ON THE COMPLEXITY OF OPTIMAL MICROAGGREGATION FOR STATISTICAL DISCLOSURE CONTROL

Anna Oganian and Josep Domingo-Ferrer

Universitat Rovira i Virgili, Dept. of Computer Engineering and Mathematics,
Av. Països Catalans 26, E-43007 Tarragona, Catalonia, Spain
E-mail {aoganian,jdomingo}@etse.urv.es

Abstract: Statistical disclosure control (SDC), also termed inference control two decades ago, is an integral part of data security dealing with the protection of statistical databases. The basic problem in SDC is to release data in a way that does not lead to disclosure of individual information (high security) but preserves the informational content as much as possible (low information loss). SDC is dual with data mining in that progress of data mining techniques forces official statistics to a continual improvement of SDC techniques: the more powerful the inferences that can be made on a released data set, the more protection is needed so that no inference jeopardizes the privacy of individual respondents' numerical data. This paper deals with the computational complexity of optimal microaggregation, where optimal means yielding minimal information loss for a fixed security level. More specifically, we show that the problem of optimal microaggregation cannot be exactly solved in polynomial time. This result is relevant because it provides theoretical justification for the lack of exact optimal algorithms and for the current use of heuristic approaches.

Keywords: Statistical database protection, Microdata protection, Microaggregation, Computational complexity.

1 Introduction

Statistical disclosure control (SDC), also termed inference control two decades ago [7, 8], is an integral part of data security dealing with the protection of statistical data bases. Statistical offices release two kinds of data through their databases: tabular data and microdata sets individual respondent records). In both cases, data dissemination should be performed in a way that does not lead to disclosure of individual information (*i.e.* it should not be possible to identify the individual respondent behind a published record) but preserves the informational content as much as possible (disclosure control problem). SDC is dual with data mining in that progress of data mining techniques forces official statistics to a continual improvement of SDC techniques: the more powerful the inferences that can be made on a released data set, the more protection is needed so that no inference jeopardizes the privacy of individual respondents.

SDC methods for microdata sets are based on one of two approaches [20, 1, 8]: sampling and perturbation. A sampling method suppresses part of the original microdata set; a perturbation method allows publication of the whole microdata set, but published values are distorted.

Microaggregation is a statistical disclosure control technique for *numerical* microdata which follows the substitution/perturbation approach [2, 5, 6, 9]. This technique is currently being used by a substantial number of European and worldwide statistical offices (e.g. see the recent UNECE survey [19]). Individual records are clustered into small aggregates or groups of size greater than or equal to a given security parameter k (usual values for k are 3 or 4). Rather

than publishing the value of a variable V_i for a given record, the average of the values of V_i over the group to which the record belongs is published. Therefore, aggregates should be as homogeneous as possible to minimize information loss. More formally, let us represent a microdata set with n records and d attributes as $X=\{x_1, x_2, ..., x_n\}$, where $x_i \in \mathbf{R}^d$ and let the security parameter k be a positive integer. A k-partition $P=\{C_1,C_2,...,C_{m(P)}\}$ of X is a partition where the size of group C_i, $1 \leq i \leq m(P)$ is at least k. Let $\mathbf{P}_k$ be the set of all k-partitions of X. Optimal microaggregation consists of finding a k-partition such that the sum of squared Euclidean distances from each x_i to the centroid $\bar{x}_{C_i} = \frac{1}{|C_i|}\sum_{x_l \in C_i} x_l$ is minimized, where C_i is the group to which x_i belongs. Formally the problem is

$$min_{P \in P_k} \sum_{i=1}^{m(P)} \sum_{x_j \in C_i} \left\| x_j - \bar{x}_{C_i} \right\|^2 \qquad (1)$$

where $\| \cdot \|$ denotes Euclidean norm and *m(P)* is *not* part of the input.

Even though no evidence was so far available in the literature that the above problem cannot in general be exactly solved in polynomial time, all proposed approaches to solving it have been of heuristic nature. Some heuristics are very simple: multivariate records are projected onto a single dimension before microaggregation and then all groups C_i (except perhaps one) are taken of the same size [2, 5, 6]. More sophisticated heuristics have been proposed in [9], which can deal with unprojected multivariate records and yield groups of variable size ≥ k; such heuristics draw on search techniques including hierarchical clustering, genetic algorithms and tabu search.

