Revisiting Distance-Based Record Linkage for Privacy-Preserving Release of Statistical Datasets

Javier Herranz[a], Jordi Nin[b], Pablo Rodríguez[b], Tamir Tassa[c]

[a] *Dept. de Matemàtica Aplicada IV, Universitat Politècnica de Catalunya, Barcelona, Spain*
email: jherranz@ma4.upc.edu
[b] *Dept. d'Arquitectura de Computadors, Universitat Politècnica de Catalunya, Barcelona, Spain*
Phone:+34 93 401 6995, Fax: +34 93 401 7055
email: {nin,pablor}@ac.upc.edu
[c] *Department of Mathematics and Computer Science, The Open University, Ra'anana, Israel*
email: tamir_tassa@yahoo.com

Abstract

Statistical Disclosure Control (SDC, for short) studies the problem of privacy-preserving data publishing in cases where the data is expected to be used for statistical analysis. An original dataset T containing sensitive information is transformed into a sanitized version T' which is released to the public. Both utility and privacy aspects are very important in this setting. For utility, T' must allow data miners or statisticians to obtain similar results to those which would have been obtained from the original dataset T. For privacy, T' must significantly reduce the ability of an adversary to infer sensitive information on the data subjects in T.

One of the main a-posteriori measures that the SDC community has considered up to now when analyzing the privacy offered by a given protection method is the Distance-Based Record Linkage (DBRL) risk measure. In this work, we argue that the classical DBRL risk measure is insufficient. For this reason, we introduce the novel *Global Distance-Based Record Linkage* (GDBRL) risk measure. We claim that this new measure must be evaluated alongside the classical DBRL measure in order to better assess the risk in publishing T' instead of T. After that, we describe how this new measure can be computed by the data owner and discuss the scalability of those computations.

We conclude by extensive experimentation where we compare the risk assessments offered by our novel measure as well as by the classical one, using well-known SDC protection methods. Those experiments validate our hypothesis that the GDBRL risk measure issues, in many cases, higher risk assessments than the classical DBRL measure. In other words, relying solely on the classical DBRL measure for risk assessment might be misleading, as the true risk may be in fact higher. Hence, we strongly recommend that the SDC community considers the new GDBRL risk measure as an additional measure when analyzing the privacy offered by SDC protection algorithms.

1. Introduction

Nowadays, a huge amount of (digital) data is collected, processed, stored and eventually released to the public, in order to be used for different purposes. Sometimes, parts of the

data contain sensitive information on individual users, and a careless dissemination of it could be inconsistent with current privacy laws. Therefore, data owners are required to protect the collected information before granting third parties access to it.

Studies in various areas of computer science have been dedicated to this problem in the last years. For instance, the *data mining* area mainly considers the interactive setting: in that setting, the data owner collects and stores the data T; external entities may then submit queries $f(\cdot)$ on the data, to which the data owner replies with an approximate answer $y \approx f(T)$ that, on one hand, should be sufficiently close to the true answer $f(T)$, and, on the other hand, should not leak any significant information on the sensitive values contained in T. This problem led to the appearance of *privacy-preserving data mining* [3, 37].

A different scenario is that of *data publishing* [17], where the data owner collects and stores some original (and sensitive) data T and then releases a modified or perturbed version T' of it. The release is made independently of the queries that could be submitted later on by external entities. The particular case in which the expected information on T that is of interest to external users is of statistical nature (means, averages, variances, correlations, etc.) has received a lot of attention and has led to the appearance of an independent area: *Statistical Disclosure Control* (SDC, for short) [13, 58]. The goal of SDC is to design and analyze different methods to protect a dataset T in such a way that: (1) the released version T' allows external entities to obtain relatively accurate statistical information on T, and (2) the released dataset T' does not introduce privacy threats for the confidential information contained in T.

The SDC community proposed in the last decade many protection methods like noise addition, rank swapping, microaggregation and others. It also considered different ways to analyze both the utility level and the privacy level offered by such methods. Regarding utility, the current measures of (probabilistic) information loss all follow the same approach of measuring the difference between computing some statistical functions on the original data T and on the released data T'. Those measures are well accepted as they capture the utility of the published data for the purposes of statistical analysis. In this paper we focus on the second aspect by which SDC methods are evaluated, i.e. the privacy which they offer.

