-Sanz (2002) is the Sum of Squares of Errors (SSE) and is defined by

$$ADM = \sum_{i=1}^{g}\sum_{l=1}^{p}\sum_{j=1}^{n_i} (x_{ilj} - m_{il})^2 \qquad (2)$$

where p is the total number of numerical attributes in the dataset and $\bar{x}_{il}$ is the mean of lth variable in the ith group. It should be noted that the sum of deviations from their mean of a set of observations is always zero, i.e. $\sum_i (x_i - \bar{x}) = 0$ and so sum of squares of deviations from mean was used to measure the similarity of each group. As in this paper we are taking sum of deviations from median, i.e. $\sum_i (x_i - m)$, it always gives a value and so we don't need to square of these deviations, rather we take

absolute value of theses deviations. Thus, given a homogeneity measure such as ADM and a security parameter *k*, which determines the minimum cardinality of the groups, the microaggregation problem can be enumerated as follows:

Definition: *Given a dataset T of n elements and a positive integer k, find a partitioning* $\boldsymbol{G} = \{G_1, G_2, \ldots, G_g\}$ *of T such that ADM is minimized subject to the constraint that* $|G_i| \geq k$ *for any* $G_i \in \boldsymbol{G}$.

Once we get the homogeneous groups, the median value over each group is computed and replaces each of the original group values. Thus we get a microaggregated microdata set and could be publish for general public use. It confirms that the microaggregated dataset is similar to the original data and preserves the privacy of individuals as well as increase the data utility.

6. ANALYSIS OF THE APPROACH

As discussed, in this paper we proposed a median based microaggregation method and proposed a distortion metric ADM to measure the homogeneity of the records in a group. In this section we will discuss some of the properties of the proposed approach and the metric.

Theorem 1 *Suppose an attribute in a dataset consists of some groups and each group consists records of at least k. Let the records of each group is replaced by the median of the corresponding group. Then the attribute consists of the original records and the attribute consists of the modified records (medians) have the same distribution.*

group	1			2			…	g		
X	x_1	…	x_k	x_{k+1}	…	x_{2k}	…	$x_{(g-1)k+1}$	…	x_{gk}
M	m_1	…	m_1	m_2	…	m_2	…	m_g	…	m_g
sign	-	…	+	-	…	+	…	-	…	+

FIGURE 3 Values of an attribute

Proof Suppose an attribute in a dataset consists of n records that are exactly divisible by k. So the attribute consists of $g = \frac{n}{k}$ groups and each group consists of k records. Suppose the attribute consists of the values, $x_1, \ldots, x_k, x_{k+1}, \ldots x_{2k}, \ldots, x_{(g-1)k+1}, \ldots, x_{gk}$, where the first group consists of first k-values, the second group consists of second k values, ..., and the last group consists of last k-values as shown in Figure 3. Also let $m_i (i = 1, \ldots, g)$ be the median of the ith group respectively, where m_i is the middle most observation of the ith group, when the values in ith group are arranged in order of magnitude. Thus the corresponding microaggregated values of the original values of the attribute are $m_1, \ldots, m_1, m_2, \ldots, m_2, \ldots, m_g, \ldots m_g$, where first k-values consists in the first group, second k- values consists in the second group and so on, if median is replace as the centroid. Thus we get match pair data and let (X_i, M_i) be n pairs of observations. We wish to test,

H_0 : X and M follow the same distribution,

H_a : The two distributions differ in location.

Let $D_i = Xi - M_i$. Under H_0, both X and M comes from the same distributions, so

$$P(D_i \text{ is positive}) = P(D_i \text{ is negative}) = \frac{1}{2}.$$

Let W be the total number of positive differences (D_i's). If X_i and M_i follow the same distribution then W follows Binomial distribution with parameters n and $\frac{1}{2}$.

