Optimal Multivariate 2-Microaggregation for Microdata Protection: A 2-Approximation

Josep Domingo-Ferrer and Francesc Sebé

Rovira i Virgili University of Tarragona
Department of Computer Engineering and Maths
Av. Països Catalans 26
E-43007 Tarragona, Catalonia
{josep.domingo,francesc.sebe}@urv.cat

Abstract. Microaggregation is a special clustering problem where the goal is to cluster a set of points into groups of at least k points in such a way that groups are as homogeneous as possible. Microaggregation arises in connection with anonymization of statistical databases for privacy protection (k-anonymity), where points are assimilated to database records. A usual group homogeneity criterion is within-groups sum of squares minimization SSE. For multivariate points, optimal microaggregation, *i.e.* with minimum SSE, has been shown to be NP-hard. Recently, a polynomial-time $O(k^3)$-approximation heuristic has been proposed (previous heuristics in the literature offered no approximation bounds). The special case $k = 2$ (2-microaggregation) is interesting in privacy protection scenarios with neither internal intruders nor outliers, because information loss is lower: smaller groups imply smaller information loss. For 2-microaggregation the existing general approximation can only guarantee a 36-approximation. We give here a new polynomial-time heuristic whose SSE is at most twice the minimum SSE (2-approximation).
Keywords: Clustering, Statistical databases, Statistical disclosure control, Privacy-preserving data mining, Microaggregation.

1 Introduction

Microaggregation [7, 8] is a technique for privacy in statistical databases, a discipline also known as statistical disclosure control (SDC). It is used to mask individual records in view of protecting them against re-identification. More generally, microaggregation can be mathematically modeled as a special kind of clustering problem where the goal is to cluster a set of p-dimensional points (the records in the SDC application) into groups of at least k points in such a way that groups are as homogeneous as possible. For the sake of concreteness, we talk about records rather than points in what follows.

Let $\mathbf{X}$ be a p-dimensional dataset formed by n records, that is, the result of observing p attributes on n individuals. Attributes will be assumed

numerical (continuous) in this paper. Microaggregation is operationally defined in terms of two steps. Given a parameter k, the first step partitions records of $\mathbf{X}$ into groups of at least k records each. The second step replaces each record by the centroid of its group to obtain the masked dataset $\mathbf{X}'$. In a microaggregated dataset, no re-identification within a group is possible, because all k records in a group are identical: the best that an intruder can hope is to track what is the group where a target individual has been masked into.

Microaggregating with minimum information loss has been known to be an important —and difficult— issue ever since microaggregation was invented as an SDC masking method for microdata. However, it was often argued that optimality in SDC is not just about minimum information loss but about the best tradeoff between low information loss and low disclosure risk. The recent application [11] of microaggregation to achieve k-anonymity [21, 20, 24, 25] for numerical microdata leaves no excuse to circumvent the problem of trying to reduce information loss as much as possible: once a value k is selected that keeps the re-identification risk low enough, the only job left is to k-anonymize (that is, to microaggregate) with as little information loss as possible.

A partition P such that all of its groups have size at least k is called a k-partition [8] and microaggregation with parameter k is sometimes denoted as k-microaggregation.

In [8], optimal microaggregation is defined as the one yielding a k-partition maximizing the within-groups homogeneity. The rationale is that, the more homogeneous the records in a group, the less variability reduction when replacing those records by their centroid (average record) and thus the less information loss. The within-groups sum of squares SSE is a usual measure of within-groups homogeneity in clustering [27, 12, 15, 16], so a reasonable optimality criterion for a k-partition $P = \{G_1, \ldots, G_g\}$ is to minimize SSE, *i.e.* to minimize

$$SSE(P) = \sum_{i=1}^{g} \sum_{j=1}^{|G_i|} (x_{ij} - c(G_i))'(x_{ij} - c(G_i))$$

where $|G_i|$ is the number of records in the i-th group, $c(G_i)$ is the mean record (centroid) over the i-th group and x_{ij} is the j-th record in the i-th group. It was shown in [8] that groups in the optimal k-partition have sizes between k and $2k - 1$.

