A Variable-MDAV-Based Partitioning Strategy to Continuous Multivariate Microaggregation with Genetic Algorithms

Agusti Solanas, *Member, IEEE*, Úrsula González-Nicolás and Antoni Martínez-Ballesté

***Abstract*—Microaggregation is a Statistical Disclosure Control (SDC) technique that aims at protecting the privacy of individual respondents before their data are released.**

Optimally microaggregating multivariate data sets is known to be an NP-hard problem. Thus, using heuristics has been suggested as a possible strategy to tackle it. Specifically, Genetic Algorithms have been shown to be serious candidates that can find good solutions on small data sets. However, due to the very nature of these algorithms and the coding of the microaggregation problem, GA can hardly cope with large data sets. In order to apply them to large data sets, the latter have to be previously partitioned into smaller disjoint subsets that the GA can handle.

In this article we summarise several proposals for partitioning data sets, in order to use GA to microaggregate them. In addition, we suggest a new partitioning strategy based on the variable-MDAV algorithm, and we compare it with the most relevant previous proposals. The experimental results show that our method outperforms the previous ones in terms of information loss.

I. Introduction

Statistical agencies, private companies and Internet search engines are just a few examples of entities that daily collect data from many people. Our habits, tastes, and hobbies are analysed by an unprecedented amount of people and machines. Most countries have legislation which compels national statistical agencies to guarantee statistical confidentiality when they release data collected from citizens or companies; see [1] for regulations in the European Union, [2] for regulations in Canada, and [3] for regulations in the United States. Thus, protecting individual privacy is a key issue for many institutions, namely statistical agencies, Internet companies, manufacturers, etc; and many efforts have been devoted to develop techniques that guarantee some degree of personal privacy. Nowadays, these efforts come from a variety of fields, namely cryptography, data mining, statistics, artificial intelligence, and so on. Notwithstanding, the field of statistical disclosure control (SDC) was the one that initially considered the problem; firstly on tabular data, and more recently on microdata (*i.e.* the personal data collected from individual respondents). In this article we focus on the protection of microdata, that is, the protection of individual respondents from re-identification through the released microdata.

With the aim to protect from re-identification of individual respondents, microdata sets are properly modified prior to their publication. The degree of modification can vary between two extremes: (i) encrypting the microdata and, (ii) leaving the microdata intact. In the first extreme, the protection is perfect (*i.e.* only the owner of a secret key can see the data), however, the utility of the data is almost nonexistent because the encrypted microdata can be hardly studied or analysed. In the other extreme, the microdata are extremely useful (*i.e.* all their information remains intact), however, the privacy of the respondents is endangered. SDC methods for microdata protection aim at distorting the original data set to protect respondents from re-identification whilst maintaining, as much as possible, some of the statistical properties of the data and minimising the information loss. The goal is to find the right balance between data utility and respondents' privacy.

microdata sets are organised in records that refer to individual respondents. Each record has several attributes. Table I is an example [1] of a microdata set with 6 records and 4 attributes that contain information about the *"Social security number"* of the respondent, the *"City"* where he/she lives, the *"Job"* he/she has, and whether he/she suffers from *"AIDS"*. More generally the attributes that can appear in a microdata set $\mathbf{X}$ can be classified in three categories as follows:

1) *Identifiers*. Attributes in $\mathbf{X}$ that *unambiguously* identify the respondent. For example, passport numbers, full names, etc. In our example (*cf.* Table I) the attribute *"social security number"* is an identifier.
2) *Key attributes*. Those in $\mathbf{X}$ that, if properly combined, can be linked with external information sources to re-identify some of the respondents to whom some of the records refer. For example, address, age, gender, etc. In our example (*cf.* Table I) the attributes *"City"* and *"Job"* are key attributes.
3) *Confidential outcome attributes*. Those containing sensitive information on the respondent, namely salary, religion, political affiliation, health condition, etc. In our example (*cf.* Table I) the attribute *"AIDS"* is a confidential outcome attribute.

Due to the fact that the re-identification of respondents is

Agusti Solanas is with the UNESCO Chair in Data Privacy in the Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, Av. Paisos Catalans 26, 43007, Tarragona, Catalonia, Spain (phone: +34 977 558867; email: agusti.solanas@urv.cat).

Úrsula González-Nicolás is with the Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, Av. Paisos Catalans 26, 43007, Tarragona, Catalonia, Spain (phone: +34 977 558270; email: ursula.gonzaleznicolas@urv.cat).

