Statistical Disclosure Control for Data Privacy Preservation

Sarat Kumar Chettri
Department of Computer Science, Saint Mary's College
Shillong, Meghalaya-793003, India

Bonani Paul
Department of Computer Science, Saint Mary's College
Shillong, Meghalaya-793003, India

Ajoy Krishna Dutta
Department of Computer Science, Saint Mary's College
Shillong, Meghalaya-793003, India

ABSTRACT
With the phenomenal change in a way data are collected, stored and disseminated among various data analyst there is an urgent need of protecting the privacy of data. As when individual data get disseminated among various users, there is a high risk of revelation of sensitive data related to any individual, which may violate various legal and ethical issues. Statistical Disclosure Control (SDC) is often applied to statistical databases for preserving the privacy of individual data. Microaggregation is an efficient Statistical Disclosure Control perturbative technique for microdata protection i.e. protection of individual data. Unlike k-Anonymity, microaggregation method modifies data without suppressing or generalizing it. But to prevent the disclosure of sensitive data it should not be modified to an extent that the data utility is affected. So, the major challenge is how to perturb the data in such a way that a balance is maintained between data utility and risk of data disclosure. Here in this paper, we have proposed a new SDC method based on multivariate data-oriented microaggregation technique for individual data protection with minimal information loss and low data disclosure risk. Experimental results show that our proposed method proves our claim as when compared with other state-of art existing methods of data protection.

General Terms
Statistical Disclosure Control, PPDM, Microaggregation

Keywords
SDC, Microaggregation, information loss, data disclosure risk, microdata, perturbative, *k*-Anonymity..

1. INTRODUCTION
In recent years with the latest advancement of information technology, there is a change in collection of huge amount of data from various sources (governmental or private) for various analyses. There is a phenomenal change in a way data are collected, stored and disseminated among various researchers, analyst and data miners for knowledge discovery. The discovered knowledge which was previously unknown, facilitates the decision making processes in different areas like in marketing and supply-chain management, medical and health care for making policies and planning strategies etc. Following are some of the scenario under which data mining techniques plays an important role for data analyses and knowledge discovery.

- If a government of a country decides for implementing various social welfare schemes for its people, then detailed study is needed to be done on the demography of the region, population etc.
- For a company to launch any new product in a market, it first needs to study the market such as consumption trend, buying habits of people etc.
- For stock market prediction, weather forecasting, web usage mining etc.

For such research analysis and planning, large amount of data sets are being shared and published, which in turn increases the risk of breaching the privacy of individuals associated with the data sets. The dissemination of records should be done in such a way that it does not violate any legal issues by limiting privacy breaches on individual records while at the same time provides meaningful analytical results applying data mining techniques. To protect individual records form identification, Statistical Disclosure Control (SDC) methods are often used for protection [1]. SDC methods are applied on the data before releasing it for different analyses. The Statistical Disclosure Control (SDC) attempts to have a balance between a person's right to privacy and the right of a society to know about the data for analyses. The definition of privacy has been formally stated in [2] as *"The right of an entity to be secure from unauthorized disclosure of sensible information that are contained in an electronic repository or that can be derived as aggregate and complex information from data stored in an electronic repository"*.

Traditionally, SDC methods have been devised to protect respondent privacy by entailing some degrees of data modification. Microaggregation is an efficient Statistical Disclosure Control perturbative technique for microdata protection i.e. protection of individual data. Unlike *k*-Anonymity, microaggregation method modifies data without suppressing or generalizing it. It was first proposed in the year 1995 by Defays and Anwar [4] as a special clustering problem where a data set is partitioned into small homogenous groups. Each group contains at least *k* records and instead of releasing the raw microdata values, the mean of the group they belong to is reported in their place prior to their publication or release. Thus, we can say that microaggregation naturally satisfies *k*-Anonymity. But microaggregation is not about simple clustering or partitioning a data set into homogenous groups where each group consists of at least *k* records. It is very crucial to group records in such a way that the data disclosure risk is kept at the minimal level while keeping the data utility high. In other words we can say that a better trade-off is required between the risks of disclosing the sensitive data and the loss of information occurred due to data modification. The microaggregation method was originally defined for continuous data by Defays and Nanopoulos [3] and also in other works as can be seen in [4, 5]. It was then extended for categorical data [7] and later for heterogeneous data [6]. The optimal microaggregation method partitions a