The optimal microaggregation problem has been shown to be NP-hard in the multivariate case, that is, when $p > 1$ ([19]). Therefore, algorithms for multivariate microaggregation are heuristic [6, 8, 22, 17, 18].

1.1 Contribution and plan of this paper

In [10], the first approximation algorithm in the literature to optimal multivariate microaggregation was described. For any integer $k \geq 2$, the SSE of the k-partition P provided by the heuristic given in [10] is shown to verify

$$SSE(P) \leq 2 \left\lceil \frac{3k-3}{2} \right\rceil (3k-3)(2k-1)SSE(P^{opt})$$

where P^{opt} is the optimal k-partition.

When using microaggregation to protect a dataset, the lower k, the lower SSE and the less information loss caused. Define an internal intruder as an intruder who has contributed one or more records to the dataset. In the presence of internal intruders or outliers, $k > 2$ should be chosen, so that an internal intruder cannot exactly guess the contribution of the other individual in her/his group. However, if internal intruders are unlikely and there are no outliers, a value as low as $k = 2$ would do for anonymity (2-anonymity): groups of records of size between $k = 2$ and $2k-1 = 3$ are formed and each record in a group is replaced by the group average record (2-microaggregation).

Thus 2-microaggregation is a relevant case deserving specific attention. For $k = 2$, the heuristic in [10] guarantees a bound $SSE(P) \leq 36 \cdot SSE(P^{opt})$, even though empirical results show that $SSE(P)$ is usually well below that bound. We propose in this paper a new heuristic for 2-microaggregation yielding a 2-partition P for which we can prove that $SSE(P) \leq 2 \cdot SSE(P^{opt})$.

Section 2 gives some background on the minimum-weight $[1, 2]$-factor problem and its algorithmic solution. Section 3 presents the 2-approximation heuristic for 2-microaggregation. The 2-approximation bound is proven in Section 4. Section 5 gives empirical results on the actual performance of the 2-approximation heuristic. Section 6 is a conclusion.

2 Background: the minimum-weight $[1, 2]$-factor problem

Given a graph $G = (V, E)$ and a function $w : E \rightarrow \mathbb{R}$ that assigns a weight to each edge, the minimum-weight $[1, 2]$-factor problem consists of finding the spanning subgraph F_{min} of G that satisfies:

- Each node in F_{min} has degree 1 or 2
- The sum of weights of edges in F_{min} is minimum.

This problem can be solved in strongly polynomial time [23], *i.e.* in running time bounded polynomially by a function only of the inherent dimensions of the problem (number of edges and nodes) and independent of the sizes of the numerical data. The graph library GOBLIN [13] solves the minimum weight factor problem over a weighted graph $G = (V, E)$ (with $|V| = n$ and $|E| = m$) by transforming the graph into a balanced flow network N_G [14] consisting of $n' = 2n+4$ nodes and $m' = 2m+4n+6$ edges and solving a minimum weight balanced flow over N_G. This solution determines a minimum weight factor over G. The Enhanced Primal Dual Algorithm [14] solves this problem in $O(n'^2 m')$ time.

In this way, a [1,2]-factor problem over a complete graph G having n nodes (and $n(n-1)/2$ edges) is solved polynomially in $O(n^4)$ time.

3 A 2-approximation algorithm for 2-microaggregation

Next, we present a 2-approximation for the multivariate 2-microaggregation problem. Our solution adapts for 2-microaggregation a corrected version of the [1] and [2] approach to 2-anonymizing categorical data through partial suppression. The algorithm in [1] and [2] could not be used as published, as it relies on [4] to solve a minimum weight $[1, 2]$-factor, a problem not dealt with by [4], but by the substantially more recent literature mentioned in Section 2.

