Facility Location and Social Choice via Microaggregation

Josep Domingo-Ferrer

Universitat Rovira i Virgili
Dept. of Computer Engineering and Mathematics
UNESCO Chair in Data Privacy
Av. Països Catalans 26
E-43007 Tarragona, Catalonia
josep.domingo@urv.cat

Abstract. Microaggregation is a cardinality-constrained clustering problem that arose in the context of data privacy. In microaggregation, the number of clusters is not fixed beforehand, but each cluster must have at least k elements. We illustrate in this paper that microaggregation can be applied for decision making in areas other than privacy. Specifically, we focus on the service facility location problem and on game theory (coalition formation and social choice).
Keywords: Microaggregation; Decision making; Service facility location problem; Cooperative games; Coalitions; Social choice.

1 Introduction

Microaggregation [2, 4] is a clustering problem that originally arose in data anonymization for privacy protection. Rather than fixing beforehand the number of desired clusters, in microaggregation one fixes the minimum size of clusters. Clusters are formed using a criterion of maximum within-cluster similarity and each cluster should contain at least k elements. In the anonymization application, elements are records corresponding to individual respondents and the records in a cluster are replaced by the centroid cluster before publication; this ensures that an individual's published record is indistinguishable from the records of at least another $k-1$ individuals (k-anonymity).

The optimal solution to the microaggregation problem is defined to be the one that maximizes the sum of within-cluster similarities, while respecting the constraint that all clusters must contain at least k elements. Finding an optimal solution for the microaggregation problem can be done in polynomial time only if the data elements are one-dimensional [7]. In the general multi-dimensional case, the problem has been shown to be NP-hard [14].

Several heuristics have been published in the literature that provide good solutions to the microaggregation problem. Most of them deal with numerical elements (*e.g.* see survey in [8]), but extensions for ordinal and nominal data have also been proposed [4–6]. Some of the proposed heuristics are approximations [3,

11] in the sense that they can be proven to yield solutions within a specific bound of the optimal solution.

The motivation of this paper is to explore applications of microaggregation other than anonymization; in particular, we describe several applications to decision making.

1.1 Contribution and plan of this paper

As pointed out above, the microaggregation problem is a well-studied NP-hard problem and several efficient heuristics aimed at solving it have been proposed.

Although microaggregation arose in the field of data anonymization, it is a general problem that can also arise in many other application areas. In this paper, we explore the use of microaggregation in facility location and game theory. Specifically, Section 2 describes how the location of service facilities can be viewed as a microaggregation problem. Section 3 illustrates two uses of microaggregation in game theory: detecting natural coalitions in cooperative games and reducing the number of strategies to facilitate rational social choice. Conclusions and avenues for future research are outlined in Section 4.

2 Microaggregation and service facility location

A well-known problem in operations research is the *simple plant location problem* (SPLP), also known under a variety of alternative names (warehouse location problem [9], uncapacitated facility location problem [10], etc.). The most popular statement of this problem is as follows:

- Let $I = \{1, \cdots, m\}$ be a set of candidate locations for industrial plants producing some product. A plant can be opened in any location $i \in I$ at a cost f_i. Each opened plant can provide an unlimited amount of product.
- Let $J = \{1, \cdots, n\}$ be a set of customers such that customer j needs an amount b_j the product.
- Let c_{ij} be the unit transportation cost from plant i to customer j.
- The problem is to decide at which locations should plants be opened and the quantity x_{ij} of product to be supplied by plant i to customer j.

Mathematically, the SPLP can be formulated as the following constrained minimization problem:

$$\min \left(\sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} x_{ij} + \sum_{i=1}^{m} f_i y_i \right)$$

subject to

$$\sum_{i=1}^{m} x_{ij} = b_j, \quad \forall j \in J$$

$$0 \leq x_{ij}/b_j \leq y_i \text{ and } y_i \in \{0,1\}, \quad \forall i \in I \text{ and } \forall j \in J$$

where y_i indicates whether a plant is opened at location i ($y_i = 1$) or not ($y_i = 0$).

Imagine now that, instead of industrial plants that produce some products, we want to find locations for service facilities that must give service to users. Examples of such service facilities could be hospitals, schools, sports facilities, etc. Let us call this problem the *service facility location problem* (SFLP). Let us point out some fundamental differences between the SPLP and the SFLP, which make the heuristics designed for SPLP unsuitable for SFLP:

- Whereas the SPLP assumes that what is transported is the product, what is transported in SFLP are the users who must reach their service facility to use it. Hence, even if transportations costs were extremely low in terms of money, users do not wish to travel long distances to reach their service facility.
- For the above reason, each user wishes their service facility to be as close as possible to them. However, for cost reasons, opening a service facility at each single user's location is not affordable. Rather, for the investment to be justified, each service facility must be shared by at least k users.