data set into groups of size lying between *k* and 2*k*-1. The user defined parameter *k* decides the degree of perturbation, large value of *k* may ensure higher data privacy but the data may not be useful for statistical analyses as information loss may be higher. Normally, for moderate size data set the *k* value is taken as 3, 4, 5 or 10 in any microaggregation method.

The rest of the paper is organized as follows. Section 2 gives background knowledge about microaggregation method and microdata protection. Section 3 gives an insight of the existing microaggregation methods. Section 4 explains the proposed method *CV-MDAV*. In Section 5, experimental data and results are presented and the effectiveness of the proposed algorithm is assessed. Finally, in Section 6 conclusions are drawn with future work directions.

2. BACKGROUND

Microdata are information about respondent individual for e.g. company data, data related to a person etc. It can be also viewed as a file which consists of *n* individual records with *m* attributes. The microdata attributes can be classified into following categories –

- Identifiers – These attributes can be used to identify individual records uniquely, for e.g. Employee ID, patient code etc.
- Quasi-identifiers – These attributes can be used to identify individual records, but not uniquely, as the records which are identified may be ambiguous. For e.g. person's age, name etc.
- Confidential attributes – These attributes contains some individual respondent information which is sensitive in nature to some extent. For e.g. patient's diagnosis report, person's community etc.
- Non-confidential attributes – The attributes which do not fall in any of the categories as mentioned above belong to this category. For e.g. person's hobbies, language skills etc. These kinds of attributes cannot be neglected as they can be a part of quasi-identifier.

The microdata file is shared among users/analysts for various research analyses which increases the risk of disclosing some sensitive information about the individuals concerning the data. There are various techniques available for protecting microdata from individual identification. It can be performed either by data modification/data masking or by generating synthetic data [12]. In both the techniques, the main aim is to get new microdata set *V'* from its original counterpart *V*. Irrespective of the techniques applied to obtain *V'*, it should serve the primary goal of low risk of disclosing data keeping its statistical information content high. The data masking technique can be broadly classified as perturbative or non perturbative methods as shown in [8, 9].

2.1 Microaggregation Concepts

Microaggregation is a Statistical Disclosure Control (SDC) method which is perturbative in nature. It is an efficient method for microdata protection and was first proposed by Defays and Anwar [4] in the year 1995. It was originally defined for continuous data by Defays and Nanopoulos [3] and also in other works as can be seen in [4, 5] and was then extended for categorical data [7] and later for heterogeneous data [6]. The microaggregation method follows mainly two steps; first it partitions the dataset into homogenous groups where each group consists of at least *k* records (where *k* is a user defined parameter) and then every record of a group is substituted with the corresponding group's mean value. There is no constraint in the number of groups that can be formed but group size should lie between *k* and 2*k*-1. Microaggregation automatically satisfies *k*-Anonymity [17] without generalizing or suppressing data. In *k*-Anonymity, every record is indistinguishable from at least (*k*-1) other records. Usually the distance measure used to determine the similarity of records in microaggregation method is Euclidean distance. To be more specific let us consider a microdata set *R* with *d*-dimensional variables on *n* individuals. Now, when microaggregation method is applied on the microdata set then *m* groups are formed with at least *k* records in each group. The centroid $\bar{x}$ is the average vector of all the records in data set *R* and $\bar{x}_i$ is computed as average vector of all the records in the cluster c_i. Optimal partition of the microdata set is measured in terms of within group sum of squares (*SSW*) in Eqn. (1) or alternatively by the between-group sum of squares (*SSB*) in Eqn. (2). For a given data set *SST* in Eqn. (3) is fixed irrespective of how microdata set *R* has been partitioned. *SSE* on the other hand varies from one cluster to other. Lower the *SSE* records are more homogenous in a cluster.

