TRANSACTIONS ON DATA PRIVACY 10 (2017) 117–144

Fine granular proximity breach prevention during numerical data anonymization

Reza Mortazavi*, **Saeed Jalili****

*School of Engineering, Damghan University, Damghan, Iran, Tel.: +98-233-5220081-6 (321), Fax: +98-233-5220414.

**Computer Engineering Department, Tarbiat Modares University, Tehran, Iran, Tel.: +98-21-82883374

E-mail: r_mortazavi@du.ac.ir, sjalili@modares.ac.ir

Received 20 November 2016; received in revised form 5 May 2017; accepted 25 August 2017

Abstract. Microaggregation is known as a successful perturbative mechanism to realize k-anonymity. The method *partitions* the dataset into groups of at least k members and then *aggregates* the group members. These aggregated values are published instead of the original ones. In conventional microaggregation methods, it is desired to produce a protected dataset similar to the original one, so close data records are grouped into the same cluster. Accordingly, the aggregation phase of the algorithms are designed to minimize the sum of within-group squared error (SSE), and therefore a simple arithmetic mean in each group is utilized within the aggregation phase to compute the centroids. However, this trivial approach does not consider the proximity of the published values to the original ones, so intruders are able to limit the range of the original values with respect to published data. In this paper, a proximity-aware microaggregation post-processing algorithm is proposed that revisits the aggregation step to remedy this deficiency. Additionally, it is possible to consider different levels of minimum required distances between original record values and their corresponding published ones. Empirical results confirm the superiority of the proposed method in achieving a better trade-off point between disclosure risk and information loss in comparison with similar microaggregation techniques.

Keywords. Microaggregation, Statistical Disclosure Control, Data Privacy, Masking.

1 Introduction

A significant amount of personal data is collected by organizations and agencies. These data are usually considered fruitful repositories for public benefit research, so they are required to be publicly available in microdata forms. However, regarding the recent developments in data processing technologies, these data publishing are obligated to be in accordance with privacy regulations.

There are some computational privacy models such as k-anonymity [1], l-diversity [2], and p-sensitivity [3]. In a k-anonymous dataset, all records are clustered into groups of at least k members, where k denotes the aggregation level enforced by the data publisher. In order to produce a k-anonymous dataset, all group members are aggregated, and the aggregated values are published.

The Statistical Disclosure Control (SDC) literature has introduced a number of anonymization methods such as noise addition [4], synthetic microdata generation [5], microaggregation [6], rank swapping [7] and hybrid methods [8]. The methods are categorized into perturbative and non-perturbative depending on the effect on the original data values [9, 10]. Additionally, they are divided into methods for continuous and categorical data based on the data type of the original values [11]. There is usually a tension between respondent privacy and data utility, i.e., more privacy results in less data quality and vice versa. The data publisher has to execute a protection algorithm (or a set of different anonymization algorithms) with different tuning parameters to capture a desired trade-off point between privacy and utility. More comprehensive information about the methods and measures can be found in [11, 12]. l Microaggregation is a perturbative mechanism that can realize k-anonymity [13], which is often utilized by statistical agencies [14]. Domingo-Ferrer and Torra have shown that microaggregation methods usually achieve promising results in terms of the conflicting measures of privacy and utility of the protected data [15].

Different microaggregation methods are compared based on Disclosure Risk (DR) and Information Loss (IL) of the protected datasets[1]. Linkage Disclosure (LD) and Interval Disclosure (ID) are two measures that quantify DR. The probability for intruders to successfully associate a protected record to its corresponding original one is quantified by LD [16]. The measure is always below $1/k$ in a k-anonymous protected dataset, so, the aggregation level k is an effective way to limit this type of disclosure. On the other hand, if intruders cannot find the exact record of a data owner, they may still be able to limit the range of some sensitive attribute values. That is called *proximity breach* in this paper. This type of disclosure is measured by ID in numerical datasets. Unfortunately, in conventional microaggregation methods, there is no effective way to decrease ID. Moreover, all data values for all respondents are assumed in the same level of privacy requirements (in terms of proximity breach).

The main contribution of this paper is to propose a Personalizable Proximity breach Prevention method through a disclosure-aware Microaggregation post-processing (P3M) algorithm to reduce ID. The distance between original and perturbed values is addressed in the proposed solution. This distance may be provided for each original attribute separately, for example, by data owners. Unfortunately, in almost all cases, it is not possible to satisfy all such *hard* requirements, so a relaxed version of the method is proposed that tries preserve these *soft* distances. The P3M usually achieves a better trade-off point between DR and IL in comparison with conventional microaggregation methods.

The remainder of this paper is organized as follows: Section 2 describes some concepts as required background. Section 3 reviews some previous microaggregation algorithms. Section 4 introduces the proposed aggregation method. Section 5 reports evaluation results. Finally, Section 6 summarizes and concludes the paper.

2 Basic concepts

In this section, we review some preliminary knowledge about the microaggregation problem (Section 2.1), statistical properties of a protected dataset by conventional microaggregation algorithms (Section 2.2), and evaluation measures (Section 2.3).

[1] The measures are introduced in Section 2.3 with more details.

2.1 The microaggregation problem

In this section, the microaggregation problem is formalized. Suppose a d-dimensional numerical dataset $V = \{v_i \in \mathbb{R}^d\}, i \in \{1, \ldots, n\}$ is given. In order to avoid the scaling problem, raw data records are scaled (normalized) and stored in $T = \{x_i\}$. The value of attribute t of the i-th record in T is computed by $x_i[t] = (v_i[t] - \mu_V[t])/\sigma_V[t]$, where $\mu_V[t]$ and $\sigma_V[t]$ are the mean and standard deviation of attribute t of V, respectively. Therefore, the mean becomes zero and standard deviation will be equal to one.

In the next steps, the algorithm partitions the dataset into m non-overlapping groups, i.e., $T = \cup_{j=1}^{m} G_j$ and $G_j \bigcap G_{j'} = \emptyset,\ \forall j, j' \in \{1, \ldots, m\}, j \neq j'$. The main criteria of a classic microaggregation algorithm during the aggregation phase is to produce a similar protected dataset to the original one for a given aggregation level k. A microaggregation method attempts to cluster similar records into groups (partitioning). The output of this phase can be shown as an assignment of records to groups, so $asn(i) = j \iff x_i \in G_j,\ \forall i \in \{1, \ldots, n\},\ \forall j \in \{1, \ldots, m\}$. For each group G_j with n_j members, the centroid C_j is computed (aggregation). In a classic aggregation algorithm, these centroids are computed to minimize the sum of within-group squared error (SSE). The measure is formulated in Equation (1).

$$SSE = \sum_{t=1}^{d} SSE_t = \sum_{t=1}^{d} \sum_{\substack{i=1 \\ j=asn(i)}}^{n} (x_i[t] - C_j[t])^2 . \tag{1}$$

The measure is minimized for the arithmetic mean of group members, i.e.,

$$C_j[t] = 1/n_j \sum_{x_i \in G_j} x_i[t],\ \forall j \in \{1, \ldots, m\}, \forall t \in \{1, \ldots, d\}.$$

The total sum of squares (TSS or SST) is also a quantity that can be rewritten in terms of n and d as shown in Equation (2)[2].

$$SST = \sum_{t=1}^{d} SST_t = \sum_{t=1}^{d} \sum_{i=1}^{n} (x_i[t] - \mu_T[t])^2 = \sum_{t=1}^{d} n = nd. \tag{2}$$

The normalized measure $L = SSE/SST$ is always between 0 and 1. Lower values of L indicate that the centroids are more similar to original records.

2.2 Some statistical results of a k-anonymous dataset produced by a conventional microaggregation method

In this section, we review some statistical properties of a k-anonymous dataset that is produced by a conventional aggregation algorithm. Let $C = \{C_j\}, j \in \{1, \ldots, m\}$ denote the set of computed centroids and $T^C = \{x_i^C\}, i \in \{1, \ldots, n\}$ shows an initial version of the protected dataset (before rescaling) where the original records are replaced by their associated cluster centroids, i.e., $x_i^C = C_j$, if $x_i \in G_j$. The application of arithmetic mean in the aggregation phase does not change attribute means (Equation (3)), but decreases variances as shown by Equation (4) [17].

[2] For simplicity, we define $Var(X) = \sigma_X^2 = 1/n \sum_{i=1}^{n} (x_i - \mu_X)^2$ where X is a set of n equally likely values x_i with $\mu_X = Mean(X)$.

$$\mu_{T^C} = \frac{1}{n}\sum_{j=1}^{m} n_j C_j = \frac{1}{n}\sum_{i=1}^{n} x_i = \mu_T = [0]^d. \tag{3}$$

$$\begin{aligned}
\sigma^2_{T^C}[t] &= \tfrac{1}{n}\sum_{j=1}^{m} n_j (C_j[t])^2 \qquad\qquad \forall t \in \{1,\ldots,d\} \\
&= \tfrac{1}{n}\sum_{i=1}^{n} (x_i[t])^2 - \tfrac{1}{n}\sum_{\substack{i=1 \\ j=asn(i)}}^{n} (x_i[t] - C_j[t])^2 \\
&= \sigma^2_T[t] - \tfrac{1}{n}\sum_{\substack{i=1 \\ j=asn(i)}}^{n} (x_i[t] - C_j[t])^2 \\
&= 1 - \tfrac{1}{n} SSE_t
\end{aligned} \tag{4}$$

The protected dataset T^C is first renormalized and saved in $T' = \{x'_i\}, i \in \{1,\ldots,n\}$ (Equation (5)) and then rescaled and shifted such that the original units are recovered (Equation (6)). The results are saved in $V' = \{v'_i\}$, with similar mean and variance of the original dataset V.