Antoni Martínez-Ballesté is with the Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, Av. Paisos Catalans 26, 43007, Tarragona, Catalonia, Spain (phone: +34 977 558876; email: antoni.martinez@urv.cat).

[1] The SS numbers in this Table are invented, any resemblance to real people, living or dead is purely coincidental.

978-1-4244-8126-2/10/$26.00 ©2010 IEEE

TABLE I
TOY EXAMPLE OF A MICRODATA SET WITH 6 RECORDS.

Row	SS. Num.	City	Job	AIDS
1	123-45-6789	BigCity	Baker	YES
2	987-65-4321	Village	Plumber	NO
3	124-35-6879	Hamlet	Doctor	YES
4	133-35-2829	Village	Teacher	NO
5	424-25-6272	Village	Farmer	NO
6	924-95-3839	BigCity	Nurse	NO

to be avoided, we can assume that the identifiers in $\mathbf{X}$ are removed or encrypted before the microdata set is released. However, unlike identifiers, key attributes and confidential outcome attributes cannot be removed from $\mathbf{X}$ without significantly degrading the quality of the information, because any attribute is potentially a key attribute. Unfortunately, removing or encrypting identifiers does not provide enough protection against re-identification to respondents. To clarify this point, consider our example in Table I, imagine that we have removed the information about the social security numbers (*i.e* the identifier attribute), so an external observer only knows the "City" where a respondent lives, his/her "Job" and, whether he/she suffers from AIDS. It is quite apparent that the respondent number 3 can be easily re-identified because the "doctor" in a given small hamlet is a well-known person and, in general, there is only one doctor in a small hamlet.

From the example above, it becomes clear that privately releasing microdata is not straightforward and requires procedures far more sophisticated than simply removing identifiers. Several techniques have been proposed to cope with this problem, namely noise addition, rank swapping, microaggregation, etc. Each of these techniques has its pros and cons, however, microaggregation is the youngest, and recent studies have shown its natural capability of preserving interesting properties such as k-anonymity.

A. Contribution and Plan of the Article

In this article we focus on microaggregation and we show how to improve its results by means of mixing the simplicity of classical microaggregation with the great exploration capabilities of genetic algorithms. We propose a novel two-step partitioning method based on the Variable-MDAV algorithm that, when properly combined with genetic algorithms, leads to clear improvements with respect to classical microaggregation.

The rest of the article is organised as follows: Section II summarises some fundamental concepts of microaggregation and some previous work on the use of genetic algorithms to improve the results of classical microaggregation. Section III elaborates on some partitioning methods for microaggregation with genetic algorithms and introduces our new proposal based on the V-MDAV algorithm. In Section IV, we present the experimental results obtained over synthetic and real data sets; and we compare the results of the previous methods with the one we propose. Finally, the article concludes in Section V with some final comments.

TABLE II
MICROAGGREGATION OF THE AGE OF 9 RESPONDENTS WITH $k = 3$.

	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9
Original	91	25	90	89	48	50	30	20	52
Part/Group	1	2	1	1	3	3	2	2	3
Micro-Aggregated	90	25	90	90	50	50	25	25	50

II. BACKGROUND AND RELATED WORK

In this section we describe the microaggregation problem. Then we summarise the main classical methods proposed to microaggregate data sets, and we finally overview the proposals that use genetic algorithms to microaggregate microdata sets.

A. The Microaggregation Problem

Microaggregating a microdata set can be understood as a clustering problem with constraints on the cardinality of the clusters. Given a privacy parameter k and a data set $\mathbf{X}$ with n records, the microaggregation problem consists in obtaining a partition[2] of the data set, so that each part in the partition has at least k records and the within-part homogeneity is maximised (*i.e.* records in the same part are as similar as possible). The resulting partition is called k-partition [4] because all parts have, at least, k records. After obtaining the k-partition, a microaggregated data set is generated by replacing each original record by the average record of the part to which it belongs. Table II shows an example in which the age of a set of 9 respondents have been microaggregated with a privacy parameter $k = 3$. In the example, records are clustered according to their similarity in subsets of cardinality 3. Then the average value of each part is computed and used to replace the original values of the microdata set.

This procedure clearly guarantees k-anonymity to respondents (*i.e.* each original record is replaced by the average of the part to which it belongs. Consequently, every record in the microaggregated data set has, at least, $k - 1$ identical, indistinguishable records). Although it is apparent that the privacy of the respondents is guaranteed when the proper value of k is used, the problem of obtaining homogeneous subsets to reduce the information loss is not so easy to solve. First, we must determine how to measure the homogeneity of the subsets – the higher the homogeneity, the lower the information loss. Second, we must try to maximise it, that is, to minimise the information loss.