$$SSW = \sum_{i=1}^{m} \sum_{j=1}^{k_i} ||x_{ij} - \bar{x}_i||^2 \qquad (1)$$

$$SSB = \sum_{i=1}^{m} k_i \, ||\bar{x}_i - \bar{x}||^2 \qquad (2)$$

$$SST = SSW + SSB = \sum_{l=1}^{n} ||\bar{x}_l - \bar{x}||^2 \qquad (3)$$

Where,

x_{ij} = *j*-th record in the *i*-th cluster.

k_i = k_i records in the *i*-th cluster.

The effectiveness of microaggregation method can be measured by evaluating the information loss caused due to data modification. The Information Loss (*IL*) is calculated as

$$IL = \frac{SSW}{SST}.100 \qquad (4)$$

To assess the security of anonymized table the data disclosure risk measurement is used. We adopt here the Distance Linkage Disclosure Risk (*DLD*) model as in [16]. It is based on the probability of inferring the original record from the anonymized table. It can be defined as for any anonymized record X_0 in an anonymized table D_0 if we compute a distance to other records in the original table *D*, we can get a nearest record X_1 and a second nearest record X_2. If X_1 or X_2 is the original record *X*, then the record *X* is called a *linked_record*. Let *num_linked_record* be the number of linked records in an anonymized table, *total_num_record* be the total number of records in an anonymized stable, *DLD* is defined as

$$DLD = \frac{num\_of\_linked\_record}{total\_num\_record}.100 \qquad (5)$$

According to the dimensionality of data in the microdata set, microaggregation method can be divided into two categories –

- Univariate micoaggregation – It is applied to each variable of a microdata set in an independent manner. The problem becomes easier, as only single variable is involved, where the idea of individual ranking can be applied as can be seen in [5]. Furthermore, in [10] we can see that there exists a polynomial-time optimal algorithm for univariate microaggregation method.
- Multivariate microaggregation – Here, the grouping process is applied to sets of variables of the microdata set. In this case, when all the variables are microaggregated together, *k*-Anonymity is automatically satisfied thereby reducing the risk of data disclosure. Thus, one can concentrate in maximising data utility. A polynomial time optimal multivariate microaggregation method is an NP hard problem as stated in [11]. Consequently, several heuristics have been proposed under this category.

Irrespective of the data dimensionality of the microdata set, the microaggregation method applied can be of fixed-size or data-oriented (variable size). The fixed-size method partitions a microdata set into groups of size *k* where each group contains *k* records except one which may contain more than *k* records when the number of records in the microdata set is not a multiple of *k*, whereas the data-oriented microaggregation method produces groups of variable sizes. The group size lies between *k* and 2*k*-1. Though fixed-size microaggregation method takes less computation time in partitioning the dataset by reducing the search space but variable size method tends be more flexible in grouping records as it can adapt to various data distribution, thus increasing within group homogeneity and incurring lesser information loss.

3. EXISITING MICROAGGREGATION METHODS

Various approaches exist in the literature in microaggregation area for microdata protection. Domingo Ferrer and Mateo Sanz proposed a multivariate fixed size microaggregation method called *MD* (maximum distance) method. Till there are more than 2*k* records the method repeatedly locates two most distant records of the data set and simultaneously forms two groups with their respective *k*-1 nearest records. A new cluster is formed with the remaining *k* records. Incase that there are less than *k* remaining records not belonging to any group then they are assigned to their respective closest clusters.

A very similar method to *MD*, the *MDAV* (Maximum Distance to Average Vector) method [12] has been proposed in the literature. *MDAV* works by finding the most distant record *r* from the global centroid and another record *s* which is most distantly located from *r*. Now two clusters are formed with *r* and *s* separately with their respective nearest (*k*-1) records. The process is repeated as long as there are less than 2*k* records. A new cluster is formed if the number of remaining records is between *k* and 2*k*-1. Incase only less than *k* records are remained then the records are assigned to their respective closest clusters.

