A COMPARATIVE STUDY OF VARIOUS MDAV ALGORITHMS

[1]Gajendra Singh Rawat, [2]Dr. Bhogeshwar Borah
[1]Department of Computer Science and Engineering, Tezpur University, India
[2]Associate Professor, Department of Computer Science and Engineering, Tezpur University, India
Email: rawatsgajendra@gmail.com

Abstract: **Microaggregation is an efficient Statistical Disclosure Control (SDC) perturbative technique for microdata protection. It is a unified approach and naturally satisfies k-Anonymity without generalization or suppression of data. Various microaggregation techniques: fixed-size and data-oriented for univariate and multivariate data exists in the literature. These methods have been evaluated using the standard measures: Disclosure Risk (DR) and Information Loss (IL). Every time a new microaggregation technique was proposed, a better trade-off between risk of disclosing data and data utility was achieved. Though there exists an optimal univariate microaggregation method but unfortunately an optimal multivariate microaggregation method is an NP hard problem. Consequently, several heuristics have been proposed but no such method outperforms the other in all the possible criteria. In this paper we have performed a study of the various microaggregation techniques so that we get a detailed insight on how to design an efficient microaggregation method which satisfies all the criteria.**

Keywords: **Statiscal Disclosure Control, Information Loss, Disclosure Risk, microdata, anonymity, microaggregation**

I. INTRODUCTION

Over the last twenty years, there has been an extensive growth in the amount of private data collected about individuals. This data comes from a number of sources including medical, financial, library, telephone, and shopping records. Such data can be integrated and analyzed digitally as it's possible due to the rapid growth in database, networking, and computing technologies. On the one hand, this has led to the development of data mining tools that aim to infer useful trends from this data. But, on the other hand, easy access to personal data poses a threat to individual privacy. This has lead to concerns that the personal data may be misused for a variety of purposes. Detailed person-specific data in its original form often contains sensitive information about individuals, and publishing such data immediately violates individual privacy. The current practice primarily relies on policies and guidelines to restrict the types of publishable data and on agreements on the use and storage of tied to a specific data mining task, and the data mining task may be unknown at the time of data publishing. Furthermore, some PPDP solutions emphasize preserving the data truthfulness at the record level but often PPDM solutions do not preserve such a

A. Clustering

From a practical perspective clustering plays an important role in data mining application. The process of grouping a set of physical or abstracts into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression. Certain fine details are lost by representing the data by fewer clusters but it achieves simplification. It models data by its clusters. Cluster analysis has wide applications, including market or customer segmentation, pattern recognition, biological studies, spatial data analysis, web document classification, scientific data exploration, information retrieval, text mining, medical diagnostics, computational biology, and many others. Cluster analysis can be used as a stand-alone data mining tool to gain

insight into the data distribution or can serve as a pre-processing step for other data mining algorithms operating on the detected clusters.

B. OBJECTIVES

Following are the objectives of work:

- To study the existing privacy preserving data mining methods.
- To analyze experimentally some of the popular preserving techniques.
- To evaluate the performance of the existing methods in terms of security and Information loss.

Microaggregation: Microaggregation is a set of procedures that distort empirical data in order to guarantee the factual anonymity of the data. At the same time the information content of data sets should not be reduced too much and should still be useful for scientific research. The goal of microaggregation is to minimize the *SSE* measure, which is defined as:

$$SSE = \sum_{i=1}^{c} \sum_{x_{ij} \in C_i} (x_{ij} - \bar{x}_i)'(x_{ij} - \bar{x}_i)$$

Where c is the total number of clusters (groups), C_i is the i-th cluster and $\bar{x}_i$ is the centroid of C_i. The total sum of square *SST* is the sum of square error within the entire dataset calculated by summing the Euclidean distance of each record x_{ij} to the centroid $\bar{x}$ as follows:

$$SST = \sum_{i=1}^{c} \sum_{x_{ij} \in C_i} (x_{ij} - \bar{x})'(x_{ij} - \bar{x})$$

Microaggregation techniques are often compared on the basis of the *SSE* or the Information loss (*IL*) measure. The *IL* measure is standardized between 0 and 1 and can be obtained from *SST* as:

$$IL = \frac{SSE}{SST}$$

Clearly, the ultimate goal of SDC techniques lies not only in reducing DR, but also in increasing of the data utility to the user.