The most common homogeneity measure is the within-part (*a.k.a.* within-group, or within-subset) *Sum of Square Errors* (SSE) [5], [6], [7]. This is the sum of squared Euclidean distances from the average record of each part (*i.e.* the centroid) to every record in the part. For a given k-partition,

[2] A partition of $\mathbf{X}$ is a set of disjoint subsets $\{\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \ldots, \mathbf{x_p}\}$ of $\mathbf{X}$ that fulfil that $\bigcup \mathbf{x_i}, \forall i = \mathbf{X}$

the SSE is computed as shown in Equation 1:

$$\mathbf{SSE} = \sum_{i=1}^{s}\sum_{j=1}^{n_i}(x_1^{ij} - \bar{x}_1^i, x_2^{ij} - \bar{x}_2^i, \ldots, x_d^{ij} - \bar{x}_d^i) \times \begin{pmatrix} x_1^{ij} - \bar{x}_1^i \\ x_2^{ij} - \bar{x}_2^i \\ \vdots \\ x_d^{ij} - \bar{x}_d^i \end{pmatrix}, \quad (1)$$

where s is the number of parts in the k-partition, n_i is the number of records in the i-th part, x_p^{ij} is the value of the p-th attribute of the j-th record in the i-th part, $\bar{x}_p^i$ is the value of the p-th component of the centroid of the i-th part, and d is the number of attributes/components of the microdata set (*i.e.* the dimension of each record). This equation could be written more compactly by using vector notation as shown in Equation 2:

$$\mathbf{SSE} = \sum_{i=1}^{s}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T, \quad (2)$$

where s is the number of parts in the k-partition, n_i is the number of records in the i-th part, $\mathbf{x}_{ij}$ is the j-th record in the i-th part, and $\bar{\mathbf{x}}_i$ is the centroid of the i-th part.

Given a microdata set $\mathbf{X}$, the optimal k-partition is the one with minimum SSE. In [8] Hansen and Mukherjee describe a method to determine the optimal k-partition of a univariate data set in polynomial time. In contrast, finding the optimal k-partition of a multivariate data set is known to be an NP-hard problem [9]. Due to the hardness of the problem, a plethora of microaggregation heuristics have been proposed. The main aim of these heuristic methods is to obtain k-partitions with low values of the SSE, in reasonable times (*i.e.* with a feasible computational cost). Multivariate microaggregation heuristics can be classified into two categories depending on the size of the generated subsets:

- Fixed-size heuristics. All parts in the resulting k-partition have the same cardinality k. Note that if the number of records in the original data set is not a multiple of k, there will be a part with a cardinality between k and $2k-1$. See [10], [4], [11], [12], [13].
- Variable-size heuristics. All parts can have different cardinality varying between k and $2k-1$. See [4], [14], [15], [13], [16], [17].

In general, variable-size heuristics are slower than fixed-size heuristics, but they improve the results of fixed-size heuristics in terms of homogeneity. This happens because variable-size heuristics have more freedom to adapt to the distribution of the records (See Figure 1 for a graphical example of this behaviour). In the example of Figure 1 we have a continuous bivariate microdata set with 11 records. Assuming that we want to microaggregate this data set with a security parameter $k = 3$, a fixed-size heuristic would return a 3-partition such as $\{\{1,2,3\},\{4,5,6,7,8\},\{9,10,11\}\}$, which is not optimal in terms of within-groups homogeneity. An optimal 3-partition would be $\{\{7,8,9,10,11\},\{4,5,6\},\{1,2,3\}\}$. It becomes apparent that the lack of flexibility in the size of the groups splits "natural" groups. As a result, the obtained k-partition is far from optimal.

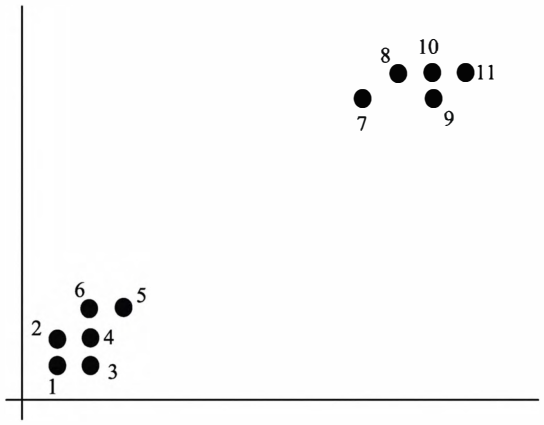


Fig. 1. Continuous, bivariate, microdata set that illustrates the lack of flexibility of fixed-size heuristics

B. Microaggregation with Genetic Algorithms

Genetic algorithms (GA) are biologically inspired optimisers based on the principles of evolution and natural selection [18]. Given a population of individuals (*i.e.* possible solutions), an environment (*i.e.* the problem space), a measure of adaptation to that environment (*i.e.* a fitness function), and a selection and recombination functions, GA look for solutions by means of mutating and recombining individuals that progressively evolve to better fit the environment.