The *MDAV-generic* [8] algorithm is a variant of the *MDAV* algorithm. This algorithm smoothly handles the remaining records after there are lesser than 3*k* remaining records. If the number of remaining records is between 2*k* and 3*k*-1 records then a cluster is formed with the (*k*-1) nearest neighbours of the most distant record from the centroid of the remaining records. If there are remaining *k* unassigned records then it is assigned to their respective closest clusters. If less than 2*k* records remain, a new cluster is formed with those remaining records.

A modified version of *MDAV* method has been proposed by Lin et. al. known as *MDAV-1* [14]. The new method *MDAV-1* differs from *MDAV* method in the sense that when the number of remaining records is between *k* and 2*k*-1 then a most distant record *r* is found from the centroid of the remaining records and a cluster is formed with *r* and its nearest (*k*-1) records. When number of records is less than *k* then the remaining records are assigned to its closest clusters.

V-MDAV (Variable-size Maximum Distance to Average Vector) is the variable-size variant of *MDAV* microaggregation method presented is presented by Solanas et. al in [13]. This algorithm extends the group that is currently formed up to a maximum size of 2*k*-1 based on some heuristics. To extend the current group it finds the closest unassigned record, e_{min} outside the group to any record inside the group and the corresponding distance between these two records is termed d_{in}. Then, the closest unassigned record to e_{min} is found with corresponding distance being termed d_{out}. If $d_{in} < \gamma d_{out}$ then the record e_{min} is inserted in the current cluster. The extension process is repeated until the group size is equal to 2*k*-1 or when a decision of inclusion is not satisfied. Here γ is a user defined constant. Values of γ close to zero are effective when the data are scattered, when the data set is clustered the best value of γ is usually close to one.

Lin et. al. has also proposed density based microaggregation method called *DBA* [14] which has a reasonable dominance over the latest microaggregation methods. The *DBA* microaggregation method has two scenarios. Initially in the first phase *DBA-1* partitions a data set into groups where each group contains at least *k* records. To partition the data set, *DBA-1* uses *k*-neighborhood of the record with the highest *k*-density among all the records that are not assigned to any group. The grouping process continues till *k* records remain unassigned. These remaining *k* records are then assigned to its nearest groups. Thus clusters are formed with no less than *k* records in each. Then in the second phase *DBA-2* tries to fine tune the clusters by either splitting the formed clusters or merging one cluster with the other. After splitting and merging still there may exist few clusters with more than 2*k*-1 records. Now to increase the data utility *MDAV-1* microaggregation method is applied to those clusters having more than 2*k*-1 records in it. This phase is known as *MDAV-2*. Regarding the complexity of the related methods, as shown in the survey paper [17] most of the methods have $O(n^2)$ except *MD* method which has a complexity of $O(n^3)$.

4. PROPOSED MICROAGGREGATION METHOD

The proposed method called Centroid based Variable size Maximum Distance to Average Vector (*CV-MDAV*) is a multivariate data oriented microaggregation method (SDC family). In order to reduce information loss of data a gain factor γ has been used to conservatively expand the group. In the experiments a fixed value of $\gamma = 1.1$ has been chosen. The proposed algorithm is stated below:

Algorithm:

CV-MDAV

Input: Data set *X*, *k*.

Output: Microaggregated data set *X'*

1. set i=1; n=$|X|$;

2. while ($n \geq 3k$) do

2.1 compute centroid $\bar{x}$ of remaining records in *X*;

2.2 find the most distant record x_r from $\bar{x}$;

2.3 find 2*k* nearest neighbours ($y_1,y_2,...,y_{2k}$) of x_r;

2.4 form cluster c_i around x_r with first (*k*-1) neighbours ($y_1,y_2,...,y_{k-1}$);

2.5 remove records ($y_1,y_2,...,y_{k-1}$) from dataset *X*;

2.6 set $n=n-k$; $j=k$;

2.7 compute centroid x_i of cluster c_i;