1.1 Our contribution

We will show that the microaggregation problem (1) is NP-hard. An NP-hard problem cannot be solved in polynomial time unless P=NP. Proving NP-hardness for microaggregation is most relevant, as it justifies the current use of heuristic approaches.

It must be pointed out that our result is different from seemingly similar NP-hardness results in the literature [14, 13, 18, 4, 12, 3, 17]. Most of such results deal with objective functions different from expression (1); even when a sum-of-squares objective function is considered, all known results concentrate on clustering problems consisting of dividing a set of points into a fixed number of clusters, and this without putting cardinality constraints on the groups. In the case of optimal microaggregation, the number of groups is *not* part of the problem input and there are cardinality constraints.

1.2 Plan of this paper

Section 2 gives background that will be useful for the proof. Section 3 contains the proof itself. Section 4 is a conclusion.

2 Background

Let $G=(V,E)$ be a graph, where $V=(v_1,v_2, ... v_n)$ is a finite set of vertices and E is a set of unordered pairs of vertices, the edges of G.

Definition 1
A *rectilinear planar layout of a graph G=(V,E)* is a representation of G in a plane, such that its vertices are mapped to horizontal line segments and the edges to vertical line segments, with all endpoints of segments at positive integer coordinates. Two horizontal segments are connected by a vertical segment if and only if the corresponding vertices are adjacent in the graph.

Proposition 1 (Rosenstiehl/Tarjan [16]).
Given a planar graph $G=(V,E)$, a rectilinear planar layout of G can be computed in time polynomial in the size of G. Moreover, the height and the width of the layout are both linear in the size of G. Without loss of generality, we may assume that all horizontal segments are at different (integer) heights.

Proposition 2 (Domingo-Ferrer/Mateo [9])
In an optimal solution of the microaggregation problem the groups have size greater than or equal to k and less than $2k$.

Proposition 3 (Edwards/ Cavalli-Sforza [11])
The sum of squares of a group of points is equal to the sum of all distances between pairs of entities of this group divided by its cardinality:

$$SSE = \sum_{x_j \in C_i} \left\| x_j - \bar{x}_{C_i} \right\|^2 = \frac{1}{|C_i|} \sum_{x_k, x_l \in C_i} \left\| x_k - x_l \right\|^2$$

3 Main result

Theorem
The microaggregation problem (1) is NP-hard.

Proof:
First we will show that microaggregation is NP-hard in the plane ($d=2$), and then we generalize this result for an d-dimensional Euclidean space with $d> 2$.

In [15] Pferschy, Rudolf and Woeginger have shown the NP-completeness of "Clustering into Triangles", that is, given a set P of points in a plane, $|P|=3p$, partition them into p triangles such that the sum of perimeters of all triangles is minimized. We will use their same construction to prove NP-hardness for microaggregation.

The Planar Exact Cover Problem by 3-Sets (Planar X3C), a special version of the Exact Cover Problem that was shown to be NP-complete by Dyer and Frieze [10] will be transformed to the microaggregation problem with $k=3$ and n a multiple of 3, say $n=3p$, where p is a positive integer. Planar X3C is defined as follows:

PLANAR EXACT COVER BY 3-SETS (Planar *X*3*C*)
Instance:
A set Q with $|Q|=3q$; a set T of triples from $Q \times Q \times Q$ such that (i) every element of Q occurs in at most three triples and such that (ii) the induced graph is planar. (This induced graph G is defined as the graph containing a vertex for every element of Q and for every triple in T. There is an edge connecting a triple to an element if and only if the element is a member of the triple. Clearly, G is bipartite with vertex bipartition Q and T).
Question:
Does there exist a subset of q triples in T which contains all the elements of Q?

