Privacy in Statistical Databases: k-Anonymity Through Microaggregation

Josep Domingo-Ferrer, *Member, IEEE,* Agusti Solanas, *Associate, IEEE,* and Antoni Martínez-Ballesté

***Abstract*— The amount of computer-stored information is growing faster with each passing day. This growth and the way in which the stored data are accessed through a variety of channels have raised the alarm about the protection of the individual privacy of the respondents whose data are being collected and stored. On the one hand, data should be available to researchers and statistical agencies so that the necessary research and planning activities can be conducted. However, on the other hand, the right of respondents to privacy must be protected.**

Statistical disclosure control (SDC) is the discipline which cares about keeping a balance between data access and privacy protection. k-Anonymity is one particular approach to SDC for individual data (microdata): the record corresponding to a specific respondent is k-anonymous if an intruder can at best link the record to a group of k respondents containing the correct one. This paper surveys the use of a special clustering technique called microaggregation to provide k-anonymity.

***Index Terms*— K-Anonymity, Privacy, Microaggregation.**

I. Introduction

Due to the new information technologies and the massive use of computers for organizing and distributing electronic data, the amount of stored information has had a spectacular increase in a very diverse set of fields, namely finance, health care and engineering. This enormous increase of stored data in sensitive areas such as health care has led to a new set of needs related to information management and privacy protection. Thus, the way to guarantee the rights of the respondents whose information is being stored has undergone a substantial change. It has become necessary to develop new techniques able to keep the balance between sharing information and protecting individual privacy. The problem is "the inevitable conflict between the individual's right to privacy and the society's need to know and process information" [1]. To that end, an array of techniques for privacy in statistical databases collectively known as Statistical Disclosure Control (SDC) have been developed. In the next sections we elaborate on SDC by paying special attention to two concepts about privacy in statistical databases: k-anonymity and microaggregation. Protecting the individual respondent information contained in statistical databases is not straightforward because data must be published in a way which cannot be linked to any specific respondent. However, this is a thorny issue: for example, a malicious user can use a combination of statistical aggregates to derive information about a single individual (tracker attacks,[2]). SDC attempts to cope with these difficulties. A possible classification of SDC techniques is as follows [3]:

- Conceptual methods: The conceptual model provides a framework for investigating the security problem at a conceptual-data-model level [4] [5].
- Query restriction: Protection is provided through the following measures: [6][7] Query set size restriction; Overlap control between successive queries; Making some data unavailable; Database partitioning; Cell suppression.
- Data perturbation: Noise is added to the data in order to modify them. [8], [9], [10]
- Output perturbation: The answer to the user queries is perturbed while leaving the underlying data unchanged.

In this paper we present a survey of methods which fall into the "data perturbation" category, with a special focus on k-anonymity and microaggregation. Three microaggregation methods are reviewed and analyzed from the computational point of view. Then, they are compared in terms of the information loss (*e.g.* perturbation) they cause to data. Section II defines k-anonymity and relates it to microaggregation. In Section III we review an early data-oriented microaggregation method, known as Maximum Distance method. In Section IV the Maximum Distance to Average Vector (MDAV) method is described. In Section V we analyze a very recent method, Variable-Size Maximum Distance to Average Vector (V-MDAV). Section VI provides a comparison of the previously analyzed methods. Finally, Section VII contains a conclusion.

II. k-Anonymity and microaggregation

k-Anonymity is a property related to privacy protection against statistical disclosure. Let $\mathbf{X}$ be a protected data set where identifiers have been removed. Given a subset of attributes —known as quasi-identifiers— in $\mathbf{X}$ known by an intruder from external identified sources $\mathbf{X}'$ (*e.g.* phone books, electoral rolls, etc.), the protected data set $\mathbf{X}$ is said to be k-anonymous if at least k records exist in $\mathbf{X}$ sharing any combination of values of quasi-identifiers. In this way, the best the intruder can achieve is to link an identified record in $\mathbf{X}'$ with a set of k records in $\mathbf{X}$. The computational procedure initially proposed to implement k-anonymity methods was based on suppressing/generalizing values of quasi-identifier attributes. Recently, it has been pointed out that

The authors are partly supported by the Spanish Ministry of Science and Education through project SEG2004-04352-C04-01 "PROPRIETAS" and by the Catalan government under grant 2005 SGR 00446.