2.8 while ($j<2k$ and $|c_i|<2k$) do

i). find distance d_1 of record x_r from $\bar{x}_i$;

ii). find distance d_2 of record y_j from $\bar{x}_i$;

iii). find k nearest neighbours ($z_1, z_2,...,z_k$)of y_j in *X*;

iv). compute centroid $\bar{z}$ of ($z_1,z_2,...,z_k$);

v). compute distance d_3 of y_j from $\bar{z}$;

vi). if ($d_2 < \gamma d_3$) then

a) insert y_j in current cluster c_i;

b) recompute centroid $\bar{x}_i$ of cluster c_i;

c) remove record y_j from *X*;

d) set n=n-1;

ix). end if

2.9 end while

2.10 set i=i+1;

3. end while

4. if (n>2*k*) then

4.1 compute centroid $\bar{x}$ of remaining records in *X*;

4.2 find the most distant record x_r from x ;

4.3 find 2*k* nearest neighbours ($y_1,y_2,...,y_{2k}$) of x_r;

4.4 form cluster c_i around x_r with its nearest (*k*-1) neighbours ($y_1,y_2,...,y_{k-1}$) ;

4.5. remove records ($y_1,y_2,...,y_{k-1}$) from dataset *X*;

4.6. set $n=n-k$; $i=i+1$;

5. end if

6. if (n>0) then

6.1 form a cluster c_i with the n remaining records;

6.2 $i=i+1$;

6.3 end if

7. end algorithm

The *CV-MDAV* algorithm iterates as long as at least 3*k* records remain unassigned. In each iteration the algorithm finds 2*k* nearest neighbours, denoted by ($y_1,y_2,...,y_{2k}$) of the farthest record x_r from the centroid $\bar{x}$ of the remaining records in dataset *X*. Current cluster, c_i is formed with the first (*k-1*) neighbours ($y_1,y_2,...,y_{k-1}$) of x_r. Each of the other *k* neighbours is tested for inclusion in the currently formed cluster by computing a heuristic. This algorithm also uses a constant gain factor γ in the heuristic to conservatively expand the formed cluster. The value of gain factor γ has been fixed to value 1.1 to reduce the complexity of determining the value of γ which is not a straight forward method.

Being a data oriented microaggregation method; it provides a flexibility of further expanding the formed group which initially consists of at least *k* records as mentioned in step 2.8 of the algorithm. The steps are-

1. If $\bar{x}_i$ be the centroid of the cluster c_i we consider the *k*-th neighbour, y_k of x_r.
2. Compute a distance d_1 of record x_r from $\bar{x}_i$ and also distance find distance d_2 of record y_j from $\bar{x}_i$.
3. Find k nearest neighbours ($z_1, z_2,...,z_k$) of y_j in *X*.
4. Compute centroid $\bar{z}$ of ($z_1,z_2,...,z_k$).
5. Compute distance d_3 of y_j from $\bar{z}$.
6. If ($d_2 < \gamma d_3$) then
    a. Expand the formed cluster with inclusion of y_j in the cluster c_i
    b. Recompute the centroid of the expanded cluster c_i.

The test is repeated for the remaining y_{2k-1} records to be included in cluster c_i . Provided the condition ($d_2 < \gamma d_3$) is satisfied cluster c_i is expanded as long as it has less than 2*k*-1 records in it.

5. EXPERIMENTAL RESULTS

In this section we present the various experimental results performed on the proposed *CV-MDAV* microaggregation method. For experimental purpose the proposed method is implemented in C under LINUX environment with a memory of 3GB and *i*3 processor of 2.13 GHz. Experiments are performed on the following three benchmark datasets proposed as reference microdata datasets during the "CASC" project [15].

- The *"Tarragona"* data set contains 834 records with13 numerical attributes.
- The *"Census"* data set contains 1,080 records with 13 numerical attributes.
- The *"EIA"* data set contains 4,092 records with 11 numerical attributes.

Attributes of the datasets are standardized by subtracting their mean and dividing by their standard deviation, so that they have equal weights when computing distances.