Algorithm 1 (2-μ-Approx)

1. *Given a dataset* $\mathbf{X}$*, build a weighted complete graph* $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ *as follows:*
   (a) *Each record of* $x \in \mathbf{X}$ *is mapped to a different vertex* $v \in \mathbf{V}$*.*
   (b) *Given two vertices* $v, v' \in \mathbf{G}$ *corresponding to records* $x, x' \in \mathbf{X}$*, the edge* $vv' \in E$ *(the one linking nodes* v *and* v'*) is assigned weight* $w(vv') = d(x, x')^2$*, where* $d(x, x')$ *is the Euclidean distance between* x *and* x'*.*
2. *Compute the minimum weight [1,2]-factor* $\mathbf{F}_{min}$ *of graph* $\mathbf{G}$ *(see Section 2). By optimality,* $\mathbf{F}_{min}$ *consists only of connected components with a single edge (two vertices) or two adjacent edges (three vertices).*
3. *The 2-partition* P *is obtained by mapping each connected component in* $\mathbf{F}_{min}$ *to the group in* P *containing the records corresponding to the vertices in the component.*
4. *Microaggregate* $\mathbf{X}$ *based on* P*.*

4 The 2-approximation bound

We exploit in this section the properties of Algorithm 2-μ-Approx to prove that it yields a 2-approximation to optimal 2-microaggregation. We first give some notation, then a preliminary lemma and finally the theorem with the approximation bound.

Given a 2-partition $P = \{G_1, \ldots, G_g\}$ of $\mathbf{X}$, such that all groups have size 2 or 3, we denote by $SSE(G_i)$ the within-group sum of squares of group G_i, that is, $SSE(G_i) = \sum_{j=1}^{|G_i|}(d(x_{ij}, c(G_i)))^2$, where $c(G_i)$ is the centroid of group G_i.

Consider the complete graph $\mathbf{G}$ built in Step 1 of Algorithm 2-μ-Approx. Define $T(G_i)$ as the minimum-weight component of a [1,2]-factor of $\mathbf{G}$ containing the vertices corresponding to records in G_i. If G_i consists of two records, $T(G_i)$ contains a single edge connecting the two corresponding vertices. If G_i consists of three records, $T(G_i)$ contains the minimum-weighted two edges connecting the three corresponding vertices.

Lemma 1. *For any group $G_i \in P$ consisting of two or three records, it holds that*

$$\frac{1}{2} \leq \frac{SSE(G_i)}{w(T(G_i))} \leq 1$$

where $w(T(G_i))$ is the sum of weights of edges in $T(G_i)$.

Proof:

i) Consider a two-record group $G_i = \{x_{i1}, x_{i2}\}$ and let $d(x_{i1}, x_{i2})$ be the Euclidean distance between both records. When microaggregating G_i, both records will be replaced by their mean vector $c(G_i)$ (*i.e.*, the centroid of G_i). It holds that

$$d(x_{i1}, c(G_i)) = d(x_{i2}, c(G_i)) = \frac{d(x_{i1}, x_{i2})}{2}$$

Thus,

$$SSE(G_i) = 2 \cdot \left(\frac{d(x_{i1}, x_{i2})}{2}\right)^2 = \frac{(d(x_{i1}, x_{i2}))^2}{2}$$

On the other hand, by construction of the graph $\mathbf{G}$ in Algorithm 2-μ-Approx, we have $w(T(G_i)) = (d(x_{i1}, x_{i2}))^2$. Thus, for any group G_i with two records it holds that

$$\frac{SSE(G_i)}{w(T(G_i))} = 1/2.$$

ii) Let us now take a three-record group $G_i = \{x_{i1}, x_{i2}, x_{i3}\}$. Its corresponding minimum weight tree $T(G_i)$ consists of three vertices v_{i1}, v_{i2}, v_{i3} and the two minimum-weight edges connecting them. Let us denote $d(x_{i1}, x_{i2}) = d_1$, $d(x_{i1}, x_{i3}) = d_2$ and $d(x_{i2}, x_{i3}) = d_3$. It is well known that, in a triangle, the sum of the squared lengths of the sides is three times the sum of the squared vertex-centroid distances. In our notation, this equality can be written as

$$d_1^2 + d_2^2 + d_3^2 = 3 \cdot SSE(G_i) \tag{1}$$

Without loss of generality, we consider that the edges of $T(G_i)$ are $e_1 = v_{i1}v_{i2}$ and $e_2 = v_{i1}v_{i3}$. By the minimality of $T(G_i)$ and using Equation (1), we get

$$w(T(G_i)) = d_1^2 + d_2^2 \leq (2/3)(d_1^2 + d_2^2 + d_3^2) = 2 \cdot SSE(G_i)$$