Josep Domingo-Ferrer is with the Dept. of Computer Engineering and Maths of the Rovira i Virgili University in E-43007 Tarragona, Catalonia. (phone: +34977558270, fax: +34977559710, e-mail:josep.domingo@urv.net)

Agusti Solanas is with the Dept. of Computer Engineering and Maths of the Rovira i Virgili University in E-43007 Tarragona, Catalonia. (e-mail:agusti.solanas@urv.net)

Antoni Martínez Ballesté is with the Dept. of Computer Engineering and Maths of the Rovira i Virgili University in E-43007 Tarragona, Catalonia. (e-mail:antoni.martinez@urv.net)

microaggregating quasi-identifiers can achieve the same goal without resulting in partially suppressed or coarser data [11]. Microaggregation has been used for several years in different countries: it started at Eurostat [12] in the early nineties, and has since then been used in Germany [13] and several other countries [14]. Microaggregation is relevant not only for SDC, but also in artificial intelligence [15]. In the latter field, the application is to increase the *knowledge* of a system for decision making and domain representation. Microaggregation techniques may also be used in data mining in order to scale down or even compress the data set while minimizing the information loss. Microaggregation satisfies the k-anonymity condition by clustering records in a data set into groups of at least k records. Such a grouping is called a k-partition. Then, each record is replaced by the centroid of its group. In order to minimize the information loss caused by microaggregation, groups should be formed so that the within-group homogeneity is maximum. There are a lot of group homogeneity measures in the literature based on different distance definitions (*e.g.* Euclidean distance, Minkowski distance, Chebyshev distance, etc.). The most common homogeneity measure for clustering is within-group sum of squares SSE [16] [17][18]. The SSE is the sum of squared distances from the centroid of each group to every record in the group. For a k-partition, SSE can be computed as:

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

where s is the number of groups in the k-partition and n_i is the number of records in the i-th group. Expression (1) is computed on the data after standardizing them, this is, after subtracting from the values of each attribute the attribute mean and dividing the result by the attribute standard deviation [1]. Optimal multivariate microaggregation, that is, with minimum SSE, was shown to be NP-hard in [19]. Thus, the only practical microaggregation methods are heuristic. In [18] it was shown that groups in an optimal k-partition contain between k and $2k-1$ records.

III. Maximum Distance

The Maximum Distance method (MD) was proposed in [18] as a multivariate microaggregation method. Next we describe the algorithm and assess its computational complexity.

A. The algorithm

The MD algorithm builds a k-partition as follows:

1) Let r and b be the two most distant records in the data set, using the Euclidean Distance; form a group with r and the $k-1$ records closest to r; form a group with s and the $k-1$ records closest to s.
2) If there are at least $2k$ records which do not belong to any of the groups formed in Step 1, go to Step 1 taking as the new data set the previous data set minus the groups formed in the previous instance of Step 1.
3) If there are between k and $2k-1$ records which do not belong to any of the groups formed in Step 1, form a new group with those points and exit the algorithm.
4) If there are less than k records which do not belong to any of the groups formed in Step 1, add them to the closest group formed in Step 1.

At this moment, we have a k-partition of the data set into a number of groups. As explained above, given a k-partition, the microaggregated data are computed by replacing each record by the centroid of the group it belongs to.

[1]Standardization does not affect attribute correlations.