Hence, let Q and $T \subset Q \times Q \times Q$ constitute an instance of planar *X*3*C*. We will construct a point set $S(Q,T)$, where $|S(Q,T)|$ is multiple of 3, that allows a partitioning into groups C_i, with $|C_i| \geq k=3$ and *SSE* minimal if and only if the planar *X*3*C* instance has a solution.

In a first step, a rectilinear planar layout for the underlying undirected graph G is computed according to Proposition 1. Then we multiply all coordinates by a factor of a large positive integer (for example 1000) in order to ensure that points on distinct horizontal (vertical) segments are sufficiently far away from each other. Next, we define the point set $S(Q,T)$. This construction is based on the triangle ∇_0 with side lengths 3,4,5 that will allow to keep all points in $S(Q,T)$ at integer coordinates. For every element of Q, there is a so-called element point in $S(Q,T)$ and for every triple in T there are three so-called triple points forming a so-called triple triangle in $S(Q,T)$. The triple triangles are that basic triangles ∇_0 with sides 3,4,5 such that the two sides of lengths 3 and 4 are axes-parallel. The element points and the triple triangles are placed somewhere at the corresponding line segments in the rectilinear layout (because of the multiplication, there is ample space to place them).

In the next step, we consider some triple $T_i = (q_1, q_2, q_3)$ in T and the corresponding three triple points t_1, t_2, t_3 that form a triangle ∇_0. For $1 \leq i \leq 3$, the point t_i is connected to the point q_i by a chain of diamonds as it is shown in Figure 1. A diamond consists of two copies of ∇_0 that are glued together either by their sides of length 5 (rectangular diamond) or by their sides of length 3 (triangular diamond). All diamonds are placed in such a way that the two shorter sides of the triangles are axes-parallel. No two rectangular diamonds occur consecutively in a chain. The chains of diamonds (roughly) follow the line segments corresponding to the two vertices t_i and q_i and to the connecting edge in the graph G.

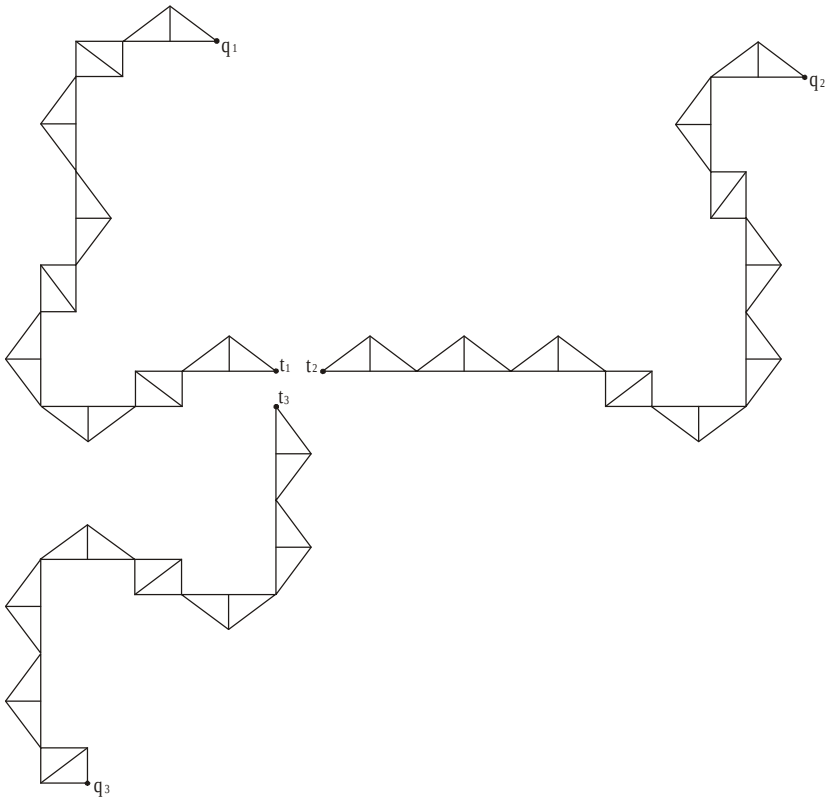

Figure 1: Chains of diamonds connecting t_i to q_i.