Thus,

$$\frac{SSE(G_i)}{w(T(G_i))} \geq 1/2$$

Another fact of elementary geometry is that, for any triangle, the sum of the squared lengths of any two sides is at least one third of the sum of the squared lengths of three sides. Using this, we can write

$$w(T(G_i)) = d_1^2 + d_2^2 \geq (1/3)(d_1^2 + d_2^2 + d_3^2) = SSE(G_i)$$

Thus,

$$\frac{SSE(G_i)}{w(T(G_i))} \leq 1$$

□

Theorem 1 (2-Approximation bound). *If P is a 2-partition found by Algorithm 2-μ-Approx and P^{opt} is the optimal 2-partition, then $SSE(P) \leq 2 \cdot SSE(P^{opt})$.*

Proof: Consider the minimum weight [1,2]-factor $\mathbf{F}_{min}$ of graph $\mathbf{G}$ computed at Step 2 of Algorithm 2-μ-Approx. Let us denote its cost, that is the sum of its edge weights, as $w(\mathbf{F}_{min})$. By Lemma 1, for any group $G_i \in P$ it holds that

$$SSE(G_i) \leq w(T(G_i)) \tag{2}$$

Extending Inequality (2) for all $G_i \in P$ and taking into account that $T(G_i)$ are the components of $\mathbf{F}_{min}$, we get

$$SSE(P) \leq w(\mathbf{F}_{min}) \tag{3}$$

Let us now take the optimal k-partition P^{opt} for the dataset $\mathbf{X}$. For each group $G_i^{opt} \in P^{opt}$, we take its corresponding vertices in $\mathbf{G}$ and connect them with one edge (if G_i^{opt} consists of two records) or the two minimum-weighted adjacent edges (if G_i^{opt} consists of three records); call the resulting graph component $T(G_i^{opt})$. The union of all $T(G_i^{opt})$ is a non-minimum weight [1,2]-factor $\mathbf{F}$ for $\mathbf{G}$. By Lemma 1 we know that $w(T(G_i^{opt})) \leq 2 \cdot SSE(G_i^{opt})$. Applying this inequality to all clusters, we get

$$w(\mathbf{F}) \leq 2 \cdot SSE(P^{opt}) \tag{4}$$

On the other hand, by definition of minimum weight [1,2]-factor

$$w(\mathbf{F}_{min}) \leq w(\mathbf{F}) \tag{5}$$

If we combine Inequalities (3),(4) and (5), the 2-approximation bound of the theorem follows. □

5 Empirical results

We will show in this section that the new 2-approximation heuristic can perform even better than the best microaggregation heuristics in the literature in terms of low within-groups sum of squares SSE. We have used two reference datasets from the European project "CASC" [3]:

- The "Tarragona" dataset contains 834 records with 13 numerical attributes corresponding to financial information on 834 companies located in the area of Tarragona, Catalonia. The "Tarragona" dataset was used in the "CASC" project and in [8, 18, 10].
- The "EIA" dataset contains 4092 records with 11 numerical attributes (plus two additional categorical attributes not used here). This dataset was used in the "CASC" project and in [5, 10] and partially in [18] (an undocumented subset of 1080 records from "EIA", called "Creta" dataset, was used in the latter paper). For the sake of speed, we have used in our experiments reported below a block with only the first 600 records of the "EIA" dataset; we call "EIA-600" the resulting dataset.

Table 1 gives the information loss under various methods for different values of k. For each case, SSE and $L_{SSE} = 100 \times SSE/SST$ are given, where SST is the total sum of squares (sum of squared Euclidean distances from all records to the dataset centroid). The advantage of L_{SSE} is that it is bounded within the interval $[0, 100]$. The methods considered in the comparison include the best heuristics in the literature, according to the comparison in [10], namely:

– An improved version of the heuristic in [8] called MDAV (Maximum Distance to Average Vector, [11]). MDAV is the microaggregation method implemented in the μ-Argus package [17] resulting from the "CASC" project.
– The μ-Approx general approximation heuristic described in [10].
– The 2-μ-Approx heuristic proposed in this paper.