B. Computational cost

The dimensionality (number of attributes) of the data set will not be considered in the analysis because it is usually much smaller than n and it only has a slight computational impact on distance calculations. For a data set with n points, the complexity of computing a k-partition using MD is the cost of forming $\lfloor \frac{n}{k} \rfloor$ groups of k points, two per iteration. The complexity of an iteration can be split as:

- Finding the two most distant points among the remaining ungrouped points in the data set. If we have an average of $\frac{n}{2}$ remaining ungrouped points, this requires computing an upper-triangular distance matrix with an average $\frac{n}{2}$ rows, that is, $\frac{n}{2}\frac{\frac{n}{2}-1}{2} = \frac{n(n-2)}{8}$ distance computations.
- Two groups with the $k-1$ closest points to each of the two points above are formed. To form one group, we must compute one row of the distance matrix with an average of $n/2$ columns and identify columns with shortest distances, which takes $\frac{(k-1)n}{2}$ computations. Thus, forming two groups takes $(k-1)n$ operations.

Therefore, since there are $\lfloor \frac{n}{2k} \rfloor$ iterations, the computation of the k-partition takes $O(n^3/k)$ operations.

IV. Maximum Distance to Average Vector

The main drawback of MD is its computational complexity because of the cost of finding the most distant records at each iteration. The Maximum Distance to Average Vector method (MDAV) improves on MD in terms of computational complexity while maintaining the performance in terms of resulting SSE. MDAV was proposed in [20], [11] as part of a multivariate microaggregation method implemented in the μ-Argus package for statistical disclosure control.

A. The algorithm

The MDAV algorithm is as follows:

1) Compute the centroid (average record) $\bar{x}$ of records in the data set. Find the most distant record r from the centroid. Also find the most distant record s from r.
2) Form two groups around r and s: the first group contains r and the $k-1$ records closest to r; The other group contains s and the $k-1$ records closest to s.
3) If there are at least $2k$ records which do not belong to any of the groups formed in Step 2, go to Step 1 taking as new set of points the previous set of records minus the groups formed in the latest instance of Step 2.

4) If there are between k and $2k-1$ records which do not belong to any of the groups formed in Step 2, form a new group with those records and exit the algorithm.
5) If there are less than k remaining records which do not belong to any of the groups formed in Step 2, add them to the group formed in Step 2 whose centroid is closest to the centroid of the remaining points.

At this moment, we have a k-partition of the data set into a number of groups. We proceed in the same way as MD so as to obtain the microaggregated data set.

B. Computational cost

The computational complexity of MDAV is much lower than the cost of running MD over a data set with n records, more precisely $O(n^2)$. Hence, MDAV can microaggregate data sets with, say, 10000 records in a few seconds. MDAV consists of forming $\lfloor \frac{n}{k} \rfloor$ groups of k points, two at each iteration. Assuming that $\frac{n}{2}$ ungrouped records remain on average, the complexity of forming one group can be split as:

- Computing the centroid of the remaining ungrouped records.
- Finding the most distant record r from the centroid, which requires $\frac{n}{2}$ distance computations.
- Finding the most distant record s from r, which requires $n/2-1$ distance computations.
- Two groups, one around r and one around s are formed, which, like for MD, takes $O((k-1)n)$ operations.

Therefore, since there are $\lfloor \frac{n}{2k} \rfloor$ iterations, the computation of the k-partition takes $O(n^2)$ operations.

V. 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 the data set, which may result in poor within-group homogeneity. Variable-size MDAV (V-MDAV) is a new algorithm that intends to overcome this limitation by computing a variable-size k-partition with a computational cost similar to the MDAV cost. A computational analysis of V-MDAV and an exhaustive set of tests can be found in [21].