From the above observations, it follows that *the SFLP is in fact a microaggregation problem*, because:

1. Clusters of at least k user locations must be formed in order to locate a service facility at the centroid of each cluster.
2. The Euclidean distance from each user location to the location of the corresponding service facility must be as small as possible. This amounts to maximizing the sum of within-cluster similarities as pointed above, for the special case of within-cluster similarity being the inverse of the sum of the Euclidean distances from the user locations in a cluster to the cluster centroid where the facility is located.

Being equivalent to a multivariate microaggregation problem (locations are bivariate), the SFLP is an NP-hard problem [14]. Any microaggregation heuristic for multivariate numerical data (*e.g.* [2, 4, 3, 11]) can be used to find a reasonably good solution to the SFLP. In fact, the approximation heuristic in [3] offers a solution that can be proven to be within a factor of $(2k-1)\max(2k-1, 3k-5)$ of the optimal one, where optimality means minimum sum of within-cluster Euclidean distances from locations to centroids. Even better, the more recent approximation heuristic in [11] offers a solution within a factor of $8(k-1)$ of the optimal one. The actual solutions returned by all the above-mentioned heuristics are usually very close to the optimal solution; in particular they are much closer than guaranteed by the theoretical approximation bounds (which are worst-case).

Since most microaggregation heuristics run in time quadratic in the number n of elements (user locations in our case), it may be necessary to use blocking for large n. In the SFLP blocking means that, rather than solving the problem

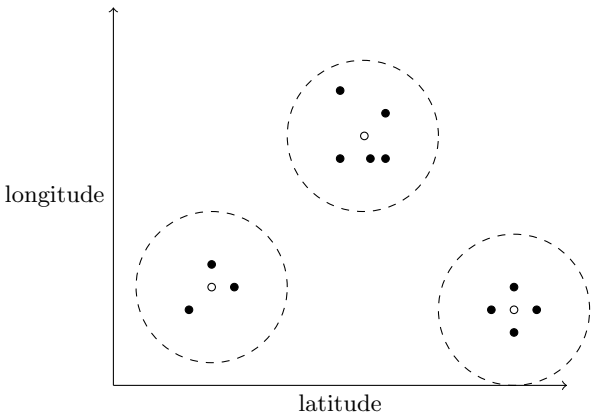

Fig. 1. Variable-size microaggregation with $k = 3$ to obtain a solution of the service facility location problem with $n = 12$ users. Black dots indicate user locations and white dots indicate proposed service facility locations.

for all users/citizens in a large geographic area (*e.g.* a country), one would independently solve instances of the problem in manageable subdivisions of the large area (*e.g.* in each state, province, county, etc.).

Figure 1 depicts an example with $n = 12$ users, a minimum of $k = 3$ users needed to justify a new service facility and service facility locations obtained with variable-size microaggregation.

3 Microaggregation and game theory

In game theory, a *cooperative game* is a game in which groups of players, called *coalitions*, may enforce cooperative behavior. Hence, the game can be viewed as a competition between coalitions of players, rather than between individual players.

We first give some background on game theory (Section 3.1). Then we describe how microaggregation can be used to detect natural coalitions in cooperative games (Section 3.2), and to reduce the number of strategies to facilitate rational social choice (Section 3.3).

3.1 Background on game theory

A game is a protocol between a set of n *players*, $\{P_1, \cdots, P_n\}$. Each player P_i has her own *set of possible strategies*, say S_i. To play the game, each player i selects a strategy $s_i \in S_i$. We will use $\mathbf{s} = (s_1, \cdots, s_n)$ to denote the vector of strategies selected by the players and $\mathbf{S} = \Pi_{i=1}^{n} S_i$ to denote the set of all possible ways in which players can pick strategies.