Suppose the values in each group are arranged in order of magnitude. Thus for each group we get first half is positive sign and the rest half is negative sign. We omit pairs for which there is no difference, this may cause when k is an odd number. Thus finally

the total number of positive sign is $\frac{n-g}{k}$, if n is odd and $\frac{n}{2}$, if n is even. That means the number of positive signs and the number of negative signs would be the same whatever k is even or odd. This is exactly what we would expect if there is no difference. Thus we can not reject H_0, showing that original values and the modified values of the attribute follows the same distribution and thus they are similar. Similarly this can be shown if n is not exactly divisible by k. This is true for each and every attribute in a microdata set. Thus if a microdata set is partition in to some groups and each record of a particular group is replaced by the corresponding median, then the microaggregated microdata set and the original dataset have the same distribution. We will now show that the homogeneity measure ADM proposed in this paper is always less than the so called homogeneity measure SSE. Before that we would like to discuss the following theorem.

Theorem 2 *Sum of absolute deviations of a set of observations from their median is always less than the deviations from mean.*

Proof Let $x_1, x_2, \ldots, x_n$ be a set of n observations. Let us assume that n is an even number and so $n = 2p$, where p is an integer. Thus median (m) lies between x_p to x_{p+1}. Also let $\bar{x}$ is the arithmetic mean which lies between x_k to x_{k+1}. Here we would like to show that

$$\sum_{i=1}^{n} | x_i - m | \leq \sum_{i=1}^{n} | x_i - \bar{x} |$$

Let us first take the absolute deviations from mean, say D_1

$$D_1 = (\bar{x} - x_1) + (\bar{x} - x_2) + \ldots + (\bar{x} - x_k)$$

$$+ (x_{k+1} - \bar{x}) + (x_{k+2} - \bar{x}) + \ldots + (x_p - \bar{x})$$

$$+ (x_{p+1} - \bar{x}) + (x_{p+2} - \bar{x}) + ... + (x_n - \bar{x}) \quad (3)$$

and the absolute deviations from median, say D_2

$$D_2 = (m - x_1) + (m - x_2) + ... + (m - x_k)$$
$$+ (m - x_{k+1}) + (m - x_{k+2}) + ... + (m - x_p)$$
$$+ (x_{p+1} - m) + (x_{p+2} - m) + ... + (x_n - m) \quad (4)$$

Therefore,

$$D_1 - D_2 = (\bar{x} - m)k - (\bar{x} + m)(p - k) - (\bar{x} - m)(n - p) + 2(x_{k+1} + x_{k+2} + ... + x_p)$$
$$= 2(x_{k+1} + x_{k+2} + ... + x_p - \bar{x}(p - k))$$
$$= 2[(x_{k+1} - \bar{x}) + (x_{k+2} - \bar{x}) + ... + (x_p - \bar{x})] \quad (5)$$

which is a positive quantity, so sum of absolute deviations from median is always less than the deviations from mean. In other words,

$$\sum_{i=1}^{n} | x_i - m | \leq \sum_{i=1}^{n} | x_i - \bar{x} |$$

without any loss of generality, we can say that

$$\sum_{i=1}^{n} | x_i - m | \leq \sum_{i=1}^{n} (x_i - \bar{x})^2$$

This is true for every group in an attribute, for every attribute and for every dataset consisting of several numeric attributes. So,

$$\sum_{i=1}^{g} \sum_{l=1}^{p} \sum_{j=1}^{n_i} | (x_{ilj} - m_{il}) | \leq \sum_{i=1}^{g} \sum_{l=1}^{p} \sum_{j=1}^{n_i} (x_{ilj} - \bar{x}_{il})^2$$

$$\Rightarrow ADM \leq SSE \quad (6)$$

Thus the proposed homogeneity measure in this paper ADM is always less than the SSE. In other words, ADM always incur less information loss than the SSE for any dataset.

7. CONCLUSION

Microaggregation is an effective method of protecting privacy in microdata. This work presents a new microaggregation method for numerical attributes. The new method consists of clustering individual records in microdata in a number of disjoint groups prior publication and then publish the median over each group instead of individual records. We showed by using statistical test that the microaggregated data and the original data have the same distribution. As it produces the similar dataset, the statistical results also produce the similar results as in the original dataset. In addition, in this paper we proposed a distortion metric to measure the homogeneity of the records in a group. The metric, called ADM can be used to measure the amount of information loss due to microaggregation. We showed that ADM always produce less information loss than the previous information loss metric. This method of microaggregation can be extremely useful for researchers, experts and the associated people to analysis data accurately and efficiently as it protects the privacy of individuals as well as produces the similar original data set.