C. Variable -Size MDAV

MDAV generates groups of fixed size k and ,thus it lacks flexibility for adapting the group size to the distribution of the records in data set ,which may result in poor within group homogeneity. Variable-size MDAV(V-MDAV) is a new algorithm that intends to overcome the limitation by computing a variable-size k-partition with a computational cost similar to the MDAV cost.

D. Privacy benefits of microaggregation

The attributes in an original unprotected micro-data set V can be classified in four categories which are not necessarily disjoint:

Identifiers These are attributes that unambiguously identify the respondent. Examples are passport number, social security number, full name, etc. Since our objective is to prevent confidential information from being linked to specific respondents, we will assume in what follows that, in a pre-processing step, identifiers in V have been removed/encrypted.

Key attributes Key attributes (called quasi-identifiers)are a set of attributes that, in combination, can be linked with external information to re-identify (some of) the respondents to whom (some of) the records in V refer. Examples of key attributes are age, gender, job, zip code, etc. Unlike identifiers, key attributes cannot be removed from V. The reason is that

II. *MDAV* MICROAGGREGATION ALGORITHM

The *MDAV* method is one of the best heuristic methods for multivariate microaggregation. The algorithm was proposed in [6, 7] as part of a multivariate microaggregation method implemented in the μ-Argus package for statistical

disclosure control. Several variant of this fixed-sized microaggregation algorithm exists with minor differences [6,7,8,23]. Some of them are presented here for comparison purpose and to build the foundation for the proposed algorithm. Firstly, we present the basic version of the *MDAV* algorithm as presented in [13] below. The dataset X with n records and value of group size k are to be provided as input to the algorithm.

Algorithm-1

1. set i=1; n=|X|;
2. while (n>=2k) 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 k-nearest neighbours $y_1,y_2,\ldots,y_k$ of x_r;

2.4 form cluster c_i with the k-neighbours $y_1,y_2,\ldots,y_k$;

2.5 remove records $y_1,y_2,\ldots,y_k$ from dataset X;

2.6 set $n = n - k$; $i = i+1$;

2.7 find the most distant record x_s from x_r;

2.8 find k-nearest neighbours $y_1,y_2,\ldots,y_k$ of x_s;

2.9 form cluster c_i with the k-neighbours $y_1,y_2,\ldots,y_k$;

2.10 remove records $y_1,y_2,\ldots,y_k$ from dataset X;

2.11 set $n = n - k$; $i = i+1$;

2.12. end while

3. if (n>=k) then

3.1 form a cluster c_i with the n remaining records;

3.2 set $n = n - n$; $i = i+1$;

3.3 endif

4. if (n>0) then

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

4.2 find the closest cluster centroid $\bar{c}_j$ from $\bar{x}$;

4.3 add the remaining records to cluster c_j;

4.4 endif

5.end algorithm

Given a dataset X with n records, *MDAV* iterates building two groups, each of size k until number of remaining unassigned records to any group becomes less than 2k (step 2). In order to build these groups, the centroid $\bar{x}$ of the remaining unassigned records is computed at the beginning of each iteration. Then the most distant record x_r from $\bar{x}$ is found and a group is built with the k-nearest neighbors of x_r including itself. These k records are removed from the dataset X. Next, the most distant record x_s from x_r is found and another group is built with the k-nearest neighbours of x_s. When the remaining records after termination of iterations is between k and 2k-1 *MDAV* simply forms a group with all of them (step 3). If less than k records remain all the records of this subgroup are assigned to its closest group determined by computing distance between centroids of the groups (step 4). All groups have k elements except only one group. Finally, given the k-partition obtained by *MDAV*, a microaggregated data set is computed by replacing each record in the original dataset by the centroid of the group to which it belongs. This step is not shown in the algorithm.