In such a construction two things have to be taken into consideration:

a) We do not want two distinct chains to come very close to each other. Again, because of the multiplication in the beginning, there are sufficiently many degrees of freedom to route the chains far away from each other. Since every element occurs in at most three triples, it is also possible to keep the chains sufficiently far from each other if they meet in an element point (Figure 2). The same holds for triple points.

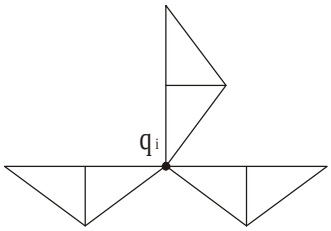

Figure 2: Placement of diamonds around an element point q_i

b) The other problem is that it is not *a priori* clear that triple and element points can indeed be connected by such a chains of diamonds: Triangular shaped diamonds shift the path by ±8 units in x-direction and 0 units in y-direction (or 0 units in x-direction and ±8 units in y-direction); rectangular shaped diamonds shift the path by a vector of (±3, ±4) or (±4, ±3). Once more, the multiplication in the beginning of the construction removes this problem. The main idea is to use only a small number of rectangular shaped diamonds in order to reach the correct remainders for the shift if divided by 8, and to use many triangular shaped diamonds for the long distances. As the distances are sufficiently large, we can mix rectangular shaped and triangular shaped diamonds without ending up in problem (a).
Connecting every triple triangle to its corresponding three element points completes the construction of the point set *S(Q,T)*. It is easy to see that this can be performed in polynomial time and $|S(Q,T)|$ is divisible by 3, say $|S(Q,T)|=3p$ (see [15]).
Now we have to prove that *S(Q,T)* can be partitioned into groups C_i , such that $|C_i| \geq 3$ and *SSE* minimal if and only if the Planar X3C has a solution.

Due to Proposition 2 we can reduce our search only to the groups of size 3, 4 and 5. So, first we will find the minimal *SSE* of a group of 3, 4 and 5 points in such a structure.
1. Groups of three points. By our construction no three points in *S(Q,T)* are at a the distance less than 3, 4 and 5 from each other, so using Proposition 3, min *SSE* of a group of three points is 50/3.
2. Groups of four points. All possible groups of four points with their relative *SSE* are presented in Figure 3. (We have considered only the groups of points in the order they come in the chain without skipping any point as it is obvious that otherwise the *SSE* of the resulting group would not be minimal). So we see that the group of four points that form a rectangular diamond has minimal *SSE*=25.
3. Groups of five points.The groups of five points are presented in Figure 4. As in the case of four points we have considered only points in the consecutive order. So the minimal *SSE* of the group of five points is 52.4.