Defining good privacy measures for SDC methods is not as simple as defining utility measures, due to their dependence on the adversarial model. The first thing that needs to be done is to define both the *goals* and the *resources* of the attacker who is trying to break the privacy barrier. Regarding the goals, the SDC community has considered two possibilities [13]: first, an *interval disclosure* attacker may try to find a good approximation for some sensitive value in T; second, a *link disclosure* attacker may try to link a perturbed record of T' with an original (non-perturbed) record obtained from another source. An interval disclosure attacker is not assumed to have additional resources (other than T'). However, when considering link disclosure attackers, one has to define what are the external resources available to them. As it happens in cryptography, the most recommendable option (in order to ensure privacy even in the worst case) is to assume that the attacker has obtained some information on *all* original records in T, and then he uses this information in order to infer links between protected records in T' and original records in T. (See for example [20, 35, 50, 51, 56, 61, 62].) The goal of the attacker is then to infer correct links between the records in T (which typically include non-sensitive information, called quasi-identifiers, such as age, location, profession etc.) and the records in T' (that may include additional sensitive information such as medical or financial data) in order to reveal sensitive information on the data subjects. A link disclosure measure can be defined as the percentage of correct links that the attacker may infer between original and protected records.

But then a new problem emerges: what is the best strategy for an attacker to find correct links? Different attackers could use very different strategies, and it is impossible to take

all such strategies into account when defining a link disclosure measure. One of the linkage strategies for the attacker which has been widely adopted by the SDC community up to now is the so-called *distance-based record linkage*. A distance-based record linkage strategy finds, for every protected record $\mathbf{v}'_j \in T'$, an original record $\mathbf{v}_i \in T$ which minimizes the distance to $\mathbf{v}'_j$, for some pre-specified distance function (for instance, the Euclidean distance if all attributes in T are numerical). The pair $(\mathbf{v}_i, \mathbf{v}'_j)$ is then added to the list of links, and then the number of correct links, divided by $|T|$, gives a distance-based record linkage (DBRL for short) measure of disclosure risk. This is the only distance-based link disclosure measure that is typically considered in the SDC area when analyzing and comparing SDC protection methods.

1.1. Our contributions

The main contribution of this work is in observing that the classical DBRL disclosure measure is insufficient in order to analyze the privacy offered by SDC protection methods. The reason is that, when computing the DBRL measure, different protected records in T' may be linked to the same original record in T. For example, it is possible that some record $\mathbf{v}$ in T will be, at the same time, the closest original record to both $\mathbf{v}'_{i_1}$ and $\mathbf{v}'_{i_2}$ in T'. However, the record $\mathbf{v} \in T$ has in T' only one true perturbed image (which could be $\mathbf{v}'_{i_1}$ or $\mathbf{v}'_{i_2}$ or even another record in T'). Namely, the true global linkage is a bijection (or a perfect matching) between the records of T and those of T': each original record in T has exactly one protected image in T'. Hence, a clever attacker may try more accurate strategies to find links between records of T and records of T'. For instance, if an attacker runs the classical DBRL process and observes that more than one protected record is linked to the same original record, he would choose different candidates for the correct links. Therefore, the classical DBRL-based privacy definition for SDC methods has to be revisited.

We would like to point out that the fact that an attacker may use his background knowledge that the correct record linkage is a bijection was already observed in [35] and was implemented in `RELAIS`, a software for record linkage that was developed at the Italian National Statistical Office. It was also used in the context of record linkage; in that context it is sometimes known as the exclusivity constraint [18]. However, and maybe even surprisingly, this type of constraint has never been taken into account by the SDC community for the purpose of risk assessment.

We make a first and important step in this direction by introducing the *Global Distance-Based Record Linkage* (GDBRL) risk measure. Namely, we propose a new privacy measure that offers a more careful assessment of the risk of link disclosure for existing (or future) SDC protection methods. The new measure is also based on distances between original and protected records, but it takes into account the fact that the true linkage between the records in T and the records in T' is a perfect matching. After formally defining the new measure and describing algorithms to compute it, we ran experiments with datasets of different sizes, which are protected with different parameterizations of several SDC methods, in order to compare the values obtained by the classical DBRL measure and the new measure. The obtained results show that the new linkage strategy (which could be launched by attackers that have a global knowledge of the table) works in the majority of the cases better (and in some cases even much better) than the classical one, and so the new privacy measure is more realistic in those cases. This means that, in practice, the level of privacy preservation offered by existing SDC methods may be (much) worse than that which was implied by using the classical DBRL measure. As many of these SDC methods are being used in real-life applications by companies, cloud services, statistical agencies, etc., there is an urgent need to modify the privacy definitions that are being currently considered in this area, since presumably protected data may continue to be published under a wrong impression of providing a sufficient level of privacy for the data subjects.