It can be seen that 2-μ-Approx yields the lowest SSE for the "Tarragona" dataset. For the "EIA-600" dataset, 2-μ-Approx ranks second after MDAV. Anyway, the differences in terms of L_{SSE} are not really meaningful. Furthermore, note that even if MDAV can slightly outperform the approximation heuristics for particular datasets, the latter have the advantage of always guaranteeing an SSE within a known multiple of the minimum SSE; this is especially valuable when that multiple is as small as twice the minimum SSE, as is the case for 2-μ-Approx.

The price paid to get the 2-approximation is that, since 2-μ-Approx basically requires to solve a minimum-weight $[1, 2]$-factor, it runs in time $O(n^4)$ (see Section 2), whereas MDAV and the general approximation μ-Approx run in $O(n^2)$. For example, the time needed to run 2-μ-Approx on the EIA-600 dataset is 81 minutes and 17 seconds, whereas MDAV and mu-Approx take a few seconds. Nonetheless, this is less painful than it would appear at first sight: all heuristics being at least quadratic-time, blocking attributes must always be used to microaggregate large datasets, so the only adaptation needed to run an $O(n^4)$ heuristic is to take smaller blocks.

Table 1. Information loss measures for the "Tarragona" and "EIA-600" datasets under various microaggregation heuristics ($k = 2$)

	Method	SSE	L_{SSE}
"Tarragona"	MDAV	1005.59	9.27499
	μ-Approx	1148.32	10.5914
	2-μ-Approx	958.496	8.84058
"EIA-600"	MDAV	59.2535	1.06927
	μ-Approx	66.8219	1.20585
	2-μ-Approx	65.8851	1.18895

Finally, we give some experimental results on how close to optimality are the partitions obtained using 2-μ-Approx. In order to be able to find

the optimal 2-partition by exhaustive search, we are constrained to using very small datasets. We have taken the third reference dataset in [3], called "Census", which contains 1080 records with 13 numerical attributes and was used in the CASC project and [9, 5, 28, 18, 11, 10]. From the "Census" dataset, we have drawn 10 random samples of $n = 15$ records with $p = 13$ attributes each. Those samples have been 2-microaggregated optimally by exhaustive search and also heuristically using 2-μ-Approx. For each sample, Table 2 shows the optimal SSE, the SSE obtained with 2-μ-Approx and the ratio between the former and the latter. It can be seen that such a ratio is 1 or close to 1 in all cases. Thus, even if the approximation bound only guarantees that the SSE obtained with 2-μ-Approx is no more than twice the optimum, it actually tends to be very close to the optimum.

Table 2. Optimal SSE vs SSE obtained with 2-μ-Approx ($k = 2$) for 10 random samples drawn from the "Census" dataset ($n = 15$ and $p = 13$).

Sample	Optimal SSE	SSE 2-μ-Approx	Ratio
1	12.042	12.042	1
2	12.2066	12.8186	0.9522
3	14.8156	14.8156	1
4	12.5545	12.5545	1
5	51.6481	52.0665	0.9920
6	74.1998	74.1998	1
7	15.6783	16.5705	0.9462
8	9.9702	9.9702	1
9	21.4293	22.7064	0.9438
10	33.3882	33.3882	1

6 Conclusion

The polynomial-time 2-approximation presented here improves for $k = 2$ on the general $O(k^3)$-approximation for multivariate microaggregation. Even though 2-microaggregation is not usable if internal intruders are likely or outliers are present, it can be an interesting option to implement 2-anonymity in other cases, because it results in low information loss and thus in high data utility. Thus, the availability of a 2-approximation for 2-microaggregation is relevant. Suggested directions for future research include: i) to devise heuristics that, for specific values of k other than

2, provide better approximations than the general $O(k^3)$-approximation; ii) to adapt Algorithm 2-μ-Approx to come up with an approximation to 2-microaggregation of non-numerical (categorical) microdata (categorical microaggregation was defined in [26]).