A. MDAV Single-group Algorithm

The above algorithm construct two groups in each iteration. The authors in [8] presented the *Centroid-based Fixed-size* microaggregation method that constructs one group in each iteration. The algorithm is adaptation of the *MDAV* algorithm so that the iteration continues until there are at least k records that are unassigned (step 2). Then, each of the remaining records is assigned to its closest group. We adapt the *MDAV*-generic algorithm (algorithm-2) in the similar way. We call it *MDAV-single-group* algorithm and present it here as it forms the basis of our proposed variable-size *MDAV* algorithm.

Algorithm-2 (The *MDAV-single-group* algorithm)

1. set i=1; n=|X|;
2. while (n>=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 k-nearest neighbours $y_1,y_2,\ldots,y_k$ of x_r;

2.4 form cluster c_i with the k-neighbours $y_1,y_2,\ldots,y_k$;

2.5 remove records $y_1,y_2,\ldots,y_k$ from dataset X;

2.6 set $n = n - k$; $i = i+1$;

2.7 end while

3. If (n>=2k) do

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

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

3.3 find k-nearest neighbours $y_1,y_2,\ldots,y_k$ of x_r;

3.4 form cluster c_i with the k-neighbours $y_1,y_2,\ldots,y_k$;

3.5 remove records $y_1,y_2,\ldots,y_k$ from dataset X;

3.6 set $n = n - k$; $i = i+1$;

3.7 end if

4. If (n>0) then

4.1 form a cluster c_i with the n remaining records;

4.2 set n=n-n; i=i+1;

4.3 end if

5. end algorithm

Computational cost of all the four *MDAV* algorithms presented above become $O(n^2)$ [13]. The main problem of these fixed-size algorithms is lack of flexibility. They only generate groups of fixed cardinality k causing higher information loss.

C. VMDAV (Variable-size Maximum Distance to Average Vector) Algorithm

V-MDAV (*Variable-size Maximum Distance to Average Vector*) is the first variable-size microaggregation method presented in [12, 13]. This algorithm extends the group that is currently formed up to a maximum size of 2k-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 2k-1 or when a decision of inclusion is not satisfied. Here γ is a user defined constant. The determination of the best value of γ for a given dataset is not straightforward. Values of γ close to zero are effective when the data are scattered, when the dataset is clustered the best value of γ is usually close to one [12]. To save time *V-MDAV* computes the global centroid of the dataset at the beginning of the algorithm and keeps it fixed instead of recomputing it in each iteration. Each of the remaining records after termination of iterations is inserted to its closest cluster. This may cause a cluster to have number of records in excess of allowed 2k-1 when the closest cluster already contains 2k-1 records because of the extension process. The algorithm for building a k-partition using *V-MDAV* is as follows:

Algorithm-3 (The *V-MDAV* algorithm)

1. set i=1; n=|X|;
2. compute centroid $\bar{x}$ of remaining records in X;
3. while (n>=k) do

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

3.2 find k-nearest neighbours $y_1,y_2,\ldots,y_k$ of x_r;

3.3 form cluster c_i with the k-neighbours $y_1,y_2,\ldots,y_k$;

3.4 remove records $y_1,y_2,\ldots,y_k$ from dataset X;

3.5 set $n = n - k$; *flag*=true;

3.6 if (n==0) then set *flag* = false;

3.7 while ($|c_i|<2k-1$ and *flag*==true)

3.7.1 find unassigned record e_{min} which is the closest to any record of the cluster c_i and let d_{in} be the distance between the two records.

