A Novel Method for Micro-Aggregation in Secure Statistical Databases Using Association and Interaction

B. John Oommen¹ and Ebaa Fayyoumi²

¹ *Chancellor's Professor; Fellow: IEEE and Fellow: IAPR.*

School of Computer Science, Carleton University, Ottawa, Canada: K1S 5B6
oommen@scs.carleton.ca

² School of Computer Science, Carleton University, Ottawa, Canada: K1S 5B6
efayyoum@scs.carleton.ca

Abstract. We¹ consider the problem of micro-aggregation in secure statistical databases, by enhancing the primitive Micro-Aggregation Technique (*MAT*), which incorporates proximity information. The state-of-the-art *MAT* recursively reduces the size of the data set by excluding points which are farthest from the centroid, and those which are closest to these farthest points, while it ignores the mutual Interaction between the records. In this paper, we argue that inter-record relationships can be quantified in terms of two entities, namely their “Association” and “Interaction”. Based on the theoretically sound principles of the neural networks (*NN*), we believe that the proximity information can be quantified using the mutual Association, and their mutual Interaction can be quantified by invoking transitive-closure like operations on the latter. By repeatedly invoking the inter-record Associations and Interactions, the records are grouped into sizes of cardinality “*k*”, where *k* is the security parameter in the algorithm. Our experimental results, which are done on artificial data and on the benchmark data sets for real-life data, demonstrate that the newly proposed method is superior to the state-of-the-art by as much as 13%.

Keywords: Information loss (*IL*), Micro-Aggregation Technique (*MAT*), Inter-record association, Interaction between micro-units.

1 Introduction

A lot of attention has recently been dedicated to the problem of maintaining the confidentiality of statistical databases through the application of statistical tools, so as to limit the identification of information on individuals and enterprises. Statistical Disclosure Control (*SDC*) seeks to balance between the confidentiality and the data utility criteria. For example, federal agencies and their contractors

¹ The first author is also an Adjunct Professor with the University of Agder, in Grimstad, Norway. This work was partially funded by NSERC, the Natural Sciences and Engineering Research Council of Canada.

who release statistical tables or micro-data files are often required by law or by established policies to protect the confidentiality of released information. However, this restriction should not affect public policy decisions which are made by only accessing non-confidential summary statistics [1]. *SDC* can be applied to information in several formats: Tables, responses to dynamic database queries and micro-data [2,3]. The protection provided by *SDC* results from either generating a set of synthetic data from a model fitted to the real data, or modifying the original data in a special way [3].

Micro-aggregation is one of the most recent techniques that has been used to mask micro-individuals in view of protecting them against the re-identification in secure statistical databases [2, 4, 5]. Moreover, it is modeled as a clustering mechanism with group size constraints, where the primitive goal is to group a set of records into clusters of size at least k , based on a proximity measure involving the variables of interest [6, 7, 8, 9, 10].

The Micro-Aggregation Problem (*MAP*), as formulated in [4, 5, 8, 9], can be stated as follows: A micro-data set $\mathcal{U} = \{U_1, U_2, \dots, U_n\}$ is specified in terms of the n “micro-records”, namely the U_i 's, each representing a data vector whose components are d continuous variables. Each data vector can be viewed as $U_i = [u_{i1}, u_{i2}, \dots, u_{id}]^T$, where u_{ij} specifies the value of the j^{th} variable in the i^{th} data vector. Micro-aggregation involves partitioning the n data vectors into m mutually exclusive and exhaustive groups so as to obtain a k -partition $\mathbb{P}_k = \{G_i \mid 1 \leq i \leq m\}$, such that each group, G_i , of size, n_i , contains either k data vectors, or between k and $2k - 1$ data vectors.

The optimal k -partition, $\mathbb{P}_k^*$, is defined to be the one that maximizes the within-group similarity, which is defined as the *Sum of Squares Error*, $SSE = \sum_{i=1}^m \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^T (X_{ij} - \bar{X}_i)$. This quantity is computed on the basis of the Euclidean distance of each data vector X_{ij} to the centroid $\bar{X}_i$ of the group to which it belongs. The *Information Loss* is measured as $IL = \frac{SSE}{SST}$, where SST is the squared error that would result if all records were included in a single group and is given $SST = \sum_{i=1}^m \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^T (X_{ij} - \bar{X})$, where $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$.

As mentioned in the literature, this problem in its multi-variate setting is known to be *NP-hard* [11], and has been tackled using different approaches such as hierarchical clustering [4, 5], genetic algorithms [4, 5, 12], graph theory [8, 9], fuzzy clustering [13, 14] and machine learning [15]. All the heuristic Micro-aggregation Techniques (*MATs*), seek to minimize the value of the *IL*. However, minimizing the loss in the data utility is an important issue that is difficult to enforce, primarily because this strategy was intended to enhance the security in an *SDC* technique. Indeed, the definition of optimality for an *SDC* is defined in the literature as being equivalent to offer the best trade-off between the *IL* and disclosure risk [16, 17]. In spite of this, the recent development of *MATs* [18] leaves the researcher no excuse to circumvent the problem of trying to reduce the value of the *IL* as much as possible [19].