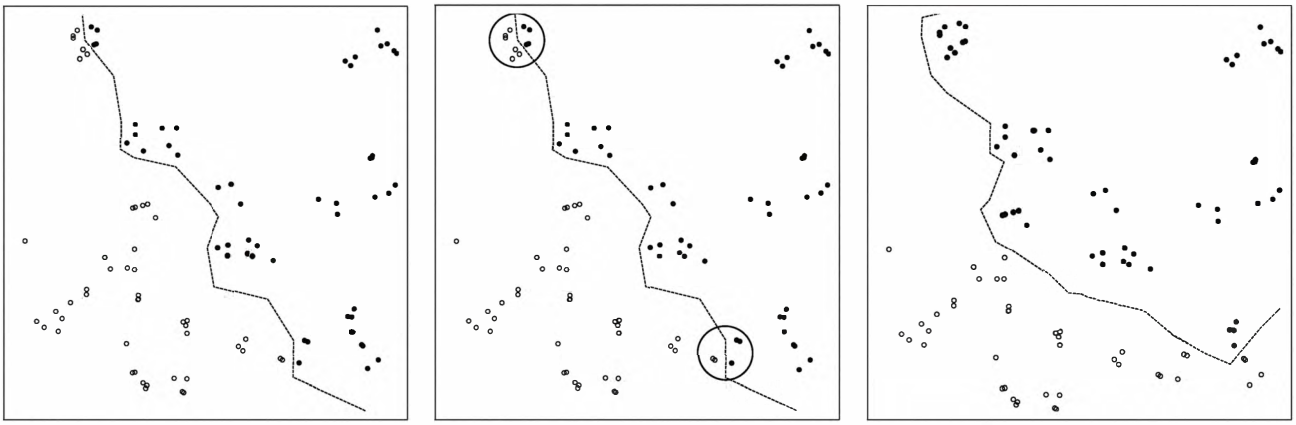
GA have been widely used in a variety of fields. In [19] Solanas et al. proposed the use of GA to microaggregate small data sets of up to a hundred records. A novel N-ary coding was proposed to cope with the multivariate nature of microaggregation and a complete set of experiments were performed to determine the best values for the main parameters of the GA, namely the population size, the crossover rate, the mutation rate, etc.) The GA described in [19] uses the *roulette wheel* selection algorithm, which is based on selecting chromosomes proportionally to their fitness, which was computed as shown in Equation 3:

$$F = \frac{1}{\mathbf{SSE} + 1}, \quad (3)$$

where the SSE is computed using Expression 1. There is a variety of selection functions, however, in this article we will mainly consider the roulette wheel algorithm and the fitness uniform selection scheme (FUSS) [20] algorithm.

Although the ideas proposed in [19] were a clear advance in the search for optimal solutions for multivariate microaggregation, they were only applicable to small data sets. In [21] Martínez-Ballesté et al. present a preliminary study on how to apply GA to medium-sized data sets. The method they proposed consists of two steps: (i) partition the original data set into rough subsets by using a fixed-size microaggregation heuristic and (ii) apply the GA on the obtained rough subsets to generate the microaggregated data set.

The main problem of [21] is that the partitioning method proposed does not consider the natural distribution of the records in the data set, and due to the very nature of fixed-



(a) V-MDAV-based 1-step partitioning (b) Split of natural subsets with the V-MDAV-based 1-step partitioning (c) V-MDAV-based 2-step partitioning

Fig. 2. Example of the different behaviour of 1-step and 2-steps partitioning of a bivariate microdata set.

sized microaggregation heuristics, naturally grouped records could be split in different subsets. In this article we propose a new partitioning method that improves the preservation of subsets containing naturally grouped records.