The results are presented in terms of Information Loss (IL) comparison of the proposed method with the other standard methods in tables 1, 2 and 3. In table 4 Data Disclosure Risk of *CV-MDAV* with three different datasets has been shown.

Table 1: Information loss comparison using Tarragona data set

Method	*k=3*	*k=4*	*k=5*	*k=10*
MDAV	16.9326	19.545	22.4615	33.1929
MDAV-1	16.9326	19.5457	22.4613	33.1924
MDAV-2	16.382	19.013	22.079	33.179
DBA-1	20.699	23.827	26.001	35.392
DBA-2	16.152	22.671	25.450	34.806
MDAV-generic	16.966	19.546	22.461	33.192
V-MDAV	16.967	19.697	22.886	33.271
CV-MDAV	16.966	19.715	**22.123**	**33.208**

Table 2: Information loss comparison using Census data set

Method	*k=3*	*k=4*	*k=5*	*k=10*
MDAV	5.692	7.494	9.088	14.155
MDAV-1	5.692	7.494	9.088	14.155
MDAV-2	5.656	7.409	9.012	13.944
DBA-1	6.144	9.127	10.842	15.785
DBA-2	5.581	7.591	9.046	13.521
MDAV-generic	5.622	7.494	9.088	14.155
V-MDAV	5.661	7.514	9.007	14.073
CV-MDAV	5.637	7.432	**8.881**	13.949

Table 3: Information loss comparison using EIA data set

Method	*k=3*	*k=4*	*k=5*	*k=10*
MDAV	0.482	0.671	1.666	3.839
MDAV-1	0.482	0.671	1.666	3.839
MDAV-2	0.411	0.587	0.946	3.16
DBA-1	1.09	0.843	1.895	4.265
DBA-2	0.421	0.559	0.818	2.08
MDAV-generic	0.482	0.671	1.666	3.839
V-MDAV	0.509	0.972	1.306	2.809
CV-MDAV	0.582	1.008	1.013	**2.640**

If the Information Loss in Eqn. (4) comparison results given in tables 1, 2 and 3 is observed one can find that the performance of *CV-MDAV* method have a dominance or on par with most of the methods with different values of *k*=3, 4,5,10. Incase of *Tarragona* data set, with *k*=5 and *k*=10 *CV-MDAV* method performs better than other listed methods. Even if the Data Disclosure Risk (*DLD*) in Eqn. (5) is measured as shown in table 4 with the same data set with *k*=5 and *k*=10, it shows satisfactory results. Further in case of *Census* data set as in table 2 similar performance exists of *CV-MDAV* and especially with *k*=5, where it dominates all the compared methods. At table 3 with *k*=10 where *EIA* data set behaves as clustered data set with the given *k* value, *CV-MDAV* outperforms all the listed microaggregation methods.

Table 4: Data disclosure risk with different data sets

Data set	*k=3*	k=4	k=5	k=10
Tarragona	52.757	36.69	28.776	3.839
Census	57.5	41.666	32.685	20.134
EIA	47.873	36.07	29.838	16.731

6. CONCLUSION

Microaggregation is a Statistical Disclosure Control (SDC) method which is perturbative in nature. It is a very popular method for micrdata protection which naturally satisfies *k*-Anonymity without generalization or suppression. The value of *k*, which is a user definable parameter, determines the degree of information loss and data anonymization. With increasing value of *k*, the information loss of microdata increases and risk of data disclosure decreases and vice versa. So the trade-off between information loss and data disclosure risk is always there in any microaggregation method. The proposed method *CV-MDAV* which is a data oriented multivariate microaggregation method shows a better performance in comparison with other existing state-of-art methods. The experiments have been performed with the standard referenced data sets namely "*Tarragona*", "*Census*" and "*EIA*". Thus the proposed method *CV-MDAV* is effective in protecting the privacy of individual data with lower information loss and moderate risk of data disclosure proving to be an efficient data oriented multivariate microaggregation method in SDC. As a future work, the method can be further expanded to protect time series and mixed data sets. Further exploration can also be done to find its effectiveness in case the data are distributed.