In general, minimizing the *IL* directly follows maximizing the similarity between records in each group. The state-of-art *MATs* depend on utilizing the “Euclidean” distance, which serves as the criterion playing a central role in estimating

the similarity between the records. However, this distance function does not completely capture the appropriate notion of similarity for any data set. Our position is that the notion of similarity should be measured by using a metric that also unravels the relationship between the inter-records. We believe that this can be quantified in terms of two quantities, namely the mutual “Association” between the individual records and their mutual “Interaction”. We propose to measure these quantities using Association Similarity Rules² (*ASRs*). In this context, we mention that the concepts of Association and Interaction are derived from the Associative Cluster Neural Network (*ACNN*) [26].

The main contribution of this paper is to integrate the foundational concepts of *ASRs* with *MATs* so as to devise a new strategy for estimating the similarity. This new method demonstrates that the *IL* can be reduced taking two measurements into consideration. First of all, we consider the mutual Association between the records. Secondly, and more importantly, we also consider the mutual Interaction between the records by using a transitive-closure like operation when $k \geq 3$. This, in turn, is achieved by invoking our newly proposed Interactive-Associative Micro-Aggregation Technique (*IAMAT*). The effect of these considerations can be seen to minimize the *IL* by up to 13% when compared to the state-of-the-art.

The structure of this paper is as follows: In the following section 2 we start with a concise survey about the reported *MATs*. Subsequently, we summarize the Associative Cluster Neural Network algorithm. In Section 3, the Interactive-Associative Micro-Aggregation Technique is presented informally and algorithmically. Then, in Section 4, we present the results of experiments we have carried out for synthetic and real data sets. The paper finishes in Section 5 with some conclusions.

2 Micro-Aggregation

As mentioned in Section 1, the *MAP* has been tackled using different techniques. Basically, a *MAT* relies on a clustering technique and an aggregation technique. *MATs* were originally used for numerical data [10], and they can be further classified as below.

- *Uni-variate vs. Multi-variate*

The difference between the uni-variate and the multi-variate *MATs* depends on the number of random variables used in the micro-aggregation process. Uni-variate *MATs* deal with multi-variate data sets by micro-aggregating one variable at a time, such as Individual ranking [27, 28]). Multi-variate *MATs* either rank multi-variate data by projecting them onto a single axis³, or dealing directly with the unprojected data. Working on unprojected multi-variate data allows simultaneous micro-aggregation of several variables so

² Association Similarity Rules are well-known data mining techniques used to discover the relationships between patterns in different application domains [20,21,22,23,24,25].

³ The multi-variate data is projected onto a single axis by using either a particular variable, the sum-z-scores or a principle component analysis prior to micro-aggregation [5].

that a single k -partition for the entire data set is obtained. Since there is no straightforward formal algorithm to sort multi-variate data without projection [4, 29, 30], many heuristic methods have been proposed such as the Maximum Distance to Average Vector (*MDAV*) [4, 18], the Minimum Spanning Tree (*MST*) [9] and the Object Migrating Micro-aggregated Automaton (*OMMA*) [15].

– *Fixed-size vs. Data-oriented*

The difference between the fixed-size and the data-oriented *MATs* depends on the number of records in each group. Fixed-size *MATs* require all groups to be of size k except for a single group whose cardinality is greater than k when the total number of records, n , is not a multiple of k . The last group is used to contain more than k records [30], but as pointed out in [4], assigning the additional records to the group containing the modal value of the data, reduces the value of the *IL* [29]. Data-oriented *MATs* allow groups to be of size greater than k and less than $2k - 1$ depending on the structure of the data. These methods yield more homogenous groups, which help to further minimize the *IL* [2, 4, 5]. Although these methods are marginally more complex than those involving fixed-size *MATs*, they are less likely to compromise the “*privacy*” of the micro-file as shown in [31]. Examples of data-oriented *MATs* are those which use a genetic algorithm [4, 5], the k -Ward *MAT* [4, 32, 5] and the Variable-size Maximum Distance to Average Vector scheme ($V - MDAV$) [33].

– *Optimal vs. Heuristic*

The first reported optimal uni-variate *MAT* with a polynomial complexity is given in [8], which solves the *MAP* as a shortest path problem on a graph. Unfortunately, the optimal *MAP* for multi-variate micro-aggregation is an *NP*-hard problem [11]. Therefore, researchers seek heuristic *MATs* that provide a good solution - close to the optimal.

2.1 Associative Clustering Neural Network (*ACNN*)

The Associative Cluster Neural Network (*ACNN*), was proposed [26] as a recurrent *NN* model that dynamically evaluates the Association of any pair of patterns through the Interaction between them and the group of patterns. The *ACNN* possesses many attractive features, such as its simple structure, its respective learning mechanism, and its efficient clustering strategy, which uses the Association as a new similarity measure. Its superiority in clustering and analyzing gene expression data has also been demonstrated [34]. The rationale behind this superiority probably lies in the inherent advantages of *ASRs*, which possess the potential to ensure that the similarities between patterns within the same cluster increase, whereas the similarities between different clusters decrease.

The *ACNN* initializes the Association between any two neurons by evaluating the relationship between them and by setting the learning ratio, α , to the most suitable value. The learning ratio should guarantee that the initial Association is large when the distance between the patterns is small. The *ACNN* studies the Interaction level of each pair of patterns based on the Association made by the other patterns, and defines the similarity threshold which ensures a robust performance.