A. The algorithm

The algorithm for building a k-partition using V-MDAV is as follows:

1) Compute the distances between the records and store them in a distance matrix.
2) Compute the centroid c of the data set.
3) While there are more than $k-1$ records not yet assigned to a group do:
   a) Let e be the most distant record to c.
   b) Form a group around e which contains the $k-1$ records closest to e.
   c) Extend the group, which consists of adding to the current group up to $k-1$ records using these steps: i) Find the unassigned record e_{min} which is closest to any of the records of the current group and let d_{in} be the distance between e_{min} and the group; ii) Let d_{out} be the shortest distance from e_{min} to the other unassigned records; iii) If $d_{in} < \gamma d_{out}$ then assign e_{min} to the current group.
4) If less than k records remain unassigned, no new group can be formed. In that case, assign the remaining records to their closest group.

B. Computational cost

In a similar way to MDAV, the computational cost is $O(n^2)$. In fact, the computational differences between MDAV and V-MDAV are limited to:

- V-MDAV computes the centroid of the data set only once. On the contrary, MDAV computes the centroid of the unassigned records at each iteration.
- However, V-MDAV has an additional step for extending a group, with computational cost $O(n)$.

VI. Comparison

In this section we compare the heuristics presented in this paper. We can summarize the previous discussion of the methods as follows:

- MD and MDAV generate a k-partition in which all the groups, except perhaps the last one, are of size k. V-MDAV generates variable-size k-partitions.
- The computational cost of MD is $O(n^3/k)$, whereas MDAV and V-MDAV have a cost $O(n^2)$. Therefore, MD is too costly even for a moderate number of records (*e.g.* 10000 records); using MD for large or moderate data sets needs blocking attributes to split the data set into several manageable blocks.

We present in Table I some experimental results regarding the information loss caused by the above three microaggregation methods. We have run the tests using different values of k and several data sets:

- "Scattered" is a simulated data set with $n = 1000$ records and $d = 2$ attributes each. This data set is *scattered* in the sense that no natural clusters are apparent. Attribute values were drawn from the $[-10000, 10000]$ range by simple random sampling.
- "Clustered" is a simulated data set with $n = 1000$ records with $d = 2$ attributes per record. This data set is *clustered* since its records form natural clusters. Attribute values were drawn as in the previous case, but forming clusters with between 3 and 5 records each.
- "Census" is a real data set that contains 1080 records with 13 numerical attributes.
- "EIA" is also a real data set that contains 4092 records with 11 numerical attributes.

The last two data sets were proposed as reference microdata data sets during the "CASC" project [22] and have been used in papers [18], [23], [11]. Regarding the results for the "Scattered" data set, it can be observed that the behavior of the heuristics is quite similar. The SSE value increases for higher values of k. In the "Clustered" data set, there is a substantial improvement when using V-MDAV with $k = 3$ and $k = 4$. Note that V-MDAV, with its increased flexibility,

TABLE I

SSE OF THE MICROAGGREGATED DATA SETS OUTPUT BY THE HEURISTICS DESCRIBED IN THIS PAPER

Data set	Method	$k=3$	$k=4$	$k=5$	$k=10$
Scattered	MD	4.71	7.21	9.67	22.38
	MDAV	4.72	7.37	9.95	21.74
	V-MDAV	4.57	6.98	9.82	22.56
Clustered	MD	3.53	5.03	7.03	18.56
	MDAV	3.57	4.79	6.86	18.73
	V-MDAV	1.52	3.85	7.51	19.99
Census	MD	803.09	1072.70	1264.51	2021.27
	MDAV	799.18	1053.78	1276.02	1997.03
	V-MDAV	798.49	1055.51	1260.56	1974.75
EIA	MD	212.60	347.45	751.44	1671.78
	MDAV	217.38	302.18	750.20	1728.31
	V-MDAV	240.70	337.87	511.20	1270.90

outperforms MDAV and MD in clustered data sets. However, since the "Clustered" data set has simulated clusters with 3, 4 or 5 records each, V-MDAV behaves no better than MD or MDAV for $k = 5$ and $k = 10$. Concerning the real data sets "Census" and "EIA", the information loss caused by V-MDAV is, in general, similar to the one caused by MD and MDAV. However, with the "EIA" data set, it can be noticed that V-MDAV yields k-partitions with lower SSE for $k = 5$ and $k = 10$. This is because "EIA" presents several natural clusters of a size ranging from 3 to 5 which are taken into account when building a k-partition with V-MDAV for the aforementioned values of k.