III. Our variable-size microaggregation-based partitioning strategy

It is not straightforward to determine how to partition an original microdata set in order to obtain parts that are suitable for a GA-based optimization. As we have explained in the previous section, in [21] Martínez-Ballesté et al. used a fixed-size microaggregation technique to partition a microdata set into smaller parts that the GA described in [19] can handle. Specifically in [21] the authors used the well-known Maximum Distance to Average Vector (MDAV) method [11][22]. In this section, we describe our approach based on a variable-size microaggregation technique, that is, Variable Maximum Distance to Average Vector (V-MDAV) [17], and we elaborate on the different ways of applying our solution (*i.e.* one-step partitioning, and two-step partitioning).

A. One-step Variable-size Partitioning

We know from [19], that a properly tuned GA can deal with up to 50 records with remarkable results. Thus, similarly to [21], our approach consists in generating subsets of cardinality smaller than 50. To do so, we propose the utilisation of the V-MDAV method. In general, V-MDAV improves the results of MDAV by means of adapting to the natural distribution of the records in a microdata set. The V-MDAV works as follows:

1) Compute the distance matrix between all records.
2) Compute the centroid c of the data set.
3) Find the most distant unassigned record r from c. Build a subset around this record, formed by r and its $k-1$ closest records.
4) Extend the subset.
5) Continue to Step 3 until the number of remaining records is smaller than $2k$.
6) If the number of remaining records is smaller than k, assign each remaining record to its closest subset. Otherwise, form the last subset with the remaining records.

The key point of V-MDAV, which makes it better and more powerful than MDAV, is the subset extension step (*i.e.* step 4 of the algorithm). If we remove this step, V-MDAV would behave almost the same as MDAV. Notwithstanding, the extension step allows V-MDAV to adapt to the natural distribution of the records. After generating a subset of k records, the extension step finds possible candidate records, which could join the subset and, if any of these candidate records are closer to the subset than to other unassigned records, they are added to the subset. The extension step works as follows:

1) Find the closest unassigned record u to the lately generated subset.
2) Let d_{in} be the distance from u to the closest record e_i in the lately generated subset.
3) Let u_{out} be the unassigned record closest to u. Let d_{out} be the distance from u_{out} to u.
4) If $d_{in} < \gamma \cdot d_{out}$ and the number of records in the subset is smaller than $2k-1$, add record u to the subset. Return to step 1.
5) Otherwise finish the extension of the subset.

The parameter γ that multiplies d_{out} in step 4 is a *gain factor* that must be tuned depending on the distribution of the records in the microdata set. Determining the best value of γ is not straightforward and is out of the scope of this paper. However, more details can be found in [17].

With the aim to illustrate the behaviour of V-MDAV as a partitioning method, we have generated a small bivariate data set consisting of one hundred records and we have partitioned it using V-MDAV with $k = 40$ (*i.e.* two subsets are to be obtained). Figure 2(a) shows the data set. Figure 2(b) shows the partition obtained by V-MDAV. It can be observed that some natural subsets are split and their records are assigned to different parts (*e.g.* see the circles in Figure 2(b)).

Due to the fact that natural subsets are split in different parts, the GA that we want to apply to each part cannot fully optimise the microdata set in terms of SSE. The breakage of natural subsets is due to the cardinality constraints of microaggregation methods (*i.e.* the generated parts must have at least k records and less than $2k$ records). To overcome this limitation, a two-step partitioning strategy is proposed in the next section.

B. Two-step Variable-size Partitioning

Instead of partitioning a microdata set in rough parts in a single step (*cf.* previous section), the two-step partitioning proceeds as follows:

1) First step
   - V-MDAV is used to partition the original microdata set $\mathbf{X}$ using a small value of k, usually 3, 4 or 5. Let us call this value k_1.
   - A new data set $\bar{\mathbf{X}}$ is generated and each part of $\mathbf{X}$ is represented by its centroid. Note that, assuming that the original microdata set has n records, the data set $\bar{\mathbf{X}}$ generated at this point has n/k_1 records.
2) Second step
   - V-MDAV is used to partition the data set $\bar{\mathbf{X}}$ generated in the previous step. This time a greater value of k is used, usually 10. Let us call this value k_2.
   - Once the partition of $\bar{\mathbf{X}}$ is finished, each record in $\bar{\mathbf{X}}$ is replaced by the original records that it represents, thus obtaining a partition of the original microdata set with a privacy parameter $k = k_1 \cdot k_2$. Note that due to very nature of V-MDAV, the cardinality of the parts will vary between k and $2k - 1$.