$$renormalization: \quad x'_i[t] = \left(\frac{x^C_i[t]}{\sigma_{T^C}[t]}\right), \forall i \in \{1,\ldots,n\}, \forall t \in \{1,\ldots,d\} \tag{5}$$

$$v'_i[t] = x'_i[t]\sigma_V[t] + \mu_V[t], \forall i \in \{1,\ldots,n\}, \forall t \in \{1,\ldots,d\} \tag{6}$$

2.3 Assessment of a masked numerical dataset

The assessment of a microdata protection method is based on the DR and IL measures[3]. Given an original dataset $V = \{v_i\}, i \in \{1,\ldots,n\}$, a microaggregation mechanism $\mathcal{F}$ computes its k-anonymous version $V' = \{v'_i\}, i \in \{1,\ldots,n\}$, where $v'_i = \mathcal{F}_k(v_i)$. Two general risk measures are computed on V' with respect to V and the average is reported as DR [18]:

A. Linkage Disclosure[4](LD): This is the standard mechanism to measure disclosure risk of a protection method [19]. Distance-based Linkage Disclosure (DLD) [20] is a kind of record linkage where original records are linked to their nearest neighbor in the masked dataset based on the Euclidean distances. In this paper, we consider a scenario in which an intruder tries to link an original record to its corresponding protected one. Assume T and T' denote the normalized versions of V and V', respectively. Each original record $x_i \in T$ is matched (linked) to its first closest protected record $x'_{i'} \in T'$. The measure counts correct links, i.e., $i = i'$. More formally, let $N^1_{T'}(x_i)$ denote the first nearest neighbor of x_i in T', i.e., $N^1_{T'}(x_i) = \min\{i' | i' = \text{argmin}_\kappa(\|x_i - x'_\kappa\|), x'_\kappa \in T'\}$ where $\|.\|$ denotes the Euclidean norm. The DLD measure is formalized in Equation (7).

$$DLD(T,T') = \frac{\#\{x_i | i = N^1_{T'}(x_i)\}}{n} \tag{7}$$

The exact value of DLD depends on the number of attributes that the intruder is assumed to know, but it is always between 0 and 1. Herranz et al. considered the most

[3] We review some general purpose DR and IL measures only for continuous data type, which is addressed in this paper. The variants of the measures for other data types can be found in [11].

[4] It is also known as identity disclosure or re-identification risk.

favorable case for the intruder: she knows all attributes of some original record and wants to link it to the corresponding protected record [21].

B. Interval Disclosure[5](ID): After microaggregation, an intruder cannot re-identify the exact record of a data subject since there are at least $k-1$ other data records with the same attribute values of the matched record. However, she may still be able to estimate the interval (range) of a sensitive numerical attribute. More precisely, if attribute t of the i-th masked record, $x'_i[t]$, falls close to its corresponding original value $x_i[t]$, the intruder would be able to estimate the original value with high accuracy. The proportion of original values that their corresponding values in the first nearest masked record fall into a predefined interval around them is reported as interval disclosure. The length of the interval may be defined in proportion to the original value (relative interval) [8], a percentage of the standard deviation [22], range of the corresponding attribute (absolute interval) [8], or the rank of the masked value among sorted attribute values (rank-based interval) [16]. For example, for a specified safety distance, $0 \leq sd \leq 1$, the original value of attribute t of record v_i, denoted by $v_i[t]$, suffers from $sd \times 100\%$ Standard Deviation-based Interval Disclosure ($SDID$) risk, if its corresponding masked value $v'_i[t]$ with the standard deviation $\sigma(att[t])$ is at most $sd \times \sigma(att[t])$ far from it. The measure is formulated in Equation (8).

$$SDID_{sd}(V,V') = \frac{\sum_{t=1}^{d} \#\{v_i | sd \geq |\frac{v_i[t]-v'_{i'}[t]}{\sigma(att[t])}|, i \in \{1,\ldots,n\}, i' = N^1_{T'}(x_i)\}}{n \times d} \tag{8}$$

Information loss quantifies the amount of the utility that is lost after microaggregation, where the original dataset is considered a baseline for comparison. There are several approaches to measure IL, such as information-theoretic measures [23, 24], statistical measures [15, 25], and probabilistic measures [11, 26]. For instance, Mateo-Sanz et al. introduced a probabilistic version of information loss called PIL [26]. We review only PIL in this section since the measure is addressed in multiple recent works such as [21, 27, 28] and [29]. Assume θ denotes a population parameter on T and Θ represents a sample statistic on T'. Let $\hat{\Theta}$ be the value of this statistic for a specific sample. The standardized sample discrepancy is shown in Equation (9).

$$Z = \frac{\hat{\Theta} - \theta}{\sqrt{Var(\hat{\Theta})}}. \tag{9}$$

For enough large number of records ($n > 100$), the discrepancy can be assumed to follow a $\mathcal{N}(0,1)$. Let us assume that $\hat{\theta}$ is the value taken by $\hat{\Theta}$ for a specific sample. The probabilistic information loss for $\hat{\Theta}$ is defined in Equation (10):

$$pil(\hat{\Theta}) = 2 \cdot P\left(0 \leq Z \leq \frac{\hat{\theta} - \theta}{\sqrt{Var(\hat{\Theta})}}\right). \tag{10}$$

In this paper, five measures are considered to quantify information loss, similar to [16, 26] and [30]. These measures are based on the following statistics ($t, t' \in \{1,\ldots,d\}$):

[5]Interval Disclosure is a special case of Attribute Disclosure for continuous datasets.

1. Quantiles for attribute A_t: The values that divide the distribution such that a given proportion of the observation are below the quantile. The quantiles from 5% to 95% with steps of 5% have been considered.
2. Mean for attribute A_t:
   $Mean(A_t) = \overline{x[t]} = \sum_{i=1}^{n} x_i[t]/n,$
3. Variance for attribute A_t:
   $Var(A_t) = \sum_{i=1}^{n} (x_i[t] - \overline{x[t]})^2/n,$
4. Covariance for attributes A_t and $A_{t'}$:
   $Cov(A_t, A_{t'}) = \frac{\sum_{i=1}^{n}(x_i[t]-\overline{x[t]})(x_i[t']-\overline{x[t']})}{n},$
5. Correlation coefficient for attributes A_t and $A_{t'}$:
   $\rho(A_t, A_{t'}) = \frac{\sum_{i=1}^{n}(x_i[t]-\overline{x[t]})(x_i[t']-\overline{x[t']})}{\sqrt{\sum_{i=1}^{n}(x_i[t]-\overline{x[t]})^2}\sqrt{\sum_{i=1}^{n}(x_i[t']-\overline{x[t']})^2}},$

The final information loss is computed based on Equation (10). The value is always between 0 and 1 and is formulated in Equation (11).

$$PIL(T, T') = (pil(Q) + pil(Mean) + pil(Var) + pil(Cov) + pil(\rho))/5. \quad (11)$$

In Equation (11), $pil(Q)$ is the average of the set of measures $pil(Q_l(A_t))$ for $l = 5\%$ to $l = 95\%$ with increment of 5%.

Generally, a decrease of DR comes with an increase in IL, so a multi-objective problem arises to find a suitable trade-off point [28]. A simple approach is to aggregate the measures based on a weighted sum method. For instance, Scoring Index SI introduced in Equation (12) may be used as an indicator to show how successful an algorithm is.

$$SI = (DR + IL)/2 \quad (12)$$

It is clear that lower values of SI are preferred.

3 Related works

Microaggregation methods are classified into fixed size and data-oriented ones [18]. A fixed size microaggregation method, clusters the dataset into groups of the same size k, but a data-oriented method partitions the dataset based on the distribution of records and results in variable size groups with at least k members. Generally, data-oriented methods are more complex, but produce more useful protected datasets [31].

Microaggregation methods are also categorized into univariate ($d = 1$) and multivariate $d > 1$. For the special case of $d = 1$, there is an optimal and polynomial time microaggregation algorithm, called MHM in this paper, that minimizes SSE[32], but the general problem is NP-hard for $d > 1$ [33]. Many heuristic algorithms are introduced in the literature. A well-known fixed size technique is Maximum Distance to Average Vector (MDAV) [34] which is the most widely-used microaggregation algorithm [35].

The pseudo-code of MDAV is shown in Algorithm 1. Initially, the method stores the normalized version of the input dataset in T (step 2) and then computes the centroid of T (step

Algorithm 1: THE PSEUDO-CODE OF MDAV

Input: V: original dataset, k: aggregation level
Output: V': protected datasets

1: Save the normalized version of V in $T = \{x_i\}$.
2: Compute the centroid μ_T of T.
3: Find the most distant record $x_r \in T$ from μ_T. Also find the most distant record $x_s \in T$ from x_r.
4: Form a cluster containing x_r and its $k-1$ nearest neighbors. Form another cluster containing x_s and its $k-1$ nearest neighbors. Put aside these clusters from the dataset.
5: If there are at least $2k$ records remaining, repeat steps 2,3, and 4.
6: If there are at least k and at most $2k-1$ records remaining, form a new cluster containing all of them
7: If there exist at most $k-1$ records in T, assign each of them to the closest clusters.
8: Aggregate group members and compute the centroids
9: Rescale the centroids to recover mean and variance of V and save them in V'
10: Return V'

2). In the next step, MDAV finds the most distant record, say x_r, from the centroid and farthest record from r, say x_s and builds two clusters including x_r and x_s and their $k-1$ nearest records in T. The records in these clusters are removed from T in step 4. Steps 2,3, and 4 are repeated until less than $2k$ records remain (step 5). All leftover records form a new cluster (step 6) or are assigned to their nearest clusters (step 7). Finally, all group members are aggregated in step 8 and the computed centroids are destandardized to recover the original units of V (step 9). Similarly, CBFS as a variant of MDAV, builds one group in each iteration [36]. The method can be implemented using the kd-tree data structure [37] to efficiently anonymize large numerical data volumes [38].