The vector of strategies $\mathbf{s} \in \mathbf{S}$ selected by the players determines the outcome for each player, which can be a payoff or a cost. In general, the outcome will be different for different players. To specify the game, we need to state for each player a preference ordering on these outcomes by giving a complete, transitive, reflexive binary relation on the set of all strategy vectors $\mathbf{S}$. The simplest way to assign preferences is by assigning, for each player, a value for each outcome representing the payoff of the outcome (a negative payoff can be used to represent a cost). A function whereby player P_i assigns a payoff to each outcome is called a utility function and is denoted by $u_i : \mathbf{S} \longrightarrow \mathbb{R}$.

For a strategy vector $\mathbf{s} \in \mathbf{S}$, we use s_i to denote the strategy played by P_i and s_{-i} to denote the $(n-1)$-dimensional vector of the strategies played by all other players. With this notation, the utility $u_i(\mathbf{s})$ can also be expressed as $u_i(s_i, s_{-i})$.

A strategy vector $\mathbf{s} \in \mathbf{S}$ is a *dominant strategy solution* if, for each player P_i and each alternate strategy vector $\mathbf{s}' \in \mathbf{S}$, it holds that

$$u_i(s_i, s'_{-i}) \geq u_i(s'_i, s'_{-i}) \tag{1}$$

In plain words, a dominant strategy $\mathbf{s}$ is the best strategy for each P_i, independently of the strategies played by all other players.

A strategy vector $\mathbf{s} \in \mathbf{S}$ is said to be a *Nash equilibrium* if, for any player P_i and each alternate strategy $s'_i \in S_i$, it holds that

$$u_i(s_i, s_{-i}) \geq u_i(s'_i, s_{-i})$$

In plain words, no player P_i can change her chosen strategy from s_i to s'_i and thereby improve her payoff, assuming that all other players stick to the strategies they have chosen in $\mathbf{s}$. A Nash equilibrium is self-enforcing in the sense that once the players are playing such a solution, it is in every player's best interest to stick to her strategy. Clearly, a dominant strategy solution is a Nash equilibrium. Moreover, if the solution is strictly dominant (*i.e.* when the inequality in Expression (1) is strict), it is also the unique Nash equilibrium. See [13] for further background on game theory.

3.2 Detecting natural coalitions in cooperative games

We assume in this section that the set of strategy vectors is finite, *i.e.*

$$\mathbf{S} = \{\mathbf{s}^1, \cdots, \mathbf{s}^m\}$$

We can represent each player P_i as an m-dimensional vector $u_i(\mathbf{S})$ whose components specify the normalized payoffs the player obtains under each strategy, according to his/her utility function $u_i(\cdot)$:

$$u_i(\mathbf{S}) = \left(\frac{u_i(\mathbf{s}^1)}{\max_{l=1}^{m} u_i(\mathbf{s}^l)}, \cdots, \frac{u_i(\mathbf{s}^m)}{\max_{l=1}^{m} u_i(\mathbf{s}^l)} \right)$$

We now can cluster vectors $u_i(\mathbf{S})$ to obtain clusters of players with "similar" interests, in the sense that they derive similar payoffs from the various strategies. Two important remarks are in order here:

– Clusters indicate "natural" coalitions, in the sense that players with similar interests may tend to rally: they wish to enforce the same strategy vectors and avoid the same strategy vectors[1].
– The "centroid player" of each cluster could be taken as the prototypical player representing the coalition of players in the cluster. In this way, a cooperative game involving the natural coalitions can be approximately transformed into a non-cooperative game between prototypical players. Finding solutions in non-cooperative games (that is, dominant strategy vectors or Nash equilibria mentioned above) is normally easier.

If a single cluster containing all players is created, the homogeneity of that cluster is likely to be low and the prototypical player of that cluster is unlikely to accurately represent the interests of all players. On the other hand, if some clusters are much smaller than others, the coalitions corresponding to the smaller clusters will have much less power to enforce dominant strategies or equilibria than the coalitions corresponding to the larger clusters. Hence, there is a tradeoff between the coalition power and the representativeness of the prototypical player. Resorting to microaggregation to form clusters ensures bounds on coalition sizes:

– If a fixed-size microaggregation heuristic is chosen (*e.g.* MDAV, [4]) then the sizes of all coalitions are k except for one coalition having size at most $2k-1$. Choosing this kind of heuristics establishes equal power (size) for all coalitions as a primary goal and prototype representativeness as a secondary goal.
– If a variable-size microaggregation heuristic is used (*e.g.* [3, 11]), then the sizes of all coalitions lie between k and $2k-1$, where the precise sizes are automatically selected by the heuristic in order to maximize within-cluster homogeneities. Choosing this kind of heuristic establishes prototype representativeness as a primary goal and equal power as a secondary goal.