By following this procedure, natural subsets of cardinality k_1 are not split. Consequently, the GA that we will apply over each part will have more chances to optimise the microaggregation in terms of SSE. Figure 2(c) illustrates the behaviour of the two-step V-MDAV partition that we propose. It can be observed that subsets that were split with the one-step method are preserved now. By construction it can be assured that no natural subsets of cardinality smaller than k_1 are going to be split. This property, however, cannot be guaranteed when the one-step method is used. Note that although it might seem that a natural subset is broken in Figure 2(c), it is not so because each subgroup has cardinality k_1.

C. Genetic Algorithm Description

Regardless of the partitioning method used (*i.e.* one-step V-MDAV partitioning or two-setp V-MDAV partitioning), we use a GA to individually microaggregate each subset. Note that the microaggregation of all subsets leads to the microaggregation of the complete microdata set. In this paper and, according to the results obtained in [19], we tune our GA with the following parameters: (i) Mutation rate: 0.1, (ii) Crossover rate: 0.4, (iii) Population size: 100 chromosomes, and (iv) Number of iterations: 5,000.

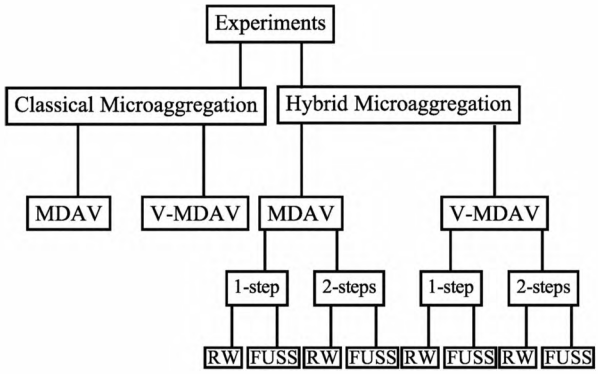


Fig. 3. Graphical representation of the different microaggregation techniques that we have used in our experiments.

In [19] the roulette wheel selection algorithm was used. However, we will also consider the fitness uniform selection scheme (FUSS) [20].

IV. Experimental Results

In this section we describe the tests that we have carried out to assess the usefulness of our proposal. We have used three real microdata sets, which have been widely used as reference microdata sets during the CASC project [23] and in [4], [24], [12], [15]. The microdata sets are the following:

- "Census" is a microdata set that contains 1,080 records with 13 numerical attributes.
- "EIA" contains 4,092 records with 12 numerical attributes.
- "Tarragona" is a microdata set containing 834 records with 13 numerical attributes.

In addition we have generated some synthetic data sets to better illustrate the behaviour of our proposal depending on the distribution of the records in the data set:

- Clustered1000x2: This is a microdata set with 1000 records and two attributes. The records have been generated so as to naturally cluster in subsets of 3, 4 and 5 records.
- Clustered1000x3: Like the Clustered1000x2 but with 3 attributes.
- Scattered1000x2: This is a microdata set with 1000 records randomly distributed. By construction, these records are not naturally clustered in subsets.
- Scattered1000x3: Like the Scatteres1000x2 but with 3 attributes.

We have microaggregated each microdata set by using the different methods explained in this article, namely classical MDAV, classical V-MDAV, Hybrid MDAV and Hybrid V-MDAV. For the hybrid methods, we have considered the partition in one step and in two steps. Moreover, we have also distinguished the use of the roulette wheel (RW) and the fitness uniform selection scheme (FUSS) selection algorithms. A graphical scheme with all the experiments that we have carried out is shown in Figure 3. We have microaggregated all data sets to guarantee 3-anonymity, that is, the results

TABLE III
SUMMARY OF THE EXPERIMENTAL RESULTS ($L = \frac{SSE}{SST}$). THE BEST RESULT FOR EACH DATA SET IS HIGHLIGHTED AND COLOURED.

Data set	Classical methods		Hybrid methods							
	MDAV	V-MDAV	MDAV 1 step		V-MDAV 1 step		MDAV 2 steps		V-MDAV 2 steps	
			Rulette	FUSS	Rulette	FUSS	Rulette	FUSS	Rulette	FUSS
Census	0.056619	0.056619	0.05986	0.059196	0.060044	0.05927	0.054655	0.054232	0.054321	**0.053717**
EIA	0.011017	0.010975	0.013155	0.012936	0.013257	0.012969	0.010055	**0.009542**	0.009916	0.009834
Tarragona	0.169507	0.158477	0.156458	0.157482	0.156276	0.157445	0.159503	0.159509	0.151785	**0.151701**
Clustered 1000x2	0.001978	0.000374	0.000723	0.000571	0.000213	0.000107	0.001004	0.000963	0.00014	**0.000088**
Clustered 1000x3	0.007195	0.000495	0.003294	0.003088	0.000525	0.000364	0.004471	0.004415	0.000434	**0.00024**
Scattered 1000x2	0.002234	0.002141	0.002548	0.002399	0.00228	0.002124	0.002107	0.002034	0.001976	**0.001868**
Scattered 1000x3	0.013171	0.013082	0.015113	0.014019	0.014946	0.014092	0.012923	0.012306	0.013487	**0.012127**