3.7.2 let, d_{out} be the distance from e_{min} to the closest unassigned record in *X*;

3.7.3 if ($d_{in} < \gamma d_{out}$) then

3.7.3.1 assign e_{min} to the current cluster c_i;

3.7.3.2 set $n = n - 1$;

3.7.3.3 if (n==0) then set *flag* = false;

3.7.3.4 else set *flag*=false;

3.8 end while

3.9 $i=i+1$;

3.10 end while

4. if ($n<k$) then

4.1 for each remaining record x_r in *X do*

4.1.1 find the closest cluster centroid $\bar{c}_j$ from x_r;

4.1.2 add the records x_r to cluster c_j;

4.1.3 end for

4.2 endif

5. end algorithm

Computational cost of *V-MDAV* algorithm also remains $O(n^2)$ [12]. Although the *V-MDAV* algorithm produces microaggregated datasets with lower information loss, other variable-size algorithms can be developed with better performance. In the next section we present a new variable-size microaggregation algorithm with lower information loss.

III. PROPOSED METHOD

Our purpose is to experiment a Privacy-Preserving Clustering technique that incur less information loss, hence provides better data utility. Privacy of the database as well as individual privacy will be protected through micro aggregation. Hence here we intend on to develop a new variable-size heuristic for the k-partition, with as much as possible homogeneous records in the same group records.

The proposed heuristics like existing in the literature micro aggregate the records in two successive steps:

1) Partitioning

   The original micro-data file is portioned into several disjoined cluster or group so that all records in the same group are similar to each other and, simultaneously, dissimilar to the records in other groups. Addistionally, each group is forced to contain at least k records. The group can posses 2k-1 records at most.

2) Aggregation

   This phase computes a certain kind of prototype (centroid in our case) for each cluster/group, and it replaces the original values in the micro-units by the computed prototype. This phase usually depends on the type of the variable concerned. In the proposed algorithm the value for the each partition is substituted by average to obscure the identification. Once the groups are formed replacing each record by the prototype becomes straight forward and so this step is not shown in the proposed algorithm presented below.

IV. EXPERIMENTAL RESULTS

In this section we present experimental results performed on the existing methods. We have implemented in C++ under LINUX environment all the three microaggregation algorithms namely *MDAV*, *MDAV-single-group*, *V-MDAV* presented in chapter 4 Experiments are performed on the following three datasets proposed as reference microdata datasets during the “CASC” project [15].

- The “Tarragona” dataset contains 834 records with13 numerical attributes.
- The “Census” dataset contains 1,080 records with 13 numerical attributes.

Table 1. Experimental results.

Dataset	*Method*	*K=3* *SSE : (IL)*	*K=4* *SSE : (IL)*	*K=5* *SSE : (IL)*	*K=10* *SSE : (IL)*
Tarragona	**1.** ***MDAV***	***1835.8318*** ***(16.9326)***	***2119.1678*** ***(19.545)***	***2435.2796*** ***(22.461)5***	***3598.7743*** ***(33.1929)***
	2. ***MDAVsingle***	***1839.4617*** ***(16.9661)***	***2139.1554*** ***(19.7303)***	***2473.9951*** ***(22.8186)***	***3601.2138*** ***(33.2154)***
	3. ***VMDAV***	***1839.6440*** ***(16.9678)***	***2135.5903*** ***(19.6974)***	***2481.3201*** ***(22.8862)***	***3607.2572*** ***(33.2711)***
Census	**1.** ***MDAV***	***799.1827*** ***(5.6922)***	***1052.2557*** ***(7.4947)***	**1276.0162** **(9.0884)**	***1987.4925*** ***(14.1559)***
	2. ***MDAVsingle***	***793.7595*** ***(5.6536)***	***1044.7749*** ***(7.4414)***	***1247.3171*** ***(8.8840)***	***1966.5216*** ***(14.0066)***
	3. ***VMDAV***	***794.9373*** ***(5.6619)***	***1054.9675*** ***(7.5140)***	***1264.5801*** ***(9.0070)***	***1975.8520*** ***(14.0730)***
EIA	**1.** ***MDAV***	***217.3804*** ***(0.4829)***	***302.1859*** ***(0.6713)***	***750.1957*** ***(1.6667)***	***1728.3120*** ***(3.8397)***
	2. ***MDAVsingle***	***215.1095*** ***(0.4779)***	***301.9676*** ***(0.6709)***	***783.0258*** ***(1.7396)***	***1580.8008*** ***(3.5120)***
	3. ***VMDAV***	***229.2986*** ***(0.5094)***	***437.8020*** ***(0.9726)***	***588.0341*** ***(1.3064)***	***1264.4328*** ***(2.8091)***

- The "EIA" dataset contains 4,092 records with 11 numerical attributes.