REFERENCES

Bezdek, J.C. (1981) Pattern recognition with fuzzy objective function algorithms. Norwell, MA: Academic Publishers.

Domingo-Ferrer, J., & Torra, V. (2005). Privacy in data mining. Data Mining and Knowledge Discovery, 11(2), 117–119.

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

Domingo-Ferrer, J., & Torra, V. (2002). Extending microaggregation procedures using defuzzification methods for categorical variables. In Proceedings of the 1st international IEEE symposium on intelligent systems, 44-49, Verna, September 2002.

Domingo-Ferrer, J., & Torra, V. (2002). Towards fuzzy *c*-means based microaggregation. In P. Grzegorzewski, O. Hryniewicz & M.A. Gil (Eds.), Advances in soft computing (pp. 289-294). Heidelberg: Physica-Verlag.

Domingo-Ferrer, J., & Torra, V. (2003). Fuzzy microaggregation for microdata protection. Journal of Advanced Computational Intelligence and Intelligent Informatics, 7(2), 153-159.

Domingo-Ferrer, J., & Torra, V. (2005). Ordinal, continuous and heterogeneous kanonymity through microaggregation. *Data Mining and Knowledge Discovery,* 11(2), 195-212.

Domingo-Ferrer, J., Martnez-Ballest, A., Mateo-Sanz, J.M., & Seb, F. (2006). Efficient multivariate data-oriented microaggregation. The VLDB Journal, 15(4), 355-369.

Domingo-Ferrer, J., Seb, F., & Solanas, A. (2008). A polynomial time approximation to optimal multivariate microaggregation. Computer and Mathematics with Applications, 55(4), 714-732.

Han, J.M., Cen, T.T., Yu, H.Q., & Yu, J. (2008). A multivariate immune clonal selection microaggregation algorithm. In Proceedings of the IEEE international conference on granular computing, pp. 252-256, Hangzhou, February 2008.

Hansen, S., & Mukherjee, S. (2003). A polynomial algorithm for optimal univariate microaggregation. IEEE Transactions on Knowledge and Data Engineering, 15(4), 1043-1044.

Laszlo, M., & Mukherjee, S. (2005). Minimum spanning tree partitioning algorithm for microaggregation. *IEEE Transactions on Knowledge and Data Engineering,* 17(7), 902-911.

Oganian, A., & Domingo-Ferrer, J. (2001). On the complexity of optimal microaggregation for statistical disclosure control. Statistical Journal of the United Nations Economic Commission for Europe, 18, 345-354.

Solanas, A. (2008). Privacy protection with genetic algorithms. In A. Yang, Y. Shan & and L.T. Bui (Eds.), Studies in Computational Intelligence (pp. 215-237). Heidelberg: Springer.

Solanas, A., Martinez-Balleste, A., & Domingo-Ferrer, J. (2006). *V-MDAV*: A multivariate microaggregation with variable group size. In Proceedings of the 17th COMPSTAT Symposium of the IASC, Rome, August 2006.

[16] Samarati, P. (2001). Protecting respondent's privacy in microdata release. IEEE Transactions on Knowledge and Data Engineering, 13(6), 1010–1027.

Sweeney, L. (2002). *k*-Anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,* 10(5), 557-570.

Torra, V. (2004). Microaggregation for categorical variables: A median based approach. In J. Domingo-Ferrer & V. Torra (Eds.), LNCS (pp. 162-174), Heidelberg: Springer.

Ward, J.H.J. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58(301), 236-244.

Willenborg, L., & Waal, T.D. (2001). Elements of statistical disclosure control. Lecture notes in statistics, 155.

Zahn, C.T. (1971). Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Transactions on Computers, C-20 (1), 68-86.

Chang, C.C., Li, Y.C., & Huang, W.H. (2007). TFRP: An efficient microaggregation algorithm for statistical disclosure control. Journal of Systems and Software, 80(11), 1866–1878.

Lin, J.L., Wen, T.H. Hsieh, J.C., & Chang, P.C. (2010). Density-based microaggregation for statistical disclosure control. Expert Systems with Applications, 37(4), 3256–3263.