7. ACKNOWLEDGMENTS

This work is partly supported by the UGC, India under Minor Research Project grant (Vide No.F.5-313/2011-12/MRP/NERO/10898 Dated 5th Dec 2011)

REFERENCES

[1] L. Willenborg and T. DeWaal, "Elements of Statistical Disclosure Control", Lecture Notes in Statistics, Springer-Verlag, New York, 2001.

[2] E. Bertino, D. Lin and W. Jiang, "A Survey of Quantification of Privacy Preserving Data Mining Algorithms", in Privacy Preserving Data Mining, Springer, US, 2008.

[3] D. Defays and P. Nanopoulos, "Panels of enterprises and confidentiality: The small aggregates method", in 92 Symposium on Design and Analysis of Longitudinal Surveys, Canada, Ottawa, 1993, 195–204.

[4] D. Defays and N. Anwar, "Micro-aggregation: A generic method, in 2nd International Symposium on Statistical Confidentiality", Eurostat, Luxembourg, 1995, 69–78.

[5] J. Domingo-Ferrer and J.M. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control", *IEEE* Transactions on Knowledge and Data Engineering, Vol. 14, No. 1, 2002, 189–201.

[6] J. Domingo-Ferrer and V. Torra, "Ordinal, continuous and heterogenerous k-anonymity through microaggregation", Data Mining and Knowledge Discovery, Vol. 11, No.2, 2005, 195–212.

[7] V. Torra, "Microaggregation for categorical variables: A median based approach", in J. Domingo-Ferrer and V.

Torra Eds. Lecture Notes in Computer Science, Vol. 3050, Springer, Berlin, Heidelberg, 162–174.

[8] J. Domingo-Ferrer and V. Torra, "Ordinal, continuous and heterogenous k-anonymity through microaggregation", Data Mining and Knowledge Discovery, Vol. 11, No. 2, 2005, 95–212.

[9] P. Samarati, "Protecting respondents' identities in microdata release", IEEE Trans. Knowledge and Data Engineering, Vol. 13, No. 6, 2001, 1010–1027.

[10] S.L. Hansen and S. Mukherjee, "A polynomial algorithm for optimal univariate microaggregation", *IEEE* Transactions on Knowledge and Data Engineering, Vol.15, No. 4, 2003, 1043–1044.

[11] A. Oganian and J. Domingo-Ferrer, "On the complexity of optimal microaggregation for statistical disclosure control", Statistical Journal of the United Nations Economic Comission for Europe, Vol. 18, No. 4, 2001, 345–354.

[12] A. Hundepool, A. V. deWetering, R. Ramaswamy, L. Franconi, A. Capobianchi, P.-P. DeWolf, J. Domingo-Ferrer, V. Torra, R. Brand & S. Giessing, (2003) "μ-ARGUS version 3.2 Software and User's Manual", Voorburg NL: Statistics Netherlands, http://neon.vb.cbs.nl/casc.

[13] A. Solanas & A. Mart´ınez-Ballest´e, "V-MDAV: A multivariate microaggregation with variable group size", Seventh COMPSTAT Symposium of the IASC, 2006, Rome.

[14] J.L. Lin, T.H. Wen, J.C. Hsieh, and P.C. Chang, "Density-based microaggregation for statistical disclosure control", Expert Systems with Applications, Vol. 37, No. 4, 2010, 3256–3263.

[15] R. Brand, J. Domingo-Ferrer & J. M. Mateo-Sanz,: "Reference data sets to test and compare sdc methods for protection of numerical microdata", European Project IST-2000-25069, 2002, CASC, http://neon.vb.cbs.nl/casc.

[16] D. Pagliuca,: Some results of individual ranking method on the system of enterprise accounts annual survey. Esprit SDC Project, Deliverable MI-3/ D, 1999.

[17] S.Chettri, B.Paul & A.Dutta, "A Comparative Study on Microaggregation Techniques for Microdata Protection" International Journal of Data Mining & Knowledge Management Process, Vol 2 (6), 2012, 27–40.