The association value between any two patterns has to be updated based on the result of the Interaction level, and this is, in turn, scaled by using the well-known sigmoidal function. This procedure has to be iteratively executed until there is no noticeable change in the successive associations. Subsequently, the *ACNN* constructs the cluster characteristic matrix to describe the cluster property at the end of the learning phase, after which it determines the number of clusters and labels the patterns with the index of the cluster that they belong to. We refer the interested reader to [26] for a complete explanation of this clustering strategy.

3 Interactive-Associative Micro-Aggregation Technique (*IAMAT*)

The state-of-the-art *MATs* use a proximity function, and in particular, the Euclidean distance, to measure the similarity between records. To the best of our knowledge, the combination of the Association *and* the Interaction between individual records has not been taken into consideration while micro-aggregating the data file. We now discuss how these two criteria are applicable to micro-aggregate the data file so as to further minimize the *IL*.

3.1 Inadequacy of Using the *ACNN* Directly

Although the basic *ACNN* dynamically evaluates the Association and the Interaction between the patterns, it is not directly applicable to the *MAP* in its virgin form for the following reasons:

- **Feature Values**

In a neural setting, the weights of the neurons are updated based on the relative relationship between them. These weights are usually updated by gradient-like considerations, so that a change in the weights leads to a better classifications. Consequently, in the *ACNN* the weights could be both positive or negative, quite simply because increasing the values of certain features may have a negative impact on the optimization problem.

As opposed to this, it is meaningless to have weights that are negative in the *MAP*. This is because the fundamental reason for tackling the problem is to determine how the records are associated with each other, and clearly, the concept of the records being negatively associated is irrelevant. Thus, if we are to use the principles of the *ACNN* to solve the *MAP*, it is imperative that the weights are never allowed to become negative. Rather, we would prefer that they stay within the interval $[0, 1]$.

- **Ineffectiveness of Sigmoidal Mappings**

When Minsky suggested the weakness of the Perceptron, he showed that it was incapable of resolving the basic *XOR* problem. However, the field of *NNs* received a huge “boost” when it was discovered that if these primitive neural units were cascaded and interconnected, the discriminant could be arbitrarily complex. To effectively model the switching and clipping effects in such complex domains, researchers introduced functions such as the

sigmoidal function whose purpose was to transform the input space using highly non-linear mappings.

It is our position that such switching and clipping effects are not pertinent to the study of the *MAP*. The reason for this is: The Associations and the Interactions between the records are, in one sense, related to their relative proximity, and we have no reason to believe that these quantities fall off or change *abruptly*. Rather, our experience is that these quantities vary smoothly with the relative proximity of the records.

– Transitive-Closure-like Properties

Obtaining the set of shortest paths on a graph can be achieved by using a transitive-closure algorithm that traverses all the edges of the graph. In this case, the shortest paths are obtained by using the operation of “Addition” on the weights of the edges along any given path, and invoking the “Minimum” operation over all the possible paths. However, the underlying algorithm has been proven to be more powerful if it is mapped using the properties of a semi-ring $(S, \oplus, \otimes)$, where (i) S is the set of weights associated with an edge, (ii) $\oplus$ represents an abstract “Addition” operation over the elements of S , and (iii) $\otimes$ represents an abstract “Multiplication” operation over the elements of S . In particular, if S is the set of reals, and $\oplus$ and $\otimes$ represent the arithmetic addition and product operations respectively, the transitive-closure algorithm would lead to a matrix multiplication scheme, which is central in determining the multi-step Markov matrix for a Markov chain.

The basic *ACNN* computes the Interaction between the neurons using the product involving a_{ip} and a_{pj} . The total two-step Interaction is thus, effectively, the contribution of the transitive-closure operation of the path from X_i to X_j via all the possible intermediate nodes, X_p . In our case, the issue of interest is *not* the total Interaction between the relevant nodes, but rather the node X^* which contributes to the maximal Interaction between X_i and X_j . Thus, unlike the *ACNN*, in our solution, we do not compute the sum of all the Interactions between the nodes. Rather, we report the one which is maximally interacting with the nodes already in the same cluster, say X_i and X_j . This is a fundamental difference between our present scheme and the *ACNN*, because it renders the computations both easier and faster, and is yet able to coalesce the nodes based on the inferred interactions.

– One-shot Training

The final difference between our present scheme and the *ACNN* is the fact that we have resorted to a one-shot “training” mechanism. This is *atypical* for most *NNs*. Indeed, most *NNs* repeatedly run the *NN* over the data set till their respective weights converge. Some families of *NNs* (for example, the Adachi’s network [35]) have been reported, which actually yield the final weights using a single pass over the data set.

In our case, we argue that repeatedly running the updating algorithm over the data set is superfluous. Rather, by initially computing the Associations, we are able to arrive at the best Interactions. The *ACNN* requires that the set of associations are then re-computed. But, since these associations are computed based on the relative proximities of the records, and since

the Interactions are computed based on the latter, it is meaningless, in the case of the MAP, to re-compute the Associations. Indeed, if we resorted to doing this, it would lead to weights that are negative and which again, as argued above, is unacceptable. It would also lead to the “rejection” of many records- which is inappropriate for the MAP. Thus, in the IAMAT, the corresponding matrices are computed in a one-shot manner. Subsequent computations are required only after the learned groups of size k are removed in each learning cycle.