There are also some multivariate extensions of MHM, where records are ordered based on various heuristics such as the next point in a TSP tour, MicTSP [39, 40], or Nearest Point Next heuristic (NPN-MHM) [41]. For instance, MicTSP maps ordering of multivariate data records into the well-known Traveling Salesman Problem (TSP) [42], in which data records are considered as the locations of cities that are to be visited. The main objective of the TSP, i.e., finding the shortest possible route that visits each city exactly once and returns to the beginning city, indirectly preserves the locality of nearby records and simulates a multivariate sort procedure to provide the input for MHM. In a more simple approach, in NPN-MHM, starting by the farthest record from the centroid of the whole dataset, all records are ordered in a nearest point next fashion, and the MHM algorithm is applied to the sequence of records to provide an optimal partitioning with respect to the input sequence. The authors also suggested a way to sort centroids of some fixed-size microaggregation algorithms such as MDAV and CBFS, and then applying MHM on the sequence of records going through their associated groups. Mortazavi and Jalili proposed a Fast Data-oriented Microaggregation algorithm (FDM) that produces k-anonymous versions of a dataset for multiple successive values of k in a single run [40]. This method applies an extension of MHM to the sequence of data records in a TSP tour. Moreover, Mortazavi and Jalili proposed a preference-based microaggregation algorithm in which the preferences of the data publishers in terms of DR and IL could direct the anonymization process [28]. Recently, Mortazavi and Jalili introduced a disclosure-aware model of aggregation where protected records are in a given minimum distance from the original ones [29]. For a more

comprehensive survey of microaggregation methods, interested readers are recommended to refer to [43, 44]. Unfortunately, none of the methods mentioned above considers proximity breach during microaggregation at the fine granularity level of the attribute value. There is no effective way for data publishers to decrease such a risk in a systematic and integrated approach and there is no difference between different attribute values during anonymization, i.e., the data owners cannot state/suggest their preferences in terms of the minimum distance of published values to their original values.

There are also some extensions of k-anonymity in the Privacy Preserving Data Publishing (PPDP) methods that consider the proximity breach. Li et al. proposed (ϵ, m)-anonymity [45]. This model demands that in each group G and for every sensitive value x in G, at most $1/m$ of records in G are allowed to have sensitive values "similar" to x, where the similarity is controlled by ϵ. Wang et al. introduced $(\epsilon, \delta)^k$-dissimilarity [46] that requires each group G has at least k members and every sensitive value in G be "dissimilar" to at least $\delta \cdot (|G| - 1)$ other ones. Two sensitive values are considered dissimilar if their distance is more than ϵ.

4 The proposed microaggregation method

In a conventional microaggregation algorithm for numerical datasets, it is desired to have similar data records partitioned into the same partitions. Moreover, the centroids are computed in favor of reducing SSE. Therefore, the simple arithmetic mean of group members is considered as the centroid of the group. The aggregation phase followed by the naïve rescaling step, does not consider the *proximity* of computed protected values to their corresponding original ones. In other words, these steps do not attempt to reduce interval disclosure risk. Instead of choosing the centroids only in favor of minimizing SSE, the aggregation phase can be enhanced to consider interval disclosure and at the same time preserve the original statistical units. As an illustrative example, Figure 1 shows a dataset with 29 two-dimensional data records in (X, Y). After microaggregation by MDAV for $k = 5$, some data points suffer from the proximity of computed centroids (shown by arrows) [29]. This problem can be avoided by a disclosure-aware aggregation algorithm that considers the proximity of the computed centroids to their associated original values during aggregation. In our example, $DR = (DLD{+}IL)/2$, IL and $SI = (DR{+}IL)/2$ of MDAV can be changed from $(10.34 + 8.62)/2 = 9.48\%$, 58.34%, and $(9.48 + 58.34)/2 = 33.91\%$ to $(6.90 + 3.45)/2 = 5.17\%$, 58.95%, and $(5.17 + 58.95)/2 = 32.06\%$, respectively.

The aggregation phase of the Personalizable Proximity breach Prevention through a disclosure-aware Microaggregation post-processing (P3M) algorithm is described in the following. Partitioning phase of the P3M utilizes the same heuristic of MDAV [34] that is shown in Algorithm 1.[6]

In order to decrease ID, without a significant loss of information utility (in terms of SSE), we enforce the simply computed centroid C_j to be shifted by δ_j. However, to preserve the similarity of the protected dataset to the original one, it is desired to minimize the total amount of these shifts. The centroids are moved around the computed values to be far *enough* from their original values, based on the minimum allowed distances Δs, as personal preferences. This formulation can be applied in the aggregation phase of any conventional microaggregation method.

[6] It is notable that the P3M can be applied as a post-processing algorithm after any conventional microaggregation algorithm without changing its partitioning phase.

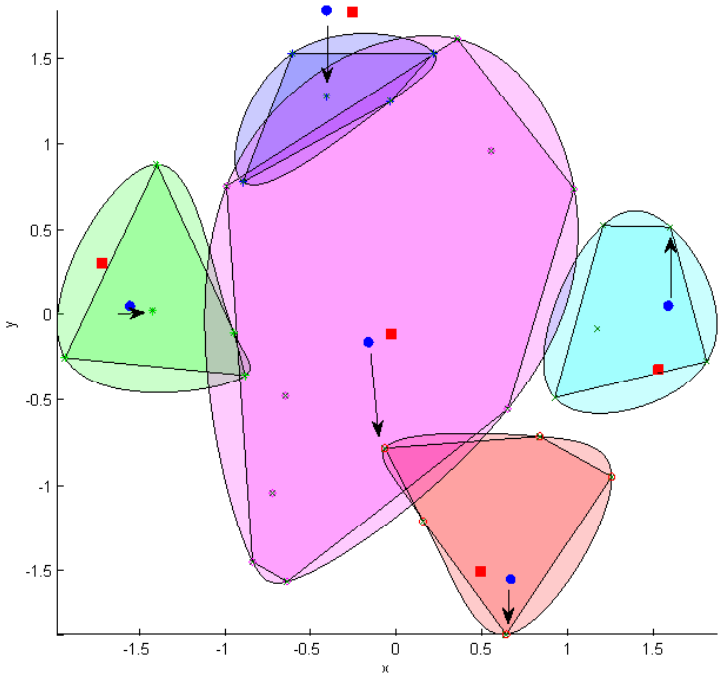


Figure 1: Simple and disclosure-aware centroids in a group of records (small markers) in $\mathbb{R}^2$ for $k = 5$ after simple (blue discs) and disclosure-aware (red squares) aggregation methods. Data points with at least one attribute close to the MDAV centroids are pointed by arrows. Arrowheads show attribute values of records that an intruder can estimate. [29]

In order to preserve the mean and variance of the original data and at the same time considering disclosure risk, we change the renormalization formula (Equation (5)) during the aggregation phase to compute the new centroids $C'_j[t]$ as shown in Equation (14) for attribute t, where obj_t is minimized with respect to a number of constraints to *simulate* renormalization step. In this equation, $\delta_j[t]$ is a variable that shows the amount of required shift of $C_j[t]$ to produce $C'_j[t] = C_j[t] + \delta_j[t]$. The value $\Delta_i[t]$ is the provided parameter by the i-th data owner that shows the minimum required distance between the original value $x_i[t]$ and its corresponding masked value $x'_i[t]$. It is desired for all of the computed centroids to be not so far from the original values, but not closer than the preferences vector Δ, so equality constraints are applied in our model formulated in Equation (14).

In this paper, we introduce the notion of Satisfaction Level (SL) as a *personalized assessment index* based on data owners' preferences. These preferences may be gathered from data owners along with the original values. Data owners are provided the chance to state their preferences in terms of the minimum required distances between their original record values and corresponding masked values in the protected dataset.[7] Consider the i-th data owner requires/suggests a minimum distance $\Delta_i[t] \geq 0$ between her original value $x_i[t]$ and its masked version $x'_i[t]$. For attribute t, denoted by $att[t]$, the value of $sl_i[t]$ is set to 1 if she is satisfied with a protection method, i.e., $\Delta_i[t] \leq |x_i[t] - x'_i[t]|$ or no such a preference is stated ($\Delta_i[t] = 0$). If the constraint is not met, $sl_i[t]$ is set to 0. All data records may not be at the same importance level. Therefore, the P3M utilizes a modified version of the constraints by introducing the notion of *importance*, shown by w. The value $w_i[t] \in [0, 1]$ is the importance of satisfying the distance constraint between $x_i[t]$ and $x'_i[t]$. The value is used as a parameter in the P3M that may be specified by data owners or data publisher. The average Satisfaction Level (SL) of a protected dataset is formulated in Equation (13). Obviously, larger values of SL are more desired, while SL is always between 0 and 1 (for completeness, we define $0/0 = 1$).

$$SL(T, T') = \frac{\sum_{i,t} w_i[t] sl_i[t]}{\sum_{i,t} w_i[t]}, i \in \{1, \ldots, n\}, t \in \{1, \ldots, d\}. \tag{13}$$

It is not always possible to satisfy all *hard* constraints of the required minimum distances, so they are relaxed by introducing $\gamma_i[t]$ in the P3M. This variable transforms the constraints to soft ones, which turns the proposed formulation more practical. For this goal, an *error* term $(1 - w_i[t])\gamma_i[t]$ is subtracted from the squared minimum distance requirement $\Delta_i[t]$. Therefore, $w_i[t] = 1$ means that the constraint cannot be violated, i.e., the constraint remains hard, but $w_i[t] = 0$ enforces less restrictions. It is notable that the parameter $w_i < 1$ does not necessarily mean preference constraints violations. It is used to put more pressure on the privacy of important respondents and making the model more practical. It is desired that our model minimizes the total sum of such errors, so they are added to the objective function for minimization. Additionally, the tuning parameter α controls the importance of reducing such errors in the first part of obj_t in comparison with the second part of obj_t that tries to minimize the total amount of shifts. The last two constraints are also added to preserve the original statistical units, mean and variance, respectively.