of all the studied methods are microaggregated data sets where each record has, at least, 2 other indistinguishable records. To do so, we have fixed the privacy parameter $k = 3$. After microaggregating the data sets, we have analysed the information loss introduced by each method. The information loss have been measured by using the following expression:

$$\mathbf{L} = \frac{\mathbf{SSE}}{\mathbf{SST}} \ , \tag{4}$$

where the SSE is computed using Expression 1 and the total sum of squares (SST) is computed as follows:

$$\mathbf{SST} = \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \ , \tag{5}$$

where n is the number of records in the data set and $\bar{\mathbf{x}}$ is the centroid of the data set. By means of dividing the SSE by the SST we obtain a normalised measure ranged between 0 and 1 that can be easily compared regardless of the used data set. For each method 10 executions have been run and the results given in Table III are the average of all these executions.

A. MDAV vs. V-MDAV

The MDAV microaggregation heuristic is a fixed-size heuristic (*i.e.* all the subsets that it generates have the same size). On the contrary, V-MDAV is able to generate subsets of variable size, thus increasing the intra-subset homogeneity.

Our intuition that V-MDAV would improve the results of MDAV as a partitioning method has been confirmed with the obtained results shown in Table III. It is apparent that V-MDAV clearly outperforms MDAV when the records in the microdata set are naturally clustered (*e.g.* see the results of Table III for the clustered data sets). The improvements achieved by V-MDAV are not so extraordinary when the records are not naturally clustered – this is the case of the real microdata sets "Census", "EIA", "Tarragona", and the scattered data sets.

Although the main improvements take place on clustered data sets, using V-MDAV on scattered data sets is also beneficial. As shown in Table III, V-MDAV (along with the GA) almost always obtains the best results.

B. One-step vs. Two-steps Partitioning

The correct partitioning of the micro data set is essential to obtain a good result (*i.e.* a low information loss). From the results shown in Table III, it becomes apparent that the 2-step partitioning generally outperforms the 1-step partitioning. This behaviour is specially remarkable when the variable-size microaggregation heuristic is used.

These results confirm that the one-step partitioning splits more natural subsets than the two-step partitioning. By construction, the two-steps partitioning guarantees that small subsets of cardinality k_1 (*cf.* Section III) will not be split into different parts during the partitioning process. Consequently, the GA can easily optimise the result.

It can be concluded that the two-step partitioning is clearly superior to the one-step partitioning.

C. Roulette Wheel vs. FUSS

The role that the GA plays in our hybrid proposal is very important. After partitioning the original microdata set in disjoint subsets, the GA is responsible for the proper optimisation of the microaggregation. In general, we have used the values recommended in previous proposals (*cf.* Section III), however, in this paper we have analysed the importance of the selection function.

It can be observed in Table III that the fitness uniform selection scheme (FUSS) outperforms the roulette wheel selection algorithm. This result is due to the superior ability of FUSS to maintain the diversity of the population when there exist many local optima, in which a classical roulette wheel selection algorithm would become stuck. As a result of the diversity provided by FUSS, the solutions to our problem are found faster and they are better.

It can be concluded that, in the context of the microaggregation problem, using the fitness uniform selection scheme is better than using the roulette wheel algorithm.

V. CONCLUSIONS

Multivariate microaggregation is an NP-hard problem that is tackled by means of heuristics. Off-the-shelf microaggregation methods protect the privacy of the respondents

by distorting the records in the original data sets. At the same time, they try to reduce the information loss whilst the privacy of the respondents is maintained. Using GA to microaggregate data sets seems to be a promising idea, however, current GA-based methods can hardly deal with large data sets. Consequently, partitioning a large data set into smaller subsets that can be properly handled by a GA is a possible solution.

In this paper we have proposed a hybrid technique that mixes the benefits of variable-size microaggregation heuristics like V-MDAV (*i.e.* they can deal with large data sets whilst maintaining the natural distribution of the subsets) with the benefits of the GA (*i.e.* they find very good solutions with small data sets). We have shown that our proposal outperforms the previous ones in terms of information loss whilst maintaining the same degree of privacy.