VII. Conclusion

In this paper, we presented three methods for k-anonymity in statistical databases using microaggregation. We analyzed these methods and reported their main advantages and shortcomings. From the comparisons presented in Section VI, we can conclude that V-MDAV overcomes the fixed group size constraint of previous heuristics with a similar computational cost. This makes the increased flexibility offered by V-MDAV really attractive. Experimental results show substantial performance improvement by V-MDAV when working on clustered data sets. A number of research issues remain open and will be addressed in future work: 1) Devise improved heuristics to outperform V-MDAV by using tree-based partitions and 2) Study the dichotomy between clustered and scattered data sets.

References

[1] M. A. Palley, "Security of statistical databases - compromise through attribute correlational modeling," in *Proceedings of the Second International Conference on Data Engineering*. Washington, DC, USA: IEEE Computer Society, 1986, pp. 67–74.

[2] D. E. Denning, P. J. Denning, and M. D. Schwartz, "The tracker: a threat to statistical database security," *ACM Transactions on Database Systems*, vol. 4, no. 1, pp. 76–96, 1979.

[3] N. R. Adam and J. C. Wortmann, "Security-control methods for statistical databases: a comparative study," *ACM Comput. Surv.*, vol. 21, no. 4, pp. 515–556, 1989.

[4] J. Pokorny, "Conceptual modeling of statistical data," in *Proceedings., Seventh International Workshop on Database and Expert Systems Application*. IEEE Computer Society, 1996, pp. 377–382.

[5] G. DeGiacomo and P. Naggar, "Conceptual data model with structured objects for statistical databases," in *Proceedings., Eighth International Conference on Scientific and Statistical Database Systems*. IEEE Computer Society, 1996, pp. 168–175.

[6] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "Hippocratic databases." in *VLDB*, 2002, pp. 143–154.

[7] R. Agrawal, R. J. B. Jr., C. Faloutsos, J. Kiernan, R. Rantzau, and R. Srikant, "Auditing compliance with a hippocratic database." in *VLDB*, 2004, pp. 516–527.

[8] J. Domingo-Ferrer, A. Martínez-Ballesté, and J. M. Mateo-Sanz, "Efficient multivariate data-oriented microaggregation," *Manuscript*, 2005.

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

[10] A. Solanas, A. Martínez-Ballesté, J. M. Mateo-Sanz, and J. Domingo-Ferrer, "Multivariate microaggregation based on a genetic algorithm," *Manuscript*, 2006.

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

[13] M. Rosemann, "Erste Ergebnisse von vergleichenden Untersuchungen mit anonymisierten und nicht anonymisierten Einzeldaten am Beispiel der Kostenstrukturerhebung und der Umsatzsteuerstatistik," in G. Ronning and R. Gnoss (editors), *Anonymisierung wirtschaftsstatistischer Einzeldaten*, Wiesbaden: Statistisches Bundesamt, 2003, pp. 154–183.

[14] E. C. for Europe, "Statistical data confidentiality in the transition countries: 2000/2001 winter survey," in *Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, 2001, invited paper n.43.

[15] J. Domingo-Ferrer and V. Torra, "On the connections between statistical disclosure control for microdata and some artificial intelligence tools," *Information Sciences*, vol. 151, pp. 153–170, May 2003.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] A. Solanas and A. Martínez-Ballesté, "V-MDAV: Variable group size multivariate microaggregation," 2006, manuscript.

[22] 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.

[23] J. Domingo-Ferrer, F. Sebé, and A. Solanas, "A polynomial-time approximation to optimal multivariate microaggregation," *Manuscript*, 2005.