[7] For simplicity, we define this formulation for normalized values only. It is trivial to transform it in the same scale of the original dataset.

$$\begin{aligned}
\min \quad & obj_t = \alpha \sum_{i=1}^{n} \gamma_i^2[t] + (1-\alpha) \sum_{j=1}^{m} n_j \delta_j^2[t] && t \in \{1,\dots,d\} \\
\text{s.t.} \quad & \\
& (C_j[t] + \delta_j[t] - x_i[t])^2 = \Delta_i^2[t] - (1-w_i[t])\gamma_i[t] \quad \forall i \in \{1,\dots,n\}, j = asn(i) \\
& \sum_{j=1}^{m} n_j \delta_j[t] = 0 \\
& \sum_{j=1}^{m} n_j (C_j[t] + \delta_j[t])^2 = n
\end{aligned} \tag{14}$$

The second part of obj_t can be rewritten as a linear equation with respect to the last constraint of the optimization problem and Equation (4), so the solution can be calculated more efficiently. The computation steps are shown in Equations (15) and (16).

$$\begin{aligned}
\sum_{j=1}^{m} n_j (C'_j[t])^2 &= \sum_{j=1}^{m} n_j (C_j[t] + \delta_j[t])^2 \\
&= \sum_{j=1}^{m} n_j (C_j[t])^2 + 2 \sum_{j=1}^{m} n_j C_j[t] \delta_j[t] + \sum_{j=1}^{m} n_j \delta_j^2[t] \\
&= (n - SSE_t) + 2 \sum_{j=1}^{m} n_j C_j[t] \delta_j[t] + \sum_{j=1}^{m} n_j \delta_j^2[t] = n
\end{aligned} \tag{15}$$

So, we have,

$$\sum_{j=1}^{m} n_j \delta_j^2[t] = SSE_t - 2 \sum_{j=1}^{m} n_j C_j[t] \delta_j[t]. \tag{16}$$

As SSE_t is fixed, we can rewrite the objective function by Equation (17):

$$obj_t = \alpha \sum_{i=1}^{n} \gamma_i^2[t] - 2(1-\alpha) \sum_{j=1}^{m} n_j C_j[t] \delta_j[t]. \tag{17}$$

After computing $\delta_j[t]$, we can simply use the new centroids $C'_j[t] = C_j[t] + \delta_j[t]$ and compute the protected values as shown in Equation (18).

$$v'_i[t] = (C'_j[t])\sigma_V[t] + \mu_V[t] \quad \forall i \in \{1,\dots,n\}, \forall t \in \{1,\dots,d\} \tag{18}$$

5 Empirical evaluations

In this section, we report the evaluation results of our proposed method. There are three real-world benchmark datasets in the SDC, which are usually used to compare different microaggregation algorithms. More details about these datasets can be found in [47]. Additionally, we have evaluated the method on a synthetic random dataset with 10 clusters of normally distributed data points around random cluster centers. All these datasets contain numerical attributes that are introduced in Table 1. A PC with Core i7, 3.50 GHz CPU, Windows 7 64-bit and 16 GB of memory is used for experiments. We utilized CONOPT solver tool to find the optimal solution of our optimization problem. All initial partitions are generated by MDAV [34] microaggregation method. We used the average of Distance-based Linkage Disclosure (DLD) and Standard Deviation-based Interval Disclosure ($SDID$) to

Table 1: Benchmark datasets for microaggregation comparison [47]

Dataset	# data records (n)	# numeric attributes (d)
Tarragona	834	13
Census	1080	13
EIA	4092	11
Synthetic	10000	10

quantify DR. The DLD is computed on all attributes using the kd-tree data structure, similar to [21]. The interval lengths are set to 5% of the standard deviation of the underlying attribute during $SDID$ computation, i.e., $sd = 5$. Additionally, the Probabilistic Information Loss, PIL [26] is used to quantify IL. In all experiments, $\Delta_i[t] = 0.1$, $w_i[t] = \epsilon > 0$ (a small constant) and $\alpha = 0.5$ are applied, unless explicitly stated.

5.1 Personalized privacy in the P3M

In this section, the P3M is assessed to show how it can be used for different values of $\Delta_i[t]$ and w. These values may be provided by different data owners to suggest the required privacy. However, these values are applied in the algorithm as the data publisher defines. In this experiment, we have assumed that all data values in each cell of the underlying datasets require relative minimum distances, i.e., $\Delta_i[t] = 0.1x_i[t],\ \forall i \in \{1, \ldots, n\}, \forall t \in \{1, \ldots, d\}$. We have also repeated the experiment for $w_i[t] \in \{0.2, 0.4, 0.6, 0.8\}$ and $k \in \{3, 4, \ldots, 10\}$. The results are shown in Figures 2-5. Figure 2 reveals the fact that the only parameter of conventional microaggregation algorithms, i.e., the aggregation level k, is not always successful to prevent the proximity breach. Therefore, the final published values may be close to the original ones even for large values of k. However, the results confirm the role of w on improving SL. The results show that for different values of k, as w increases, satisfaction increases in general. For example, in Tarragona dataset for $k = 5$ and $w = 0.6$, we have $SL = 90.48\%$, while this value increases to $SL = 93.05\%$ for $w = 0.8$. The results also show that w is more effective for smaller values of k, except for EIA, as a clustered dataset, in which w plays a more important role for larger values of k.

5.2 The effect of the proposed method on improving SL

In this section, we compare the results of MDAV with the P3M in terms of the satisfaction level SL. Tables 2 reports SL measures for $k = \{3, 4, 5, 10\}$ for $w_i[t] = 0.001$. Additionally, Table 3 gives the same SL measures for $w_i[t] = rand$ where $rand$ is a number between 0 and 1 that is generated using MATLAB uniform random number generator seeded by 1. As expected, the best SL values are achieved by the P3M. For instance, based on Table 2, the P3M improves SL of MDAV for Tarragona in $k = 5$ from 37.00% to 54.55%, which means a 47.44% relative improvement. The results also show that the improvements are more significant for sparse datasets such as Tarragona than the clustered datasets such as EIA. In brief, even without any special tuning of the parameters in the P3M, the results show that there are opportunities to increase SL, significantly. There is no significant difference between the results of MDAV in Tables 2 and 3, while the latter confirms the superiority of the P3M in comparison with MDAV in all cases.

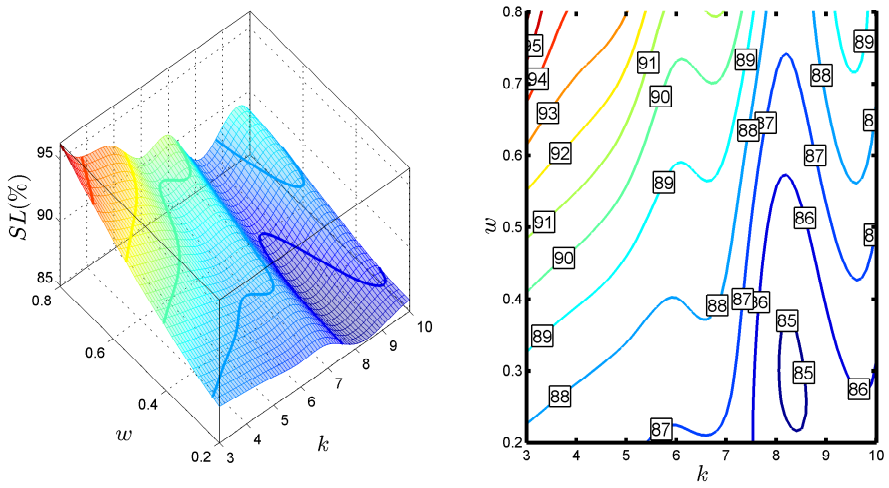


Figure 2: Satisfaction level of the P3M for different values of k and w for Tarragona dataset

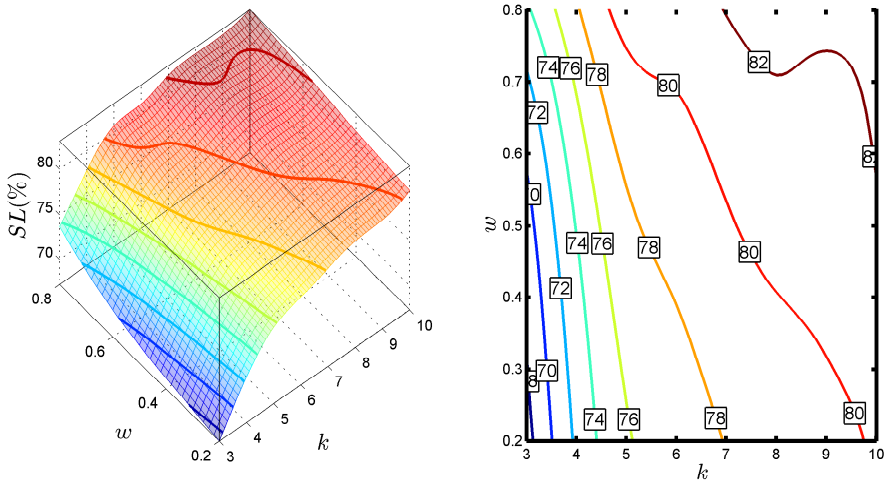


Figure 3: Satisfaction level of the P3M for different values of k and w for Census dataset