A. Standardizing the Numeric attributes

Attributes of the datasets are standardized by subtracting their mean and dividing by their standard deviation, so that they have equal weight when computing distances. Standardization/normalization helps to prevent attributes with large ranges (eg.Salary) from outweighing attributes with smaller ranges. It helps to speed up distance measurements in classification or clustering. In we adopted the standardization based on the mean and standard deviation of A.

B. Information Loss Measures

The goal of microaggregation is to minimize the *SSE* measure, which is defined as:

$$SSE = \sum_{i=1}^{c} \sum_{x_{ij} \in C_i} (x_{ij} - \bar{x}_i)'(x_{ij} - \bar{x}_i)$$

Where c is the total number of clusters (groups), C_i is the i-th cluster and $\bar{x}_i$ is the centroid of C_i. The total sum of square *SST* is the sum of square error within the entire dataset calculated by summing the Euclidean distance of each record x_{ij} to the centroid $\bar{x}$ as follows:

$$SST = \sum_{i=1}^{c} \sum_{x_{ij} \in C_i} (x_{ij} - \bar{x})'(x_{ij} - \bar{x})$$

Microaggregation techniques are often compared on the basis of the *SSE* or the information loss (*IL*) measure, which is standardized between 0 and 1 can be obtained from *SST* as

V. CONCLUSION AND FUTURE WORK

In this dissertation we have proposed an inproved variable-size *MDAV* algorithm named that produces lower information loss with little increase in computational complexity ($O(kn^2)$). Fixed-size algorithms have complexity $O(n^2)$. This is acceptable as k is usually a small integer.

Proposed algorithm is a modification of the *MDAV* algorithm to make it variable-size. The algorithm computes $2k$ nearest neighbours of the farthest record from the centroid of the remaining unassigned records in the dataset. First k of the $2k$ neighbours form a cluster and it is extended up to a size of $2k$-1 records by including some of the remaining k neighbours based on a heuristic. The *IVMDAV* algorithm requires a user defined factor γ to be used for the cluster extension process. It can be easily determined as it need to be slightly greater than 1.0 (possible values in the range 1.0 – 1.20).

In future the following considerations can be made to further improve the algorithm. To form a single cluster $2k$ nearest neighbours of the currently selected record for cluster formation is considered. It is possible to consider $3k$ neighbours instead of $2k$ as the algorithm iterates so long as there are at least $3k$ neighbours yet to be assigned to any cluster. This will increase computation time slightly while producing better results as more records are considered for inclusion in the cluster extension. Another possibility for modification of the algorithm is to test whether the current record considered for group formation i.e. the furthest record from the centroid of the remaining records in the dataset is a outlier or not. If it is a outlier than the group formed by the record will remain as a group of k records and it should not be extended to contain upto 2k-1 records

REFERENCES

[1] Agrawal R., Srikant "R. Privacy-Preserving Data Mining". *ACM SIGMOD Conference*,2000.

[2] CHARU C. AGGARWAL and PHILIP S. YU "PRIVACY-PRESERVING DATA MINING:MODELS AND ALGORITHMS"

[3] Sweeney L.: Replacing Personally Identifiable Information in Medical Records, the Scrub System. *Journal of the American Medical Informatics Association*, 1996.

[4] Sweeney L.: Guaranteeing Anonymity while Sharing Data, the Datafly System. *Journal of the American Medical Informatics Association*,1997.

[5] J.M. Mateo-Sanz and J. Domingo-Ferrer, "A Method for Data Oriented Multivariate Microaggregation," *Proc. Statistical Data Protection'* 98,pp. 89-99,1999.