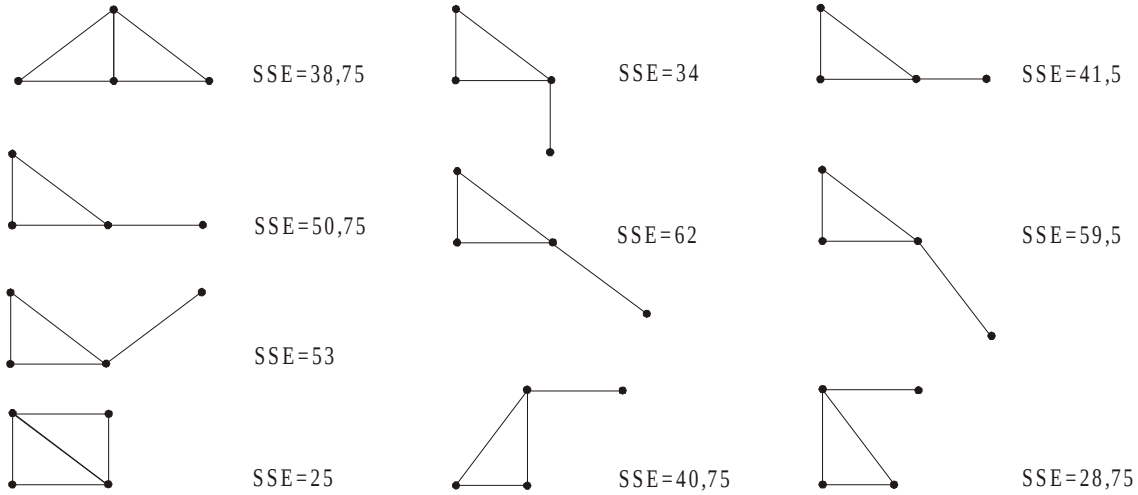

Figure 3. Clusters of 4 points

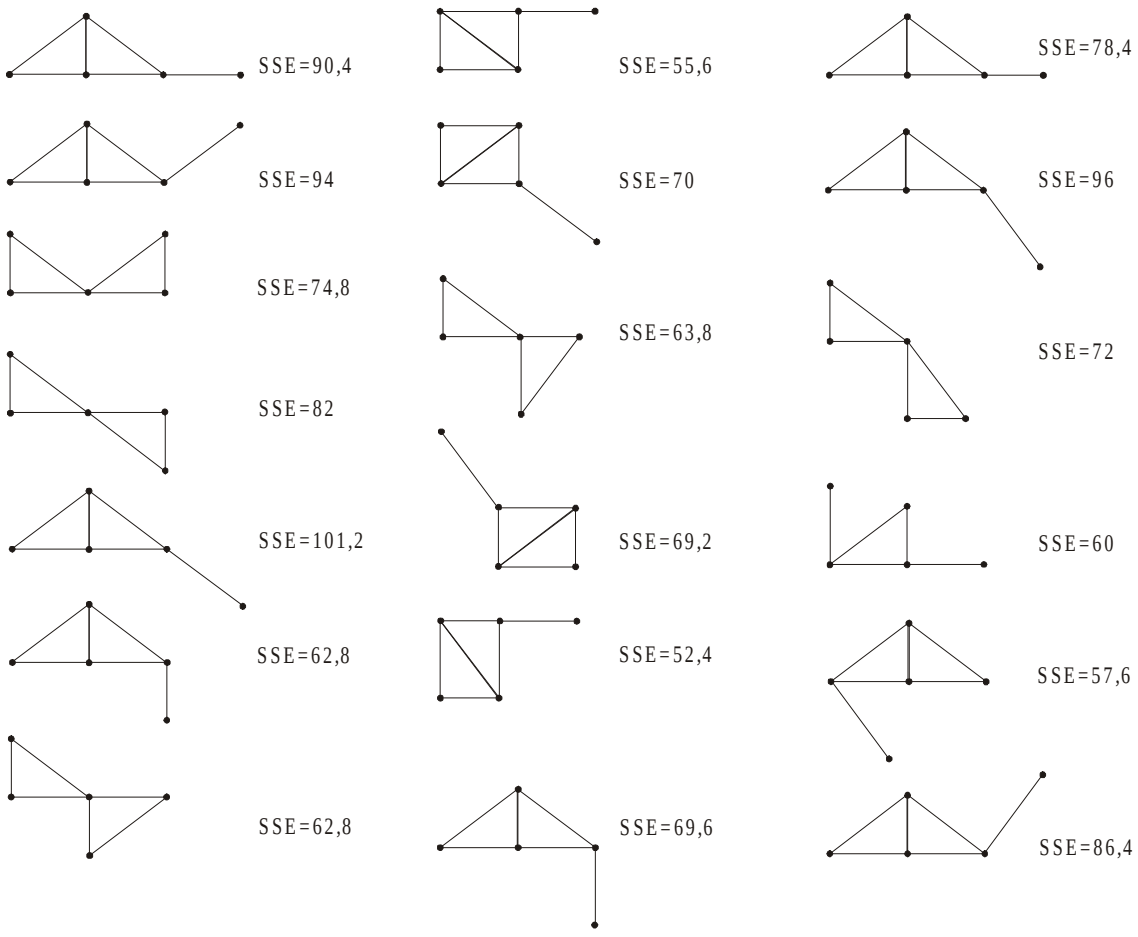

Figure 4. Clusters of 5 points

We claim that (i) the partitioning of *S(Q,T)* has the minimal *SSE* if all the groups are of size 3, and (ii) the partitioning into triangles with this minimal *SSE* is possible if and only if planar X3C allows a solution.