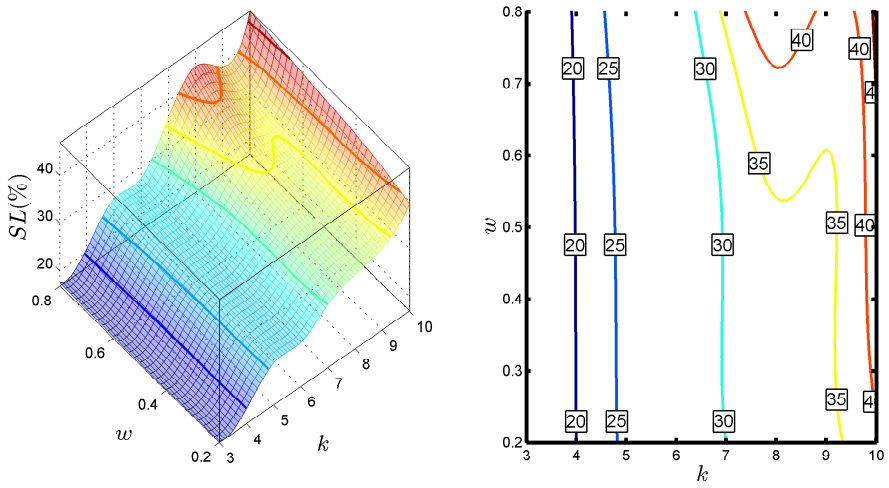


Figure 4: Satisfaction level of the P3M for different values of k and w for EIA dataset

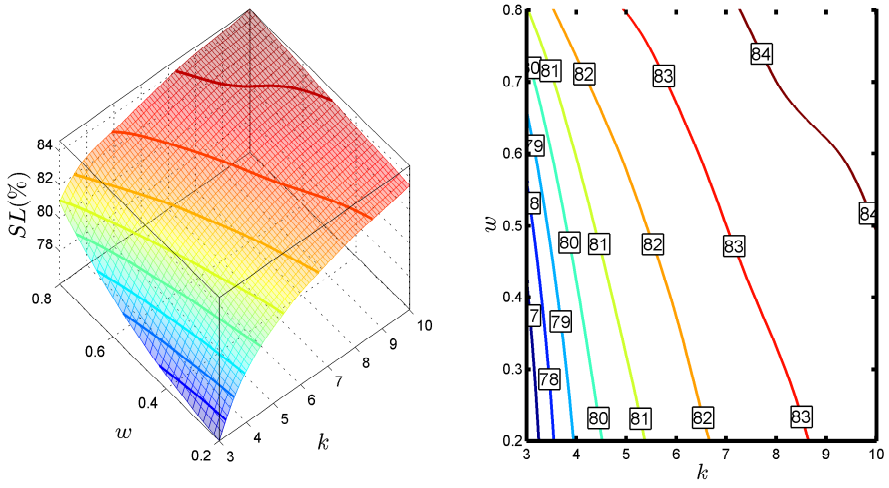


Figure 5: Satisfaction level of the P3M for different values of k and w for Synthetic dataset

Table 2: The effect of the P3M in improving SL in comparison with MDAV for $w_i[t] = 0.001$. Best values are shown in boldface.

k	Method	$SL(\%)$			
		Tarragona	Census	EIA	Synthetic
3	MDAV	29.33	53.67	5.58	65.84
	P3M	**47.92**	**55.19**	**5.67**	**67.65**
4	MDAV	33.78	59.10	7.35	70.24
	P3M	**51.38**	**61.10**	**7.52**	**71.96**
5	MDAV	37.00	62.44	10.64	72.46
	P3M	**54.55**	**65.18**	**10.89**	**74.00**
10	MDAV	47.20	69.34	17.15	76.98
	P3M	**57.64**	**73.10**	**17.81**	**78.24**

Table 3: The effect of the P3M in improving SL in comparison with MDAV for $w_i[t] = rand$. Best values are shown in boldface.

k	Method	SL (%)			
		Tarragona	Census	EIA	Synthetic
3	MDAV	29.37	53.74	5.60	65.85
	P3M	**66.16**	**60.49**	**6.81**	**71.23**
4	MDAV	33.36	59.22	7.42	70.27
	P3M	**67.29**	**67.06**	**9.37**	**75.34**
5	MDAV	36.70	62.41	10.78	72.41
	P3M	**69.42**	**70.24**	**13.50**	**76.80**
10	MDAV	47.09	69.60	17.22	76.74
	P3M	**76.26**	**76.29**	**27.93**	**80.43**

5.3 Improving the achieved trade-off points through parameter tuning

In this section, we repeat the previous experiments of Section 5.2 but for a range of different parameters of the P3M, i.e., $0 \leq \alpha \leq 1$ and $0 \leq \Delta \leq 1$ both with a step of 0.2. For each dataset and k, 16 experiments are conducted. Table 4 illustrates the best achieved SI measures along with their α and $\Delta_i[t] = \Delta$ values, and compares them with the results of MDAV.

Table 4: Comparison of the P3M and MDAV based on achieved SI

Dataset	k	$SI(\%)$		P3M Improvement (%)
		P3M (α, Δ)	MDAV	
Tarragona	3	16.69 (0.6, 0.8)	37.05	54.95
	4	17.20 (0.6, 0.8)	34.76	50.52
	5	16.44 (0.8, 0.8)	34.01	51.65
	10	21.34 (0.8, 0.8)	33.21	35.75
Census	3	24.37 (0.8, 0.6)	28.37	14.10
	4	24.00 (0.8, 0.6)	27.58	12.97
	5	25.29 (0.8, 0.6)	26.88	5.89
	10	25.23 (0.6, 0.8)	26.20	3.71
EIA	3	18.56 (0.6, 0.6)	34.41	46.05
	4	17.98 (0.6, 0.6)	33.17	45.80
	5	19.78 (0.6, 0.6)	34.44	42.57
	10	21.89 (0.8, 0.4)	35.46	38.26
Synthetic	3	27.41 (0.6, 0.8)	31.09	11.85
	4	28.12 (0.8, 0.8)	30.31	7.24
	5	28.34 (0.8, 0.6)	30.11	5.88
	10	28.66 (0.8, 0.6)	29.33	2.30

The results of all experiments indicate that usually larger values of α and Δ than 0.5 lead to an improved SI. The improvements are considerable in most cases. For instance, the proposed method has achieved $SI = 16.69\%$ for $k = 3$ in Tarragona, which means 54.95% relative improvement in comparison with MDAV. The improvements decrease as k increases since smaller values of k usually result in more compact groups, so the original values are close to the computed centroids in MDAV and large values of ID make DR increase, but in the P3M, the centroids are moved around so the risk decreases.

5.4 On the role of w on the achievements of the P3M

In this section, we report all of the micro indexes of the protected datasets by the P3M for different values of $w \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $k \in \{3, \ldots, 10\}$[8]. Figures 6-9 show the results for different datasets.

The figures show that as the importance parameter w increases, satisfaction level (SL) improves. This is rational since the optimization algorithm in the P3M puts more pressure on satisfying the minimum distances allowed, i.e., $\Delta_i[t]$. Meanwhile, SI improves only in Tarragona and EIA datasets for larger values of w. This is related to the distribution of the two datasets since Tarragona is known as a sparse dataset and EIA has clustered data points. Therefore, increasing w can effectively reduce ID in both of the datasets. For two other datasets with more uniform distributions, smaller values of w usually result in better values of SI.

For the first three datasets, the P3M yields an increased value for SI as k increases since more data records are aggregated and IL of the protected dataset increases. For Synthetic dataset, as k increases, DLD decreases significantly and produces better protected datasets in terms of SI.

5.5 Running time of the P3M

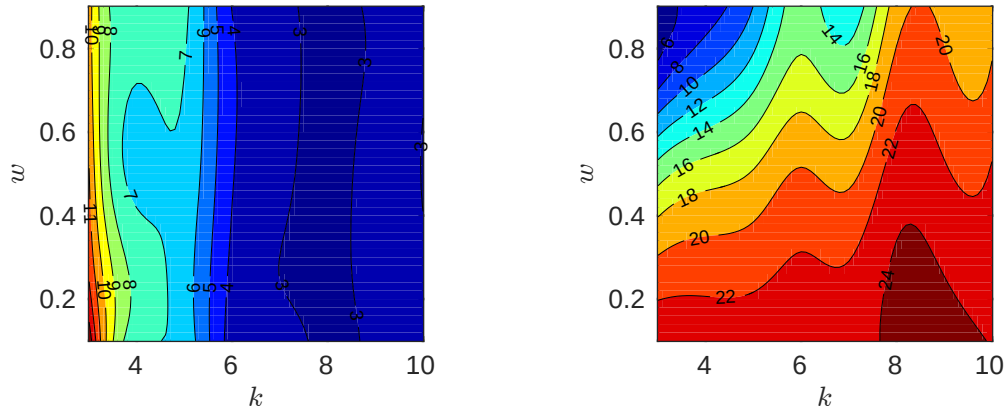
The improvements of the P3M are not at no cost. Table 5 shows the running time of the proposed method[9]. It is fair to say that the simple aggregation method in conventional microaggregation methods such as MDAV is faster than the method applied in the P3M. However, in microaggregation, the performance in terms of the running time seems less important than the achieved privacy and satisfaction level of data owners since the whole process of anonymization is an offline task.