Although this proposal clearly improves the results of the previous ones, there are still several points to be studied and improved in the future:

- Propose methods to optimally determine the value of γ: This is a tough task. We believe that a value of γ that smoothly adapts to the data could be the best approach.
- Propose methods to find the best values for k_1 and k_2: Similarly to γ, the best values of k_1 and k_2 should vary according to the microdata set. Thus, data-oriented methods seem promissing.
- Study and improve the coding of the GA so as to allow it to cope with larger data sets.

ACKNOWLEDGEMENTS

This work was partly supported by the Spanish Government through projects TSI2007-65406-C03-01 "E-AEGIS" and CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", and by the Government of Catalonia under grant 2009 SGR 1135. The views of the authors with the UNESCO Chair in Data Privacy do not necessarily reflect the position of UNESCO nor commit that organisation.

REFERENCES

[1] European Parliament, "DIRECTIVE 2002/58/EC of the European Parliament and Council of 12 july 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications)," 2002, http://europa.eu.int/eur-lex/pri/en/oj/dat/2002/l_201/l_20120020731en00370047.pdf.

[2] Canadian Privacy, "Canadian privacy regulations," 2005, http://www.media-awareness.ca/english/issues/privacy/canadian_legislation_privacy.cfm.

[3] US Privacy, "U.S. privacy regulations," 2005, http://www.media-awareness.ca/english/issues/privacy/us_legislation_privacy.cfm.

[4] 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, pp. 189–201, 2002.

[5] A. W. F. Edwards and L. L. Cavalli-Sforza, "A method for cluster analysis," *Biometrics*, vol. 21, pp. 362–375, 1965.

[6] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, pp. 236–244, 1963.

[7] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1967, pp. 281–297.

[8] S. L. Hansen and S. Mukherjee, "A polynomial algorithm for optimal univariate microaggregation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 1043–1044, July-August 2003.

[9] 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, pp. 345–354, 2001.

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

[11] A. Hundepool, A. V. de Wetering, 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.

[12] 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.

[13] E. Fayyoumi and B. J. Oommen, "A fixed structure learning automaton micro-aggregation technique," in *Privacy in Statistical Databases-PSD 2006*, ser. Lecture Notes in Computer Science, J. Domingo-Ferrer and L. Franconi, Eds., vol. 4302, Berlin, 2006, pp. 114–128.

[14] G. Sande, "Exact and approximate methods for data directed microaggregation in one or more dimensions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 459–476, 2002.

[15] M. Laszlo and S. Mukherjee, "Minimum spanning tree partitioning algorithm for microaggregation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 7, pp. 902–911, 2005.

[16] J. Domingo-Ferrer, A. Martínez-Ballesté, J. M. Mateo-Sanz, and F. Sebé, "Efficient multivariate data-oriented microaggregation," *The VLDB Journal*, vol. 15, no. 4, pp. 355–369, 2006.

[17] A. Solanas and A. Martínez-Ballesté, "V-MDAV: Variable group size multivariate microaggregation," in *COMPSTAT'2006*, Rome, 2006, pp. 917–925.

[18] J. Holland, *Adaptation in Natural and Artificial Systems*. University of Michigan Press, 1975.

[19] A. Solanas, A. Martínez-Ballesté, J. M. Mateo-Sanz, and J. Domingo-Ferrer, "Multivariate microaggregation based on genetic algorithms," in *3rd IEEE Conference On Intelligent Systems*. Westminster: IEEE Computer Society Press, 2006, pp. 65–70.

[20] M. Hutter, "Fitness uniform selection to preserve genetic diversity," IDSIA, Manno-Lugano, Switzerland, Tech. Rep. IDSIA-01-01, 2001.

[21] A. Martínez-Ballesté, A. Solanas, J. Domingo-Ferrer, and J. M. Mateo-Sanz, "A genetic approach to multivariate microaggregation for database privacy," in *ICDE Workshops*. IEEE Computer Society, 2007, pp. 180–185. [Online]. Available: http://dx.doi.org/10.1109/ICDEW.2007.4400989

[22] A. Solanas, *Studies in Computational Intelligence: Success in Evolutionary Computation*. Springer, 2008, ch. Privacy Protection with Genetic Algorithms, pp. 215–237.

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

[24] J. Domingo-Ferrer, F. Sebé, and A. Solanas, "A polynomial-time approximation to optimal multivariate microaggregation," *Computers & Mathematics with Applications*, vol. 55, no. 4, pp. 714–732, 2008.