Based on the above principles, we now present the design of our newly-proposed scheme, the Interactive-Associative Micro-Aggregation Technique (IAMAT).

3.2 Design of the IAMAT

We propose IAMAT to micro-aggregate the records in the data set by using a new methodology to evaluate the similarity between them. This similarity is intuitively expressed by their inter-record relationships, and is estimated by measuring the “Association” and “Interaction” as *modeled* in the ACNN. The resulting measurements are similar to the ones that cluster the records based on the distance between them. Consequently, instead of merely assigning relatively “close” records to be in the same group, we choose to “estimate” the Association *and* the Interaction between them, and if the combination of these indexes is relatively high, we assign them to be in the same group. Otherwise, we determine that they should be in two different groups. We believe that using this pair of measurements will help to achieve a more robust performance than other existing measures, which is a claim that we have verified. From a top level, we can describe it as below.

The IAMAT is a consequence of incorporating the above considerations into the elegant MDAV strategy. Consider the IAMAT for any specific value of k . The IAMAT uses the centroid of the data set to relatively determine the farthest record, say X_r . Subsequently, we achieve a quick search to obtain the record that is most associated to X_r , say X_s . After this, we propose to choose $k - 2$ records based on the mutual *Interaction* between each record inside the group and the remaining unassigned records. Consequently, the next step consists of creating a cluster that comprises the associated pair $\langle X_r, X_s \rangle$ and the most interactive $k - 2$ records. At the end of this stage, the cluster is micro-aggregated and removed from the original data set. The above steps are iteratively repeated until no more than $k - 1$ records remain in the original data set. The IAMAT terminates by assigning the remaining unassigned records to the last group. The scheme is algorithmically described below in Algorithm 1, after which each step is explained in greater detail.

Unlike the MDAV, instead of measuring the distance between the records, the IAMAT utilizes the Association as per the ACNN. The ACNN classifies the records as being associated if the value of the association index, a_{ij} , is positive. Otherwise the neurons will be classified as being unrelated, leading to its “rejection”. Clearly, rejecting records will not comply with the spirit and goal of the MAP whose aim is to minimize the IL. We believe that an Association

Algorithm 1. Interactive-Associative Micro-Aggregation Technique (*IAMAT*)

Input: The original micro-data file, $\mathcal{D}$, that contains n unassigned records, and the parameter, k .

Output: The micro-aggregated micro-data file, $\mathcal{D}'$.

Method:

- 1: Compute the centroid of $\mathcal{D}$ as $\mu = \frac{1}{n} \sum_{i=1}^n X_i$.
- 2: Compute the scaling factor α as related to the mean square distance as $\alpha = \frac{\sqrt{n}}{\frac{1}{n} (\sum_{i=1}^n \|X_i - \mu\|^2)}$.
- 3: Compute the association values between μ and each record, X_i , in $\mathcal{D}$ as $a_{\mu i} = e^{-\frac{\|X_i - \mu\|^2}{\alpha}}$.
- 4: Initialize the number of groups to zero.
- 5: **while** there are more than $(k - 1)$ *Unassigned* records in $\mathcal{D}$ **do**
- 6: Increment the number of groups by unity.
- 7: Initialize the number of records inside the group to zero.
- 8: Select the least associated *Unassigned* record, X_r , to the centroid μ as follows $X_r = \text{Min } a_{\mu i}$.
- 9: Mark X_r as *Assigned* record.
- 10: Compute the association values between X_r and each *Unassigned* record, X_i , in $\mathcal{D}$.
- 11: Select the most associated *Unassigned* record, X_s , to X_r as follows $X_s = \text{Max } a_{ri}$.
- 12: Mark X_s as an *Assigned* record.
- 13: Compute the association values between X_s and each *Unassigned* record, X_i , in $\mathcal{D}$.
- 14: Add X_r and X_s to the group and increment the number of records inside the group by two units.
- 15: **while** the number of records inside the group is less than k **do**
- 16: **for all** *Unassigned* records, X_p , in $\mathcal{D}$ **do**
- 17: Initialize the Interaction of X_p , η_p , to 1.
- 18: **for all** *Assigned* records inside the group, X_i **do**
- 19: Update the value of Interaction as follows $\eta_p = \eta_p * a_{ip}$.
- 20: **end for**
- 21: **end for**
- 22: Let X^* be the record which has the highest value for η_p .
- 23: Mark X^* as an *Assigned* record.
- 24: Add X^* into this group and increment the number of records inside the group by unity.
- 25: Compute the association values between the most interactive record, X^* and each *Unassigned* record, X_i , in $\mathcal{D}$.
- 26: **end while**
- 27: Remove the present cluster from the set $\mathcal{D}$.
- 28: **end while**
- 29: Assign the remaining *Unassigned* records to the last group.
- 30: Build the micro-aggregated data file, $\mathcal{D}'$.
- 31: **return** $\mathcal{D}'$.
- 32: **End Algorithm** Interactive-Associative Micro-Aggregation Technique (*IAMAT*)

between any pair of records exists regardless of its value, and this could be very small when it is close to zero, or very large when it is close to unity. Therefore, the *IAMAT* quantifies the value of the Association between two records, say X_i and X_j , to belong to the interval $[0, 1]$, and this is computed as follows:

$$a_{ij} = a_{ji} = r(X_i, X_j) = e^{-\frac{\|X_i - X_j\|^2}{\alpha}}.$$

where $r()$ is the identical function used in the definition of the *ACNN*, which evaluates the relationship between any two records, and which also involves α . The value of α is assigned so as to guarantee that the initial Association is large when the distance between X_i and X_j is small and vice versa. It is given as:

$$\alpha = \frac{\sqrt{n}}{\frac{1}{n}(\sum_{i=1}^n \|X_i - \frac{1}{n}(\sum_{i=1}^n X_i)\|^2)}.$$