5.6 Comparison of the P3M and a personalized microaggregation algorithm

In this section, we compare the proposed method with a recent personalized microaggregation algorithm, DREAM [29]. The DREAM uses the similar idea of the P3M, but focuses only on record level preferences of data owners to improve SI. More precisely, the DREAM tries to put anonymized values far from the original ones as much as possible, i.e., maximize $|x_i[t] - C'_j[t]|$. While both methods try to compute the centroids more intelligently, the P3M is more fine granular, since it focuses on attribute level privacy requirements. This results in more flexibility for data publisher to satisfy enforced requirements (for instance stated by data owners) and more tunable handles (parameters) to achieve a better trade-off between privacy and utility in terms of SI (Equation (12)). Table 6 shows the best achieved SI of the P3M and DREAM on Tarragona, Census, and EIA for $k = \{3, 4, 5, 10\}$. These results show that the P3M is more successful in scattered or clustered datasets such as Tarragona and EIA than uniform ones such as Census. It seems that the P3M can compute new centroids more effectively in non-uniform datasets to be far enough from distinct attribute values in compare with DREAM.

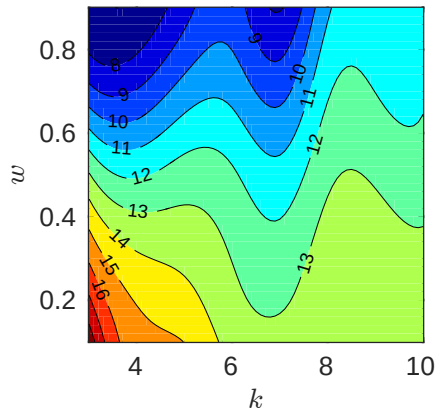
[8] In these experiments, the same value of $\Delta = 0.1$ is used for all values in contrast with the experiments in Section 5.1.

[9] In all cases, the execution time of the simple aggregation method using the arithmetic mean is less than 1 second, so it is not shown in the table

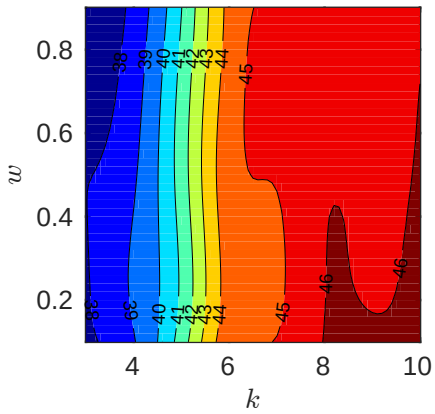


(a) DLD

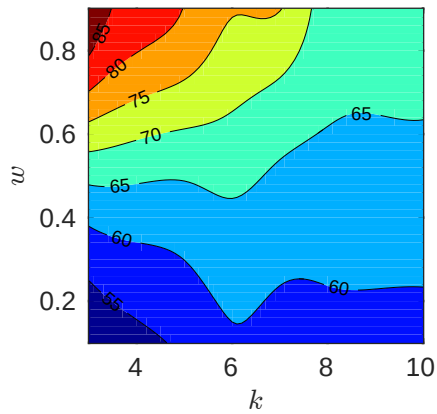
(b) ID



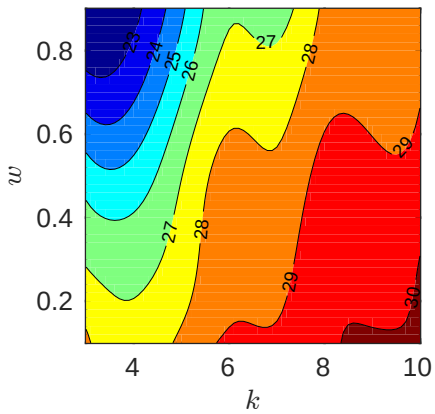
(c) DR



(d) IL

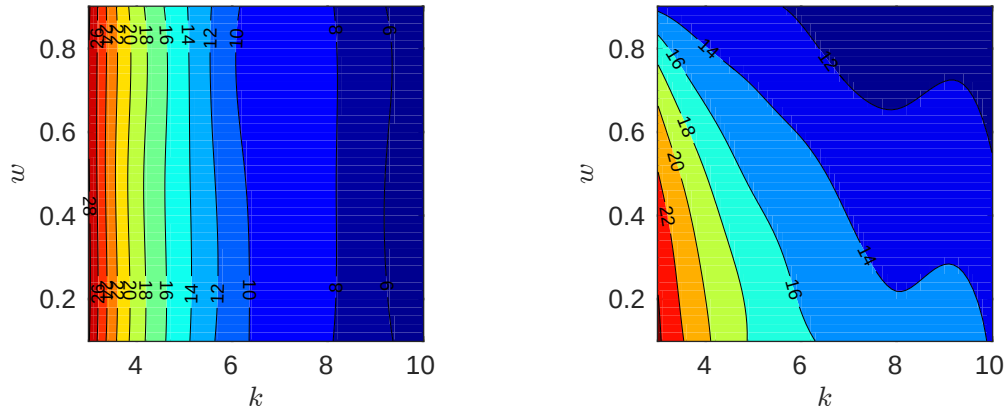


(e) SL



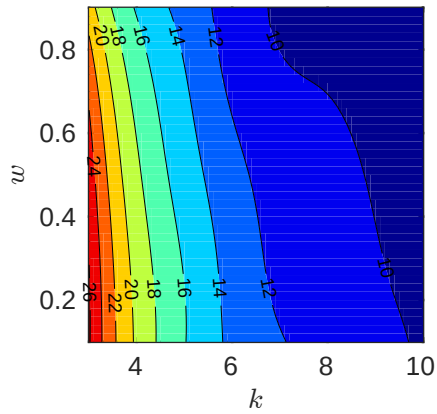
(f) SI

Figure 6: Evaluation of the P3M based on DLD, ID, DR, IL, SL, and SI for different k and w on Tarragona dataset

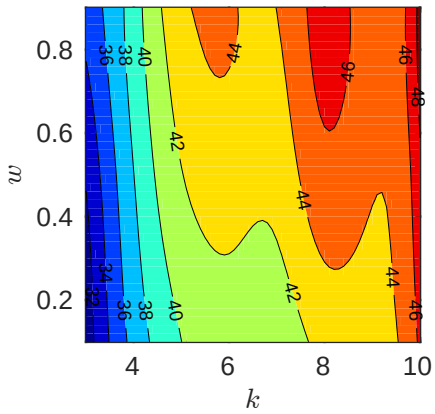


(a) DLD

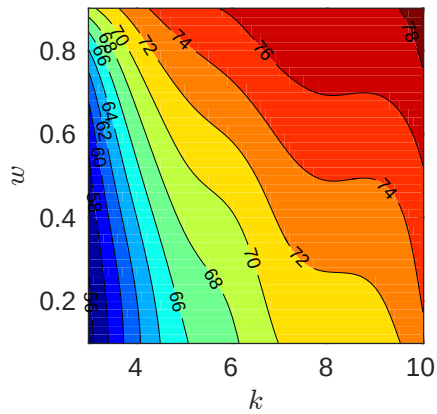
(b) ID



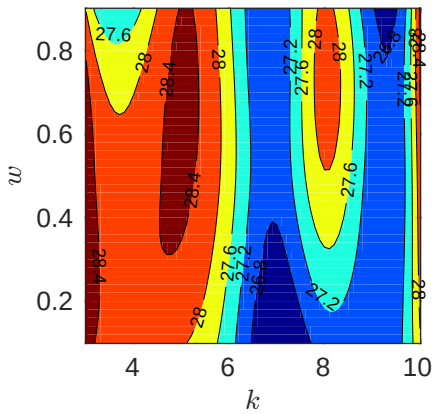
(c) DR



(d) IL

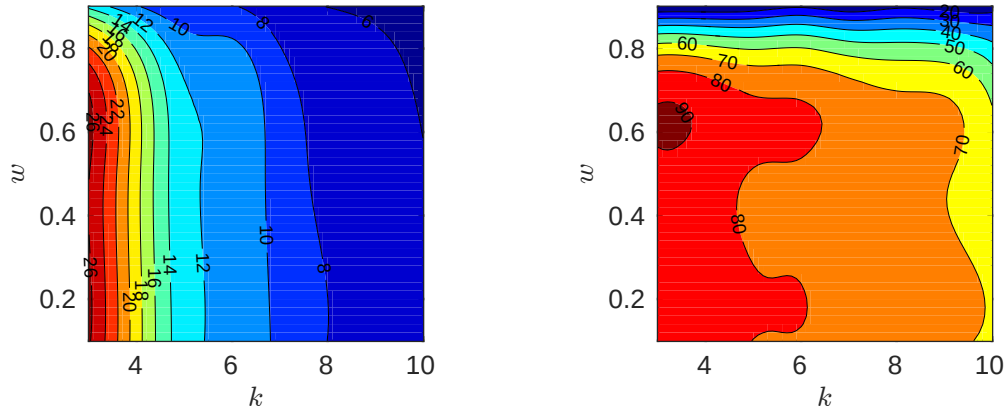


(e) SL



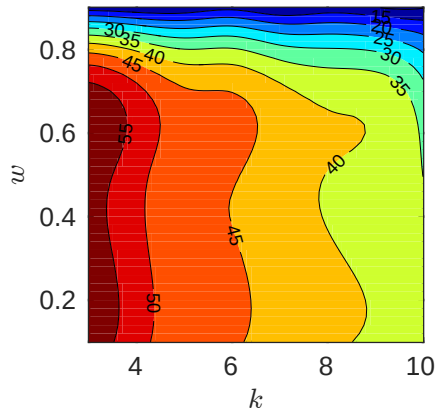
(f) SI

Figure 7: Evaluation of the P3M based on DLD, ID, DR, IL, SL, and SI for different k and w on Census dataset TRANSACTIONS ON DATA PRIVACY 10 (2017)