First we demonstrate the second claim and then the first.

ii) As we have noted above the minimal *SSE* of a group of 3 points is 50/3, when the points form a triangle with sides 3,4,5. So the minimal *SSE* that a partitioning into the groups all of size 3 may have is (50/3)*p*, where *p* is an integer that represents the number of groups.

(if) We claim that the subset T' of *T* that contains all triples for which the corresponding triple points form a cluster constitutes a solution to the planar X3C.

Consider some element point $q_i \in Q$. The corresponding element point is contained in exactly one cluster, and this cluster belongs to a chain of diamonds. In this chain of diamonds, every other triangle must form a cluster; therefore, the corresponding point t_i on the other end of the chain cannot be covered by any triangle cluster in the chain, and the corresponding three triangle points must form another cluster.

The clusters in the other (one or two) chains going away from q_i cover the corresponding triangle points, and these triangle points cannot form a cluster in the clustering. This way we may assign to each $q_i \in Q$ a unique triple in *T*. On the other hand, if we assign one q_i to some triple, the other two elements in this triple must be assigned to this triple, too. Clearly, this yields a solution to the X3C.

(only if) Now assume that the planar X3C has a solution $T' \subseteq T$. We construct a clustering into triangles as follows. All triangles corresponding to triples in T' are clusters, all triangles in $T \setminus T'$ are not. This completely determines which triangles on the chains have to be used in the clustering. Since T' is a solution of X3C, every element point is in exactly one cluster.

So the *SSE* of total construction with all the groups of size 3 is:

$$SSETOT_3 = \sum_{T_i \in T'} \sum_{j=1}^{3} (50/3) N_{ij} + (50/3)|T'| + \sum_{T_i \in T \setminus T'} \sum_{j=1}^{3} (50/3) N_{ij} = (50/3)(q + \sum_{i=1}^{|T|} \sum_{j=1}^{3} N_{ij}) \quad (2)$$

where N_{ij} is the number of diamonds in the chain *j* of the triple T_i,
for $1 \le j \le 3$ and $1 \le i \le |T|$.

i) Now we will show that the partitioning of *S(Q,T)* has the minimal *SSE* if all groups are of size 3.

Suppose, for the sake of contradiction, that a partitioning into the groups of size 3,4,5 exists with $SSE \le SSETOT_3$. Suppose that there are *x* groups of size 3, *y* groups of size 4 and *z* groups of size 5, and $y \neq 0$ or $z \neq 0$ in such a partition. As noted above, the minimal *SSE* is 25 for a group of 4 points and is 52.4 for a group of 5 points; on the other hand , in our construction $|S(Q,T)| = 3p$, so we have:

$$\begin{cases} 3x + 4y + 5z = 3p \\ 50/3\, x + 25y + 52.4z \le 50/3\, p \end{cases}$$

$$\begin{cases} p = x + 4/3\, y + 5/3\, z \\ p \ge x + 1.5y + 3.144z \end{cases}$$

$\frac{4}{3}y + \frac{5}{3}z \geq 1.5y + 3.144z$ - A contradiction has been reached.

So, we have shown that if a partitioning of *S(Q,T)* into the groups C_i such that $|C_i| \geq 3$ has the minimal *SSE* (equal to $SSETOT_3$ IN EXPRESSION (2)) then all groups are of size three. And a partitioning of *S(Q,T)* into the groups of size three with minimal *SSE* can be obtained if and only if Planar X3C has a solution. So the partition into the groups C_i with $|C_i| \geq 3$ has a minimal *SSE* if and only if Planar X3C has a solution. Thus, an NP-complete problem (Planar X3C) is no harder than the microaggregation problem. Therefore microaggregation is NP-hard in a plane.