The rationale for incorporating the Association with the Interaction between the records inside a group, is that it leads to more homogeneous groups. The concept of the *Interaction* turns out to be crucial in forming the cluster, because we believe that merely being close to the farthest records is not a reason that is sufficiently important for any record to be grouped with the most distant one. Rather, we propose that the Interaction with respect to all the records inside the group has to be taken into consideration while clustering the records. As mentioned earlier, the latter is computed by invoking transitive-closure like operations. Finding the most interactive record with the associated pair is achieved by searching for the maximum product of the Association between the unassigned records, say X_p , and each record in the associated pair, $\langle X_i, X_j \rangle$, as follows:

$$\eta_{ij} = \begin{cases} a_{ip}(t-1) \times a_{pj}(t-1) & p \neq i, j \text{ and } i \neq j \\ 0 & i = j. \end{cases}$$

The equation above is valid when $k = 3$. By increasing the value of k , the transitive-closure is applied by adding one unassigned record at a time. The decision of grouping the unassigned record with other records in the group depends on the Interaction of that record with respect to other records inside the group. Logically, the most interactive unassigned record has been chosen as follows:

$$\text{Index Maximum}_{1 \leq p \leq n} \eta_p,$$

where η is computed as follows:

$$\eta_p = \prod_{i=1}^{n_i} a_{ip}$$

where X_i represents the record inside the group, G_j , of size n_j and X_p represents the unassigned record.

4 Experimental Result

4.1 Data Sets

The *IAMAT* has been rigorously tested and the results obtained seem to be very good, where the “goodness” of a scheme refers to the combination of its

being efficiently computed, and its leading to a minimum value for the IL . We have tested it using the two real-life benchmark reference data sets used in previous studies: (i) The *Tarragona* data set which contains 834 records with 13 variables [4]. (ii) The *Census* data set which contains 1080 records with 13 variables [36].

To further investigate the performance of the new scheme, many experiments have been carried out using various simulated data sets involving vectors with dimensions ranging from 10 up to 80, and sets of cardinality from 10,000 up to 100,000. The simulated multi-variate data sets were generated using *Matlab*'s built-in-functions: (i) Uniform distribution ($\mu=0$; $\max=1000$). (ii) Normal distribution ($\mu=0$; $\sigma=0.05$).

The experiments were also used to investigate the scalability of the *IAMAT* with respect to the size of the data, its dimensionality, and the group size.

4.2 Results

For a given value of the security parameter k , which represents the minimum number of records per group, we compared the percentage value of the $IL = (SSE/SST)$ times 100 (as defined in Section 1) resulting from the *IAMAT* and the *MDAV* strategies. It is important to mention that the *MDAV* was implemented based on the centroid concept and not a diameter concept.⁴ All the programs were written in the *C++* language, and the tests were performed on an Intel(R) Pentium (R)M. Processor 1.73 GHz., with 512 MB of RAM.

Table 1 shows the improvement of the solution obtained by using the *IAMAT* as opposed to the *MDAV* on the multi-variate real data sets, where all the 13 variables are used simultaneously during the micro-aggregation process. The reduction in the value of the IL attains up to 8% on the *Tarragona* data set, and 5.12% on the *Census* data set when the group size is equal to 3. It is clearly evident that the impact of the group size on the solution is minimized by increasing the number of records per group. To be fair, we also mention that computational time required to execute the *IAMAT* is almost double the computational time required for the *MDAV*, although, in every case, the time was less than 0.5 second. In term of comparison, we believe that minimizing the loss in the data utility is more important than minimizing the extremely small computational time, especially because the micro-aggregation is usually performed off-line where the additional time requirement is less crucial. However, the question of how the decrease of IL is related to the increase in the computational time is still open.

We undertook a comprehensive evaluation of the performance of the *IAMAT* scheme so as to investigate the it scalability with respect to the size of the data set, its dimensionality and the group size as shown in Table 2 in Appendix A.

– The scalability of the *IAMAT* with respect to the data set size

We tested both the *IAMAT* and the *MDAV* schemes using data based on the uniform and the normal distributions with cardinalities ranging from

⁴ We did not program the *MDAV* scheme. We are extremely thankful to Dr.Francesc Seb for giving us his source code.

Table 1. Comparison of the percentage of the *IL* and the computational time between the *MDAV* and the *IAMAT* on the Tarragona and Census data sets

Data Set	k value	<i>MDAV</i>		<i>IAMAT</i>		Improv. (%)
		IL	Time	IL	Time	
Tarragona	3	16.9593	0.17	15.6023	0.31	8.00
	4	19.7482	0.12	19.2872	0.22	2.33
	5	22.8850	0.12	22.7164	0.23	0.74
Census	3	5.6535	0.22	5.3639	0.41	5.12
	4	7.4414	0.19	7.2170	0.44	3.02
	5	8.8840	0.17	8.8428	0.42	0.46
	6	10.1941	0.17	9.9871	0.42	2.03

10,000 records up to 100,000 with 10 variables. The percentage of the improvement achieved by invoking the *IAMAT* in the *IL*, (when $k = 4$) reached as high as 13.73% for the normal distribution and 13.67% for the uniform distribution. It is fair to state that the *IAMAT* requires almost triple the computational time needed to execute the *MDAV* scheme. In general, increasing the size of the data set tends to minimize the *IL* value.