1.2. Organization of the paper

In Section 2 we describe the traditional framework of SDC protection (by perturbation), we sketch four specific and very popular SDC methods, and we recall the utility and privacy measures that are typically considered in this setting. In Section 3 we introduce the new global distance-based link disclosure measure. After formally defining the measure, we explain how it can be computed for any given pair of datasets — T (original) and T' (protected). Then, we describe in Section 4 the experiments that we ran in order to compare the classical DBRL measure with the new GDBRL measure, when analyzing the level of linkage privacy that is offered by given SDC protection methods. The results show that the proposed linkage strategy is more effective than the classical ones, in the majority of the cases. We conclude the paper in Section 5 with some final remarks and possible lines for future work in this area.

2. Overview of perturbation-based methods for protecting statistical datasets

2.1. The usual framework

A statistical database is a table $T = \{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$, where for each $1 \leq n \leq N$, the record $\mathbf{v}_n$ consists of $M+L$ attributes, $\mathbf{v}_n = (\mathbf{v}_n(1), \ldots, \mathbf{v}_n(M), \mathbf{v}_n(M+1), \ldots, \mathbf{v}_n(M+L))$. Some of the attributes in the table may be found also in other publicly available databases (for example, the age of the individual who is the subject of the corresponding record, or the number of children she or he has). Such attributes may be used by an adversary in order to link records in the released de-identified database to records in publicly accessible identified databases. Those attributes are called *quasi-identifiers*. Of the remaining attributes, some may be of sensitive nature. The privacy goal is to limit an adversary from using his knowledge of the quasi-identifier values of some target individuals in order to infer those individuals' sensitive information.

Herein, we assume that the first M attributes are the quasi-identifiers, while the last L attributes are the remaining ones. For concreteness (especially in the experimental part of the paper) we will assume that the quasi-identifiers are numerical attributes; however, the ideas in the paper can easily be extended to nominal attributes.

A typical mechanism of statistical disclosure limitation alters the quasi-identifier entries of the table, but retains the remaining attributes (and, in particular, all sensitive ones). The resulting table, which is released in lieu of T, is a table $T' = \{\mathbf{v}'_1, \ldots, \mathbf{v}'_N\}$, where, for all $1 \leq n \leq N$, $(\mathbf{v}'_n(1), \ldots, \mathbf{v}'_n(M)) \in \mathbb{R}^M$ is a modification of the quasi-identifier part of $\mathbf{v}_n$, i.e. $(\mathbf{v}_n(1), \ldots, \mathbf{v}_n(M))$, while $\mathbf{v}'_n(j) = \mathbf{v}_n(j)$ for all $M+1 \leq j \leq M+L$.

2.2. Some classical SDC perturbative methods

There are three main families of SDC protection procedures [58]: (i) perturbative methods, which introduce some kind of error/change into each element of the original data, (ii) generalization methods, that replace the original values with less specific ones, and (iii) synthetic data generators, which produce synthetic data that resembles the original one.

Our newly suggested measure is relevant for methods belonging to procedures (i) or (iii), but it is less relevant for methods in procedure (ii), as we explain in Section 3.4 (see item (c) there) and Section 4.1.2. Let us sketch the working of some specific SDC protection methods (in (i) or (iii)) that process numerical attributes, and that will be used later in our experiments.

2.2.1. Additive noise

This is perhaps the simplest and most intuitive data perturbation method. In methods that are based on additive noise [30], each value $\mathbf{v}_n(j)$ of the attribute j in the original dataset T is replaced with $\mathbf{v}'_n(j) = \mathbf{v}_n(j) + \epsilon_{n,j}$, for $1 \leq n \leq N$ and $1 \leq j \leq M$, where all noise terms $\epsilon_{n,j}$ are drawn from the normal distribution $\mathcal{N}(0, \sigma_\epsilon^2)$, in the simplest implementation of noise addition. The variance σ_ϵ^2 is proportional to the variance of the corresponding original attribute. Namely, if σ_j^2 is the variance of the distribution of the values along the jth attribute in the table, then the variance that is used in generating