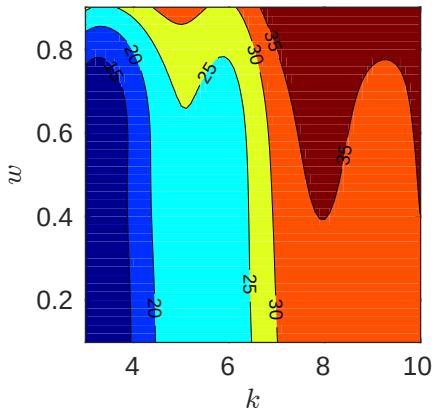


(a) DLD

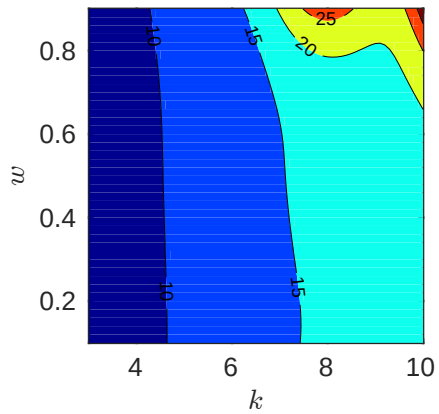
(b) ID



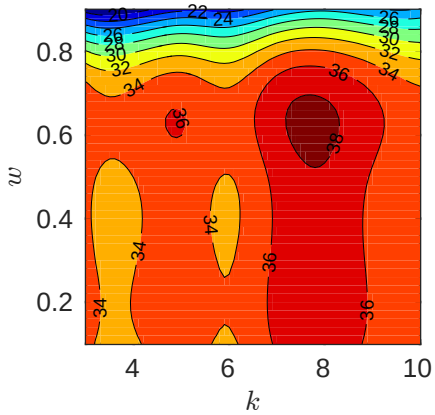
(c) DR



(d) IL

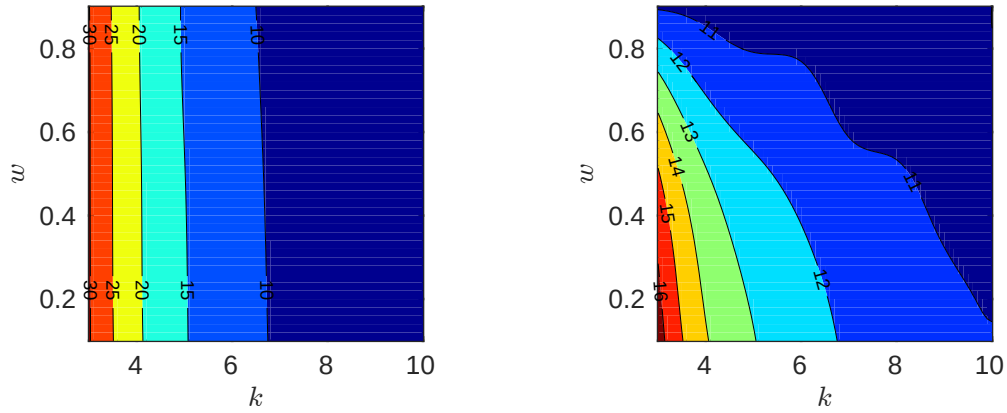


(e) SL



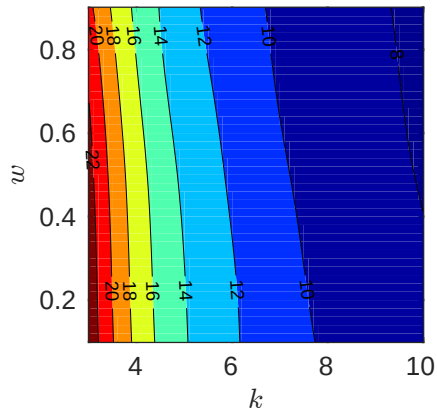
(f) SI

Figure 8: Evaluation of the P3M based on DLD, ID, DR, IL, SL, and SI for different k and w on EIA dataset

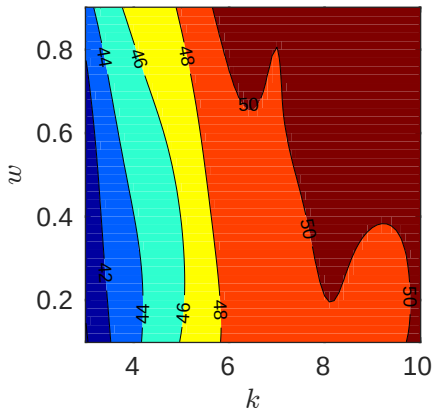


(a) DLD

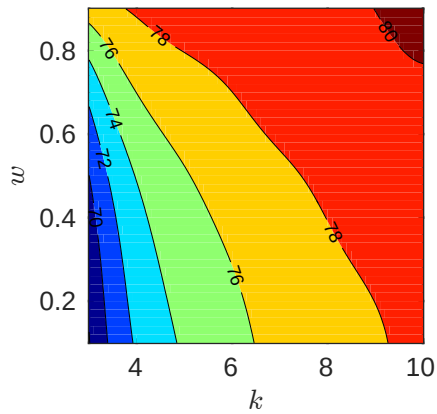
(b) ID



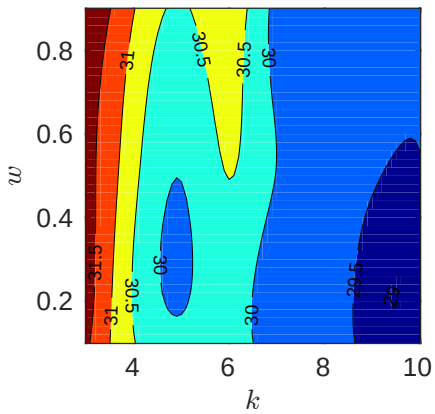
(c) DR



(d) IL



(e) SL



(f) SI

Figure 9: Evaluation of the P3M based on DLD, ID, DR, IL, SL, and SI for different k and w on Synthetic dataset

Table 5: Running time of disclosure-aware aggregation phase of the P3M

dataset	k	running time (sec)
Tarragona	3	< 1
	4	< 1
	5	< 1
	10	< 1
Census	1	< 1
	4	< 1
	5	< 1
	10	< 1
EIA	3	1
	4	1
	5	1
	10	1
Synthetic	3	3
	4	2
	5	2
	10	2

Table 6: Comparison of the P3M and DREAM based on achieved SI

Dataset	k	$SI(\%)$	
		P3M (α, Δ)	DREAM
Tarragona	3	**16.69** (0.6, 0.8)	28.67
	4	**17.20** (0.6, 0.8)	34.76
	5	**16.44** (0.8, 0.8)	34.01
	10	**21.34** (0.8, 0.8)	33.21
Census	3	24.37 (0.8, 0.6)	**22.86**
	4	24.00 (0.8, 0.6)	**21.82**
	5	25.29 (0.8, 0.6)	**22.25**
	10	25.23 (0.6, 0.8)	**21.91**
EIA	3	**18.56** (0.6, 0.6)	31.48
	4	**17.98** (0.6, 0.6)	31.70
	5	**19.78** (0.6, 0.6)	29.93
	10	**21.89** (0.8, 0.4)	30.32

It is worth mentioning the P3M is more efficient than DREAM. For instance, the P3M running time for EIA and $k = \{3, 4, 5, 10\}$ is about half of the running time of DREAM. This improvement is more noticeable in large datasets (see Table 5 of [29]). This is the result of simple yet effective constraints used in the P3M.

6 Summary and Conclusion

In conventional microaggregation methods in the SDC, cluster centroids are computed to minimize SSE. Although this is a successful measure to assess the similarity of protected datasets to the original ones, the measure does not reflect anything about the proximity of computed perturbed values to their original equivalents, so these methods may produce a useful protected dataset while most attribute values suffer from interval disclosure risk. Unfortunately, conventional methods only consider univariate sensitive values where proximity is defined among grouping data values, but final published values (centroids) are not involved directly. Additionally, all of these values are protected by the same level of privacy without considering the preferences of data owners. In contrast, we incorporate these requirements within our algorithm.

The P3M is designed as an integrated part of such methods to enable the data publishers to decrease ID. The results presented in Section 5 prove the superiority of the method in comparison with classic methods. The improvement of SI as a trade-off measure becomes more significant for smaller values of k. For example, the proposed method reduces SI of $MDAV$ in EIA dataset for $k = 5$ from 34.44% to 19.78% (Table 4), which means a 42.57% relative improvement, but this enhancement increases to 46.05% for $k = 3$. In brief, our results in Section 5 show that microaggregation techniques may achieve a better trade-off in terms of SI while the main privacy parameter of the underlying privacy model i.e., k is fixed. This would be an interesting attainment in practice.

Additionally, evaluation of the P3M for different values of w and k in Section 5.4 showed that the P3M can effectively control SL by tuning w. This means that the proposed method is completely personalizable for data owners since they can set the pressure of the optimization algorithm on satisfying their preferences during microaggregation.

The experiments also suggest that the best values for SI, as a trade-off measure, are attained for larger values of w in sparse or clustered datasets (especially for small values of k), but for more uniform datasets, small values for w with larger values of k usually produce more interesting datasets in terms of SI.

It is notable that if all respondents choose large values for minimum allowed distances, the solution of the proposed method does not yield an acceptable Satisfaction Level SL. Depending on the statistical agency policy, this may prohibit data publication at all, which is more acceptable than a data breach from the view point of data respondents. On the other side, if a data respondent does not care about the published value of a special attribute which is assumed to be sensitive in general, the statistical agency can use this information to produce more useful data.