Acknowledgments

Thanks go to David Hartvigsen and Gérard Cornuéjols for their advice on the $[1, 2]$-factor literature. We also acknowledge partial support by the Government of Catalonia under grant 2005 SGR 00446, by the Spanish Ministry of Science and Education under project SEG2004-04352-C04-01 "PROPRIETAS" and by Eurostat-European Comission under contract "CENEX-SDC".

References

1. G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Anonymizing tables. In T. Eiter and L. Libkin, editors, *Proceedings of ICDT'2005*, volume 3363 of *Lecture Notes in Computer Science*, pages 246–258, Berlin Heidelberg, 2005.
2. G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Approximation algorithms for k-anonymity. *Journal of Privacy Technology*, 2005. Paper no. 20051120001.
3. 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.
4. G. P. Cornuéjols. General factors of graphs. *Journal of Combinatorial Theory*, B45:185–198, 1988.
5. R. Dandekar, J. Domingo-Ferrer, and F. Sebé. Lhs-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 153–162, Berlin Heidelberg, 2002. Springer.
6. D. Defays and N. Anwar. Micro-aggregation: a generic method. In *Proceedings of the 2nd International Symposium on Statistical Confidentiality*, pages 69–78, Luxemburg, 1995. Eurostat.
7. 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*, pages 195–204, Ottawa, 1993. Statistics Canada.
8. J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189–201, 2002.
9. J. Domingo-Ferrer, J. M. Mateo-Sanz, and V. Torra. Comparing sdc methods for microdata on the basis of information loss and disclosure risk. In *Pre-proceedings of ETK-NTTS'2001 (vol. 2)*, pages 807–826, Luxemburg, 2001. Eurostat.
10. J. Domingo-Ferrer, F. Sebé, and A. Solanas. A polynomial-time approximation to optimal multivariate microaggregation. *Manuscript*, 2005.

11. J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogenerous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195–212, 2005.
12. A. W. F. Edwards and L. L. Cavalli-Sforza. A method for cluster analysis. *Biometrics*, 21:362–375, 1965.
13. C. Fremuth-Paeger. Goblin: A library for graph matching and network programming problems. release 2.7, 2005. http://www.math.uni-augsburg.de/opt/goblin.html.
14. C. Fremuth-Paeger and D. Jungnickel. Balanced network flows. vii: Primal-dual algorithms. *Networks*, 39(1):35–42, 2002.
15. A. D. Gordon and J. T. Henderson. An algorithm for euclidean sum of squares classification. *Biometrics*, 33:355–362, 1977.
16. P. Hansen, B. Jaumard, and N. Mladenovic. Minimum sum of squares clustering in a low dimensional space. *Journal of Classification*, 15:37–55, 1998.
17. A. Hundepool, A. Van 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*. Statistics Netherlands, Voorburg NL, may 2005. http://neon.vb.cbs.nl/casc.
18. M. Laszlo and S. Mukherjee. Minimum spanning tree partitioning algorithm for microaggregation. *IEEE Transactions on Knowledge and Data Engineering*, 17(7):902–911, 2005.
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*, 18(4):345–354, 2001.
20. P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.
21. P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI International, 1998.
22. G. Sande. Exact and approximate methods for data directed microaggregation in one or more dimensions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):459–476, 2002.
23. A. Schrijver, editor. *Combinatorial Optimization: Polyhedra and Efficiency. Volume A*. Springer Verlag, Berlin, 2003.
24. L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 10(5):571–588, 2002.
25. L. Sweeney. k-anonimity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 10(5):557–570, 2002.
26. V. Torra. Microaggregation for categorical variables: a median based approach. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases*, volume 3050 of *Lecture Notes in Computer Science*, pages 162–174, Berlin Heidelberg, 2004. Springer.
27. J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244, 1963.
28. W. E. Yancey, W. E. Winkler, and R. H. Creecy. Disclosure risk assessment in perturbative microdata protection. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 135–152, Berlin Heidelberg, 2002. Springer.