Figure 2 depicts an example with two strategies, 12 players and three natural coalitions that can be formed when using variable-size microaggregation with $k = 3$.

3.3 Facilitating social choice

Social choice is a theoretical framework that studies how to combine individual preferences, interests or welfares to reach a collective decision or social welfare in some sense [1]. The Nakamura number [12] measures the degree of rationality of collective decision rules, such as voting rules. If the number of alternatives (candidates, options, etc.) to choose from is less than the Nakamura number,

[1] We discard here coalitions including players with similar utilities for the highest-paying strategy vectors *only*. The reason is that these are weaker coalitions, because they will break up if players not in the coalition manage to enforce a strategy vector that is not among the highest-paying ones for the coalition.

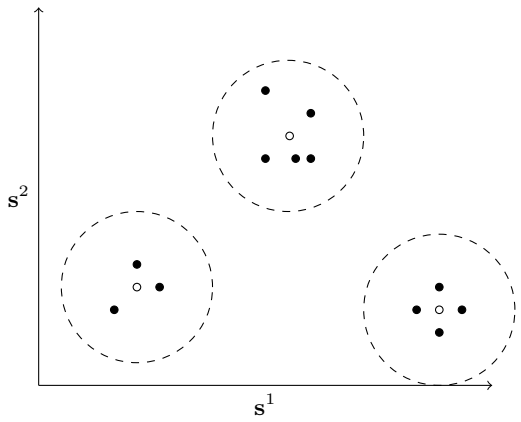

Fig. 2. Toy example with 12 players being clustered into 3 natural coalitions, for $m = 2$ strategy vectors $\mathbf{s}^1$ and $\mathbf{s}^2$, and variable-size microaggregation with $k = 3$. Black dots indicate actual players; white dots indicate the prototype player for each coalition; all coordinates are assumed to be normalized.

then the voting rule will identify "best" alternatives without any problem. In contrast, if the number of alternatives is greater than or equal to the Nakamura number, the voting rule will fail to identify "best" alternatives for some pattern of voting (*i.e.* for some tuple of voters' preferences), because a voting paradox will arise: a cycle of preferences will appear, like alternative a being socially preferred to alternative b, b to c and c to a.

The above discussion motivates the *relevance of being able to reduce the number of alternatives* in such a way that the new alternatives are as representative as possible of the old alternatives. We propose to use microaggregation to implement such a reduction.

Let us assimilate voters to players and alternatives to strategy vectors. We can represent each strategy vector $\mathbf{s}^j$ as an n-dimensional vector $\mathbf{u}(\mathbf{s}^j)$ whose components specify the normalized payoffs $\mathbf{s}^j$ brings to each player, according to the players' utility functions $u_1(\cdot)$ to $u_n(\cdot)$:

$$\mathbf{u}(\mathbf{s}^j) = \left(\frac{u_1(\mathbf{s}^j)}{\max_{l=1}^{m} u_i(\mathbf{s}^l)}, \cdots, \frac{u_n(\mathbf{s}^j)}{\max_{l=1}^{m} u_i(\mathbf{s}^l)} \right)$$

We now can cluster vectors $\mathbf{u}(\mathbf{s}^j)$ to obtain clusters of "similar" strategy vectors, in the sense that they provide similar payoffs to players/voters. The "centroid strategy" of each cluster can be taken as the prototypical strategy that will be used to replace the strategies in the cluster, thereby reducing the total number of strategies/alternatives.

If a single cluster containing all strategies is created, the homogeneity of that cluster is likely to be low and the prototypical strategy of that cluster is unlikely to accurately represent the interests of all players/voters (for example,think of

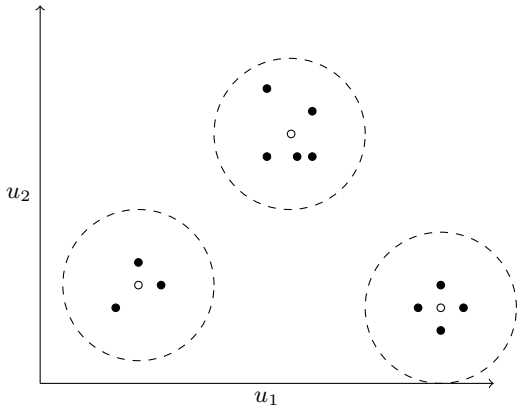

Fig. 3. Toy example with $n = 2$ voters/players with utility functions u_1 and u_2 in which 12 original alternatives/strategy vectors are reduced to 3 prototype alternatives/strategy vectors, using variable-size microaggregation with $k = 3$. Black dots indicate original strategies; white dots indicate prototype strategies; all coordinates are assumed to be normalized.