The NP-hardness of microaggregation in a *d*-dimensional Euclidean space follows from the NP-hardness of the same problem in the plane. (By induction, if the problem is NP-hard in a *(d-1)*-dimensional space we can construct an instance of the problem in an *d*-dimensional space by adding a zero *d*-th coordinate to all points. This instance of the microaggregation problem in dimension *d* has a solution if and only if the initial *d-1*-dimensional problem has a solution. So *d*-dimensional microaggregation is NP-hard for *d>2* as well). QED

4 Conclusions

Microaggregation is a technique widely used for preserving confidentiality of individual respondents' numerical data. We have shown that the problem of optimal microaggregation, where optimal means yielding minimal information loss for a fixed security level, is NP-hard for dimension greater than or equal to two. This provides theoretical justification for the use of heuristic approaches (e.g. [9]) when attempting to solve this problem, since in general, no exact and efficient methods can be devised.

Acknowledgments

This work was partly funded by the European Commission under 5th FP RTD project IST-2000-25069 "CASC". Thanks go to Ulrich Pferschy, Ricard Gavaldà and Francesc Sebé for useful advice and discussions during the preparation of this paper.

References

[1] N.R. Adam and J.C. Wortmann, "Security-control methods for statistical databases: a comparative study", *ACM Computing Surveys*, vol.21, pp. 515-556, 1989.

[2] N.Anwar, *Micro-Aggregation - The Small Aggregates Method*, internal report, Luxembourg: Eurostat, 1993.

[3] E. Boros and P.L. Hammer, "On clustering problems with connected optima in Euclidean spaces", *Discrete Mathematics*, vol. 75, pp. 81-88, 1989.

[4] P.Brucker, "On the complexity of clustering problems", in *Optimization and Operations Research*, eds. R. Hehn, B. Korte and W. Oettli, Berlin: Springer-Verlag, pp. 45-54, 1977.

[5] 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, pp. 195-204, 1993.

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

[7] D.E. Denning and P.J. Denning, "Data security", *ACM Computing Surveys*, vol. 11, pp. 227-249, 1979.

[8]D.E. Denning, *Cryptography and Data Security*, Reading MA: Addison-Wesley, 1982.

[9] J.Domingo-Ferrer and J.M. Mateo-Sanz "Practical data-oriented microaggregation for statistical disclosure control", *IEEE Transactions on Knowledge and Data Engineering* (to appear, March 2002).

[10] M.E. Dyer and A.M.Frieze, "Planar 3DM is NP-complete", *Journal of Algorithms*, vol.7, pp.174-184, 1986.

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

[12] D.S. Johnson, "The NP-completeness column: an ongoing guide", *Journal of Algorithms*, vol. 7, pp. 174-184, 1986.

[13] N. Megiddo and A.Tamir, "On the complexity of locating linear facilities in the plane", *Operations Research Letters*, vol.13, pp.194-197, 1982.

[14] N. Megiddo and K.J. Supowit, "On the complexity of some common geometric location problems", *SIAM Journal of Computing*, vol.13, pp.182-196, 1984.

[15] U.Pferschy, R.Rudolf, G.J.Woeginger, "Some geometric clustering problems", *Nordic Journal of Computing*,, vol.1, pp.246-263, 1994.

[16] P.Rosenstiehl and R.E. Tarjan, "Rectilinear planar layout of planar graphs and bipolar orientations", *Discrete Computational Geometry*, vol.1, pp. 343-353, 1986.

[17] T. Schreiber, "Clustering for data reduction and approximation", *Computer Graphics Geometry*, vol. 2, no. 2, 2000. http://eos.wdcb.ru/cgg/1/1.htm

[18] K.J. Supowit, *Topics in Computational Geometry*, Ph. D. Thesis, Dept. of Computer Science, University of Illinois at Urbana-Champaign, Report UIUCDCS-R-81-1062, 1981.

[19] United Nations Economic Commission for Europe Secretariat, "Statistical data confidentiality in the transition countries: 2000/2001 winter survey", in *2nd UN/ECE and Eurostat Joint Work Session On Statistical Data Confidentiality*, Skopje: March 2001.

[20] L. Willengorg and T. De Waal, *Elements of Statistical Disclosure Control*, New York: Springer-Verlag, 2001.