– **The scalability of the *IAMAT* with respect to dimensionality**

We also tested the *IAMAT* and the *MDAV* on the uniform and the normal distributions for various dimensions ranging from 10 to 80, when the data size was set to 10,000 records, and the value of k was set to 3. The highest percentage of the improvement in the *IL* was about 10% for both the uniform and normal distributions. The computational time required to micro-aggregate all the individual records was directly proportional to the dimensionality in both schemes, and the required computational time in the *IAMAT* was almost double the required time for the *MDAV*. As expected, increasing the dimensionality implies increasing the *IL*. This is intuitively appealing because increasing the dimensionality tends to minimize the similarity between the individual records and, at the same time, reduce the Association and the Interaction between the different multi-variate records.

– **The scalability of the *IAMAT* with respect to the group size**

The scalability of the *IAMAT* and the *MDAV* with regard to the group size was studied for both the uniform and the normal distributions, where the group size ranged from 3 to 10, and for the cardinality of the data set being 10,000 records, with a dimensionality of 10 variables. The percentage of improvement in reducing the value of the *IL*. This reduction reached 12.74% for the normal distribution when the group size was 5, and reached 12.31% for the uniform distribution when the group size was 4. This is, again, a fair observation, because having many records in the group tends to minimize the within-group similarity, and to simultaneously maximize the similarity between the groups. This also tends to increase the *IL* value.

5 Conclusions

In this paper, we have considered the problem of achieving micro-aggregation in secure statistical databases. The novelty of our method involves enhancing the primitive *MAT* that merely incorporates proximity information. The state-of-the-art *MAT* recursively reduced the size of the data set by excluding points which were farthest from the centroid and those which were closest to these farthest points. Thus, although the state-of-the-art method was extremely effective, we have argued that it uses only the proximity information, and ignores the mutual Interaction between the records. In this paper, we have proved that inter-record relationships can be quantified in terms of two entities, namely their “Association” and “Interaction” that can be measured by invoking transitive-closure like operations, and by mapping the problem into a neural setting using the *ACNN*. By repeatedly invoking the inter-record Associations and Interactions, we have shown that the records can be grouped into sizes of cardinality “ k ”. Our experimental results, which were done on artificial data and on the benchmark data sets for real life data, demonstrate that the newly proposed method is superior to the state-of-the-art by as much as 13%. Thus, we believe that our strategy leads to a very promising tool for solving the *MAP*.

We foresee two avenues for future work. The first avenue is to extend the *IAMAT* towards data-oriented micro-aggregation, where the group size, n_i , satisfies $k \leq n_i < 2k$. The second involves investigating the effect of having a dynamic α ⁵ on the compactness of each group and on the value of the *IL*.

Acknowledgements

The authors would like to thank Dr. Josep Domingo-Ferrer for all his support and advice, and for providing us with the data sets. We are also thankful to him, Dr. Josep Mateo-Sanz, and Dr. Francesc Sebé for answering our queries, and for providing the code for the *MDAV* scheme. Without their help, we would not have been able to produce these results.

References

1. Adam, N., Wortmann, J.: Security-Control Methods for Statistical Databases: A Comparative Study. *ACM Computing Surveys* 21(4), 515–556 (1989)
2. Cuppen, M.: Secure Data Perturbation in Statistical Disclosure Control. PhD thesis, Statistics Netherlands (2000)
3. Willenborg, L., Waal, T.: Elements of Statistical Disclosure Control (ILL Number: 2132712). Springer, Heidelberg (2001)

⁵ We believe that we can compute the dynamic value of α in any iteration based on the remaining *Unassigned* records at the end of every iteration in the clustering phase. As opposed to this the fixed value is computed once before invoking the clustering phase, and is based on the entire data set.