a country with a single party). On the other hand, if one goes for a less dramatical reduction of alternatives, it would seem fair to reduce the granularity of all original alternatives to a similar level, perhaps with more similar alternatives being included in larger clusters. This is exactly what variable-size microaggregation heuristics (*e.g.* [3, 11]) offer: the sizes of all clusters lie between k and $2k - 1$, where the precise sizes are automatically selected by the heuristic in order to maximize within-cluster homogeneities. Choosing this kind of heuristic seeks to obtain prototype strategies with similar representativeness of the original strategies. The smallest possible value of k ought to be taken that brings the final number of alternatives/strategy vectors below the desired threshold (for example, the Nakamura number of the game).

Figure 3 shows a toy example with $n = 2$ voters/players where 12 original alternatives/strategy vectors are reduced to 3 prototype alternatives/strategies, using variable-size microaggregation with $k = 3$.

4 Conclusions

Although microaggregation was a problem that arose and was studied in the context of data anonymization, we claim that it is relevant in other application domains. In this paper, we have sketched its application to decision making. Specifically, microaggregation heuristics have been shown to offer solutions to the service facility location problem. Also, microaggregation can be helpful in game theory. Indeed, in cooperative games it helps detecting natural coalitions (players with similar interests). In social choice it can be used to reduce the

number of alternatives with minimum loss of information, in order to facilitate rational voting.

Acknowledgments and disclaimer

This work was partly supported by the Government of Catalonia under grant 2009 SGR 1135, by the Spanish Government through projects TIN2011-27076-C03-01 "CO-PRIVACY" and CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", and by the European Comission under FP7 projects "DwB" and "Inter-Trust". The author is partially supported as an ICREA Acadèmia researcher by the Government of Catalonia. The author is with the UNESCO Chair in Data Privacy, but he is solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization.

References

1. K. J. Arrow. *Social Choice and Individual Values*. New Haven CT: Yale University Press, 1951.
2. 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.
3. J. Domingo-Ferrer, F. Sebé and A. Solanas. A polynomial-time approximation to optimal multivariate microaggregation. *Computers & Mathematics with Applications*, 55(4):714-732, 2008.
4. J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195-212, 2005.
5. J. Domingo-Ferrer. Marginality: a numerical mapping for enhanced exploitation of taxonomic attributes. In *Modeling Decisions for Artificial Intelligence-MDAI 2012* (eds. V. Torra, Y. Narukawa, B. López and M. Villaret), LNCS 7647, pp. 367-381. Berlin: Springer, 2012.
6. J. Domingo-Ferrer, D. Sánchez and G. Rufian-Torrell. Anonymization of nominal data based on semantic marginality. *Information Sciences*, 242:35-48, 2013.
7. S. L. Hansen and S. Mukherjee. A polynomial algorithm for optimal univariate microaggregation. *IEEE Transactions on Knowledge and Data Engineering*, 15(2): 1043-1044, 2003.
8. A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte Nordholt, K. Spicer, and P. P. De Wolf. *Statistical Disclosure Control*. New York: Wiley, 2012.
9. B. M. Khumawala. An efficient branch and bound algorithm for the warehouse location problem. *Management Science*, 18(12):B-718-B-731, 1972.
10. J. Krarup and P. M. Pruzan. The simple plant location problem: survey and synthesis. *European Journal of Operational Research*, 12(1):36-81, 1983.
11. M. Laszlo and S. Mukherjee. Approximation bounds for minimum information loss microaggregation. *IEEE Transactions on Knowledge and Data Engineering*, 21(11):1643-1647, 2009.

12. K. Nakamura. The vetoers in a simple game with ordinal preferences. *International Journal of Game Theory*, 8(1):55-61, 1979.
13. N. Nisan, T. Roughgarden, É. Tardos and V. V. Vazirani (eds.). *Algorithmic Game Theory*. New York: Cambridge University Press, 2007.
14. A. Oganian and J. Domingo-Ferrer. On the complexity of optimal microaggregation for statistical disclosure control. *Statistical Journal of the United Nations Economic Commission for Europe*, 18(4):345-354, 2001.