In this paper, we have proposed a disclosure-aware aggregation model where the data publisher can effectively control the minimum distance between original and protected values based on data owners' preferences. This enables conventional microaggregation algorithms to achieve a decreased interval disclosure risk. The relative improvement of SI reaches up to 50% in some cases. In all experiments, a large group of data owners is satisfied with respect to their stated preferences. We are going to improve the formulation of the aggregation problem to find more efficient solutions. This is a critical task in anonymizing

large scale numerical datasets. It makes the solution less dependent on commercial tools such as CONOPT and helps to make it publicly available. Extending the proposed model for non-numeric datasets based on recent ideas of TBM [48] is an unexplored area. Comparing the running time of the P3M and the DREAM confirms the efficiency of the proposed method. However, we have not considered the case of very large and huge datasets. More research in this direction may be interesting.

References

[1] L. Sweeney, k-anonymity: A model for protecting privacy, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10 (5) (2002) 557–570.

[2] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, L-diversity: Privacy beyond k-anonymity, ACM Transaction on Knowledge Discovery from Data (TKDD) 1 (1) (2007) 1–52.

[3] T. M. Truta, B. Vinay, Privacy protection: p-sensitive k-anonymity property, in: Proceedings 22nd International Conference on Data Engineering Workshops, IEEE, 2006, pp. 94–94.

[4] R. Brand, Microdata protection through noise addition, in: J. Domingo-Ferrer (Ed.), Inference Control in Statistical Databases, Vol. 2316 of Lecture Notes in Computer Science, Springer, 2002, pp. 97–116.

[5] J. Burridge, Information preserving statistical obfuscation, Statistics and Computing 13 (4) (2003) 321–327.

[6] D. Defays, P. Nanopoulos, Panels of enterprises and confidentiality: the small aggregates method, in: Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys, 1993, pp. 195–204.

[7] R. A. Moore Jr, Controlled data-swapping techniques for masking public use microdata sets, Tech. Rep. 96-04, Statistical Research Division Report Series, US Bureau of the Census, Washington D.C. (1996).

[8] J. Herranz, J. Nin, M. Solé, More hybrid and secure protection of statistical data sets, IEEE Transactions on Dependable and Secure Computing 9 (5) (2012) 727–740.

[9] L. C. Willenborg, T. De Waal, Elements of statistical disclosure control, Vol. 155, Springer Verlag, 2001.

[10] J. Domingo-Ferrer, V. Torra, Disclosure protection methods and information loss for microdata, Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies (2001) 91–110.

[11] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, P.-P. De Wolf, Statistical disclosure control, Wiley, 2012.

[12] G. Navarro-Arribas, V. Torra, Information fusion in data privacy: A survey, Information Fusion 13 (4) (2012) 235 – 244.

[13] J. Domingo-Ferrer, A. Solanas, A. Martinez-Balleste, Privacy in statistical databases: k-anonymity through microaggregation, in: Proceedings of International Conference on Granular Computing, IEEE, 2006, pp. 774–777.

[14] W. E. Winkler, Re-identification methods for masked microdata, in: Privacy in statistical databases, Springer, 2004, pp. 216–230.

[15] J. Domingo-Ferrer, J. M. Mateo-Sanz, V. Torra, Comparing SDC methods for microdata on the basis of information loss and disclosure risk, in: Pre-proceedings of ETK-NTTS, Vol. 2, 2001, pp. 807–826.

[16] J. Domingo-Ferrer, V. Torra, A quantitative comparison of disclosure control methods for microdata, Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies (2001) 111–134.

[17] A. Oganian, A. F. Karr, Combinations of SDC methods for microdata protection, in: Privacy in Statistical Databases, Springer, 2006, pp. 102–113.

[18] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, P.-P. De Wolf, Cenex SDC handbook on statistical disclosure control, version 1.01 (2006).

[19] J. Nin, J. Herranz, V. Torra, On the disclosure risk of multivariate microaggregation, Data & Knowledge Engineering 67 (3) (2008) 399 – 412.

[20] D. Pagliuca, G. Seri, Some results of individual ranking method on the system of enterprise accounts annual survey, Report, Esprit SDC Project, Deliverable MI-3 D (1999).

[21] J. Herranz, J. Nin, M. Solé, Kd-trees and the real disclosure risks of large statistical databases, Information Fusion 13 (4) (2012) 260 – 273.

[22] J. Mateo-Sanz, F. Sebé, J. Domingo-Ferrer, Outlier protection in continuous microdata masking, in: J. Domingo-Ferrer, V. Torra (Eds.), Privacy in Statistical Databases, Vol. 3050 of Lecture Notes in Computer Science, Springer, 2004, pp. 201–215.

[23] J. Domingo-Ferrer, D. Rebollo-Monedero, Measuring risk and utility of anonymized data using information theory, in: Proceedings of the EDBT/ICDT Workshops, EDBT/ICDT, ACM, New York, NY, USA, 2009, pp. 126–130.

[24] M. Askari, R. Safavi-Naini, K. Barker, An information theoretic privacy and utility measure for data sanitization mechanisms, in: Proceedings of the second ACM conference on Data and Application Security and Privacy, CODASPY, ACM, New York, NY, USA, 2012, pp. 283–294.

[25] W. Yancey, W. Winkler, R. Creecy, Disclosure risk assessment in perturbative microdata protection, in: J. Domingo-Ferrer (Ed.), Inference Control in Statistical Databases, Vol. 2316 of Lecture Notes in Computer Science, Springer, 2002, pp. 135–152.

[26] J. Mateo-Sanz, J. Domingo-Ferrer, F. Sebé, Probabilistic information loss measures in confidentiality protection of continuous microdata, Data Mining and Knowledge Discovery 11 (2) (2005) 181–193.

[27] J. Herranz, S. Matwin, J. Nin, V. Torra, Classifying data from protected statistical datasets, Computers & Security 29 (8) (2010) 875 – 890.

[28] R. Mortazavi, S. Jalili, Preference-based anonymization of numerical datasets by multi-objective microaggregation, Information Fusion 25 (2015) 85–104.

[29] R. Mortazavi, S. Jalili, Enhancing aggregation phase of microaggregation methods for interval disclosure risk minimization, Data Mining and Knowledge Discovery 30 (3) (2016) 605–639.

[30] S. Ladra, V. Torra, On the comparison of generic information loss measures and cluster-specific ones, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 16 (2008) 107–120.

[31] Y. Li, S. Zhu, L. Wang, S. Jajodia, A privacy-enhanced microaggregation method, in: T. Eiter, K.-D. Schewe (Eds.), Foundations of Information and Knowledge Systems, Vol. 2284 of Lecture Notes in Computer Science, Springer, 2002, pp. 148–159.

[32] S. Hansen, S. Mukherjee, A polynomial algorithm for optimal univariate microaggregation, IEEE Transactions on Knowledge and Data Engineering 15 (4) (2003) 1043–1044.

[33] A. Oganian, 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) (2001) 345–354.

[34] J. Domingo-Ferrer, V. Torra, Ordinal, continuous and heterogeneous k-anonymity through microaggregation, Data Mining and Knowledge Discovery 11 (2) (2005) 195–212.

[35] A. Solanas, Privacy protection with genetic algorithms, in: A. Yang, Y. Shan, L. Bui (Eds.), Success in Evolutionary Computation, Vol. 92 of Studies in Computational Intelligence, Springer, 2008, pp. 215–237.

[36] M. Laszlo, S. Mukherjee, Minimum spanning tree partitioning algorithm for microaggregation, IEEE Transactions on Knowledge and Data Engineering 17 (7) (2005) 902–911.

[37] J. L. Bentley, Multidimensional binary search trees used for associative searching, Communications of the ACM 18 (9) (1975) 509–517.

[38] M. Solé, V. Muntés-Mulero, J. Nin, Efficient microaggregation techniques for large numerical data volumes, International Journal of Information Security 11 (4) (2012) 253–267.

[39] B. Heaton, New record ordering heuristics for multivariate microaggregation, Book, Nova Southeastern University (2012).

[40] R. Mortazavi, S. Jalili, Fast data-oriented microaggregation algorithm for large numerical datasets, Knowledge-Based Systems 67 (2014) 195 – 205.

[41] J. Domingo-Ferrer, A. Martínez-Ballesté, J. M. Mateo-Sanz, F. Sebé, Efficient multivariate data-oriented microaggregation, The VLDB Journal 15 (4) (2006) 355–369.

[42] K. L. Hoffman, M. Padberg, G. Rinaldi, Traveling salesman problem, in: Encyclopedia of operations research and management science, Springer, 2013, pp. 1573–1578.

[43] E. Fayyoumi, B. J. Oommen, A survey on statistical disclosure control and microaggregation techniques for secure statistical databases, Software: Practice and Experience 40 (12) (2010) 1161–1188.

[44] J.-L. Lin, P.-C. Chang, J. Y.-C. Liu, T.-H. Wen, Comparison of microaggregation approaches on anonymized data quality, Expert Systems with Applications 37 (12) (2010) 8161–8165.

[45] J. Li, Y. Tao, X. Xiao, Preservation of proximity privacy in publishing numerical sensitive data, in: Proceedings of the ACM SIGMOD international conference on Management of data, ACM, 2008, pp. 473–486.

[46] T. Wang, S. Meng, B. Bamba, L. Liu, C. Pu, A general proximity privacy principle, in: IEEE 25th International Conference on Data Engineering (ICDE), IEEE, 2009, pp. 1279–1282.

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

[48] M. Salari, S. Jalili, R. Mortazavi, Tbm, a transformation based method for microaggregation of large volume mixed data, Data Mining and Knowledge Discovery 31 (1) (2017) 65–91.