[6] 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.

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

[8] Domingo-Ferrer J, Mateo-Sanz J., Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering* 2002; **14**(1):189–201

[9] Malin B., Sweeney L.: Determining the identifiability of DNA database entries. *Journal of the American Medical Informatics Association*, pp. 537–541, November 2000

[10] Laszlo, M., Mukherjee, S.: Minimum spanning tree partitioning algorithm for microaggregation. IEEETrans. Knowl. Data Eng. **17**(7), 902–911 (2005)

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

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

[13] J. Domingo-Ferrer, A. Solanas & A. Mat´_nez-Ballest´e, 2006 "Privacy in statistical databases: kanonymity through microaggregation", in IEEE Granular Computing' 06. Atlanta. USA, pp. 774-777. 118 Computer Science & Information Technology (CS & IT).

[14] Newton E., Sweeney L.,Malin B.: Preserving Privacy by De-identifying Facial Images. *IEEE Transactions on Knowledge and Data Engineering, IEEE TKDE*, February 2005.

[15] A. Hundepool, A. V. deWetering, R. Ramaswamy, L. Franconi, A. Capobianchi, P.-P. DeWolf, J. Domingo-Ferrer, V. Torra, R. Brand, and S. Giessing, *µ-ARGUS version 4.0 Software and User's Manual*. Voorburg NL: Statistics Netherlands, May 2005, http://neon.vb.cbs.nl/casc.

[16] Sweeney L.: Privacy-Preserving Bio-terrorism Surveillance. *AAAI Spring Symposium, AI Technologies for Homeland Security*, 2005

[17] Sweeney L.: Privacy Technologies for Homeland Security. *Testimony before the Privacy and Integrity Advisory Committee of the Deprtment of Homeland Scurity*, Boston, MA, June 15, 2005

[18] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam "*l*-diversity: Privacy beyond *k*-anonymit", In *Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE 2006)*, 2006

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

[20] Sweeney L.: AI Technologies to Defeat Identity Theft Vulnerabilities. *AAAI Spring Symposium, AI Technologies for Homeland Security*, 2005

[21]Benjamin C. M. Fung *Concordia University, Montreal*, Rui Chen *Simon Fraser University, Burnaby* and Philip S. Yu *University of Illinois at Chicago"* Privacy-Preserving Data Publishing: A Survey of Recent Developments" ACM Computing Surveys, Vol. 42, No. 4, Article 14, Publication date: June 2010

[22] Privacy-Preserving Data Mining, *Models and Algorithms* Edited by Charu C. Aggarwal *IBM T.J. Watson Research Center, USA* and Philip S. Yu *University of Illinois at Chicago, USA, Springer 2008*

[23] Ebaa Fayyoumi and B. John Oommen " A survey on statistical disclosure control and micro-aggregation techniques for secure statistical databases." *Softw. Pract. Exper.* 31 May 2010;

[24] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, CMU, SRI, 1998.

[25] Josep domingo-Ferrer, Agusti Solanas. "Privacy in Statistical Databases:k-Anonymity Through Microaggregation,"IEEE 2006

[26] Domingo-Ferrer, J., Sebé, F., & Solanas, A. (2008). A polynomial-time approximation to optimal multivariate microaggregation. Computer and Mathematics witt Applications, 55(4), 714–732.

[27] 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.

[28] Ebaa Fayyoumi and B. John Oommen "A survey on statistical disclosure control and micro-aggregation techniques for secure statistical databases" Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/spe.992

[29] Sweeney L., Gross R.: Mining Images in Publicly-Available Cameras for Homeland Security. *AAAI Spring Symposium, AI Technologies for Homeland Security*, 2005.

[30] Domingo-Ferrer, J., Martínez-Ballesté, A., Mateo-Sanz, J. M., & Sebé, F. Efficient multivariate data-oriented microaggregation. The VLDB Journal, 15(4), 355–369. (2006)