4. Domingo-Ferrer, J., Mateo-Sanz, J.: Practical Data-Oriented Microaggregation for Statistical Disclosure Control. *IEEE Transactions on Knowledge and Data Engineering* 14(1), 189–201 (2002)
5. Mateo-Sanz, J., Domingo-Ferrer, J.: A Method for Data-Oriented Multivariate Microaggregation. In: *Proceedings of Statistical Data Protection98*, Luxembourg: Office for Official Publications of the European Communities, pp. 89–99 (1999)
6. Domingo-Ferrer, J.: Statistical Disclosure Control in Catalonia and the CRISES Group. Technical report (1999)
7. Domingo-Ferrer, J., Torra, V.: Aggregation Techniques for Statistical confidentiality. In: *Aggregation operators: new trends and applications*, Germany: Heidelberg, Physica-Verlag GmbH, pp. 260–271 (2002)
8. Hansen, S., Mukherjee, S.: A Polynomial Algorithm for Univariate Optimal Microaggregation. *IEEE Transactions on Knowledge and Data Engineering* 15(4), 1043–1044 (2003)
9. Laszlo, M., Mukherjee, S.: Minimum Spanning Tree Partitioning Algorithm for Microaggregation. *IEEE Transactions on Knowledge and Data Engineering* 17(7), 902–911 (2005)
10. Torra, V.: Microaggregation for Categorical Variables: A Median Based Approach. In: Domingo-Ferrer, J., Torra, V. (eds.) *PSD 2004*. LNCS, vol. 3050, pp. 162–174. Springer, Heidelberg (2004)
11. Oganian, A., Domingo-Ferrer, J.: On The Complexity of Optimal Microaggregation for Statistical Disclosure Control. *Statistical Journal of the United Nations Economic Commission for Europe* 18(4), 345–354 (2001)
12. Solanas, A., Martinez-Balleste, A., Mateo-Sanz, J., Domingo-Ferrer, J.: Multivariate Microaggregation Based Genetic Algorithms. In: *3rd International IEEE Conference on Intelligent Systems*, pp. 65–70 (2006)
13. Domingo-Ferrer, J., Torra, V.: Fuzzy Microaggregation for Microdata Protection. *Journal of Advanced Computational Intelligence and Intelligent Informatics (JACIII)* 7(2), 153–159 (2003)
14. Torra, V., Domingo-Ferrer, J.: Towards Fuzzy C-Means Based Microaggregation. In: Grzegorzewski, P., Hryniewicz, O., Gil, M. (eds.) *Advances in Soft Computing: Soft Methods in Probability, Statistics and Data Analysis*, Germany: Heidelberg, Physica-Verlag, pp. 289–294 (2002)
15. Fayyouni, E., Oommen, B.: A Fixed Structure Learning Automaton Micro-Aggregation Technique for Secure Statistical Databases. In: *Privacy Statistical Databases*, Italy: Rome, pp. 114–128 (2006)
16. Crises, G.: Trading off Information Loss and Disclosure Risk in Database Privacy Protection. Technical report (2004)
17. Mateo-Sanz, J., Domingo-Ferrer, J., Seb, F.: Probabilistic Information Loss Measures in Confidentiality Protection of Continuous Microdata. *Data Mining and Knowledge Discovery* 11(2), 181–193 (2005)
18. Domingo-Ferrer, J., Torra, V.: Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation. *Data Mining and Knowledge Discovery* 11(2), 195–212 (2005)
19. Domingo-Ferrer, J., Seb, F.: Optimal multivariate 2-microaggregation for microdata protection: a 2-approximation. In: *Privacy Statistical Databases*, Italy: Rome, pp. 129–138 (2006)
20. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules Between Sets of Items in Large Databases. In: *Proceedings of ACM SIGMOD*, USA: Washington DC, pp. 207–216 (1993)

21. Agrawal, R., Mannila, H., Srikant, H., Toivonen, R., Verkamo, I.: Fast Discovery of Association Rules. In: *Advances in Knowledge Discovery and Data Mining*, pp. 307–328 (1996)
22. Bacao, F., Lobo, V., Painho, M.: Self-organizing Maps as Substitutes for K-Means Clustering. In: *International Conference on Computational Science*, pp. 476–483 (2005)
23. Feng, L., Dillon, T., Weigana, H., Chang, E.: An XML-Enabled Association Rule Framework. In: Mařík, V., Štěpánková, O., Retschitzegger, W. (eds.) *DEXA 2003. LNCS*, vol. 2736, pp. 88–97. Springer, Heidelberg (2003)
24. Markey, M., Lo, J., Tourassi, G.: Self-Organizing Map for Cluster Analysis of A Breast Cancer Database. *Artificial Intelligence in Medicine* 27, 113–127 (2003)
25. Park, J., Chen, S., Yu, P.: Using A Hash-Based Method with Transaction Trimming for Mining Association Rules. *IEEE Transactions on Knowledge and Data Engineering* 9, 813–826 (1997)
26. Yao, Y., Chen, L., Chen, Y.: Associative Clustering for Clusters of Arbitrary Distribution Shapes. *Neural Processing Letters* 14, 169–177 (2001)
27. Defays, D., Anwar, M.: Masking Micro-data Using Micro-Aggregation. *Journal of Official Statistics* 14(4), 449–461 (1998)
28. Defays, D., Anwar, N.: Micro-Aggregation: A Generic Method. In: *Proceedings of the 2nd International Symposium on Statistical Confidentiality*, Luxembourg: Office for Official Publications of the European Communities, pp. 69–78 (1995)
29. Domingo-Ferrer, J., Mateo-Sanz, J., Oganian, A., Torra, V., Torres, A.: On The Security of Microaggregation with Individual Ranking: Analytical Attacks. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10(5), 477–491 (2002)
30. Mas, M.: Statistical Data Protection Techniques. Technical report, Eustat: Euskal Estatistika Erakundea, Instituto Vasco De Estadística (2006)
31. Li, Y., Zhu, S., Wang, L., Jajodia, S.: A privacy-enhanced microaggregation method. In: Eiter, T., Schewe, K.-D. (eds.) *FoIKS 2002. LNCS*, vol. 2284, pp. 148–159. Springer, Heidelberg (2002)
32. Fayyumi, E., Oommen, B.: On Optimizing the k -Ward Micro-Aggregation Technique for Secure Statistical Databases. In: *11th Australasian Conference on Information Security and Privacy Proceeding*, Australia: Melbourne, pp. 324–335 (2006)
33. Solanas, A., Martínez-Ballesté, A.: V-MDAV: A Multivariate Microaggregation With Variable Group Size. In: *17th COMPSTAT Symposium of the IASC*, Rome (2006)
34. Yao, Y., Chen, L., Goh, A., Wong, A.: Clustering Gene Data via Associative Clustering Neural Network. *The 9th International Conference on Neural Information Processing (ICONIP 2002)* 5, 2228–2232 (2002)
35. Adachi, M., Aihara, K.: Associative Dynamics in a Chaotic Neural Network. *Neural Networks* 10(1), 83–98 (1997)
36. Domingo-Ferrer, J., Torra, V.: A Quantitative Comparison of Disclosure Control Methods for Microdata. In: Doyle, P., Lane, J., Theeuwes, J., Zayatz, L. (eds.) *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, Amsterdam: North-Holland, Berlin, pp. 113–134. Springer, Heidelberg (2002)

Appendix A

Table 2. Comparison of the percentage of the *IL* and the computation time between the *MDAV* and the *IAMAT* on simulated data. The results demonstrate the scalability of the *IAMAT* with respect to the size of the data set, dimensionality and the group size.

Investigate scalability with respect to the size of the data set										
Data size	Normal Distribution					Uniform Distribution				
	<i>MDAV</i>		<i>IAMAT</i>		improv. (%)	<i>MDAV</i>		<i>IAMAT</i>		improv. (%)
	IL	Time	IL	Time		IL	Time	IL	Time	
10	13.9376	12.05	12.2106	29.37	12.39	13.9202	12.24	12.2069	29.96	12.31
20	11.9354	48.90	10.3653	120.37	13.15	11.9617	63.15	10.4275	145.09	12.83
30	10.8849	111.84	9.4488	277.20	13.19	10.9228	143.17	9.4976	328.02	13.05
40	10.2238	200.42	8.8732	494.94	13.21	10.2733	222.74	8.9063	526.37	13.31
50	9.7234	315.04	8.4382	773.92	13.22	9.7694	389.45	8.4574	903.39	13.43
60	9.3513	457.50	8.1119	1,115.69	13.25	9.3707	644.32	8.1338	1386.01	13.20
70	9.0358	623.89	7.8243	1,520.94	13.41	9.0776	629.73	7.8523	1500.12	13.50
80	8.7898	809.397	7.6063	1,981.06	13.46	8.8183	1002.77	7.6247	1959.25	13.54
90	8.5584	1,044.77	7.4010	2,505.86	13.52	8.5885	1252.61	7.4140	2485.42	13.68
100	8.3799	1,512.68	7.2293	3,413.98	13.73	8.3933	1656.90	7.2456	3546.72	13.67
Investigate scalability with respect to the dimensionality of the data set										
Number of variables	Normal Distribution					Uniform Distribution				
	<i>MDAV</i>		<i>IAMAT</i>		improv. (%)	<i>MDAV</i>		<i>IAMAT</i>		improv. (%)
	IL	Time	IL	Time		IL	Time	IL	Time	
10	10.5383	13.53	9.4821	30.14	10.02	10.5437	13.75	9.4400	33.87	10.47
20	24.3766	24.30	22.3577	41.26	8.28	24.4692	25.03	22.2454	46.40	9.09
30	32.6044	35.50	29.9938	69.79	8.01	32.5113	36.67	29.9711	60.00	7.81
40	37.5679	47.00	34.8732	64.98	7.17	37.4627	47.18	34.8052	72.68	7.09
50	40.8705	40.87	38.2173	67.87	6.49	40.8427	58.78	38.2470	85.98	6.36
60	43.3104	70.01	40.7761	89.88	5.85	43.3912	69.48	40.8478	98.81	5.86
70	45.3212	82.78	42.7483	104.6	5.68	45.3778	81.45	42.8263	112.48	5.62
80	46.8119	93.78	44.3459	113.17	5.27	46.8533	97.10	44.4367	132.42	5.16
Investigate scalability with respect to the group size										
k value	Normal Distribution					Uniform Distribution				
	<i>MDAV</i>		<i>IAMAT</i>		improv. (%)	<i>MDAV</i>		<i>IAMAT</i>		improv. (%)
	IL	Time	IL	Time		IL	Time	IL	Time	
3	10.5383	13.53	9.4821	32.96	10.02	10.5437	13.75	9.4400	33.87	10.47
4	13.9376	11.93	12.2106	28.79	12.39	13.9202	12.24	12.2069	29.96	12.31
5	16.5221	11.17	14.4167	26.67	12.74	16.4020	11.45	14.4091	29.32	12.15
6	18.7797	10.68	16.4650	26.86	12.33	18.5484	10.96	16.4152	28.78	11.50
7	20.3782	10.32	18.0344	25.48	11.50	20.2887	10.56	18.0431	28.28	11.07
8	21.6869	10.12	19.1562	26.67	11.67	21.5464	10.31	19.1668	26.32	11.04
9	22.8931	9.87	20.4255	24.92	10.78	22.7587	10.07	20.3952	26.02	10.39
10	23.8439	9.07	21.4969	26.06	9.84	23.6922	9.92	21.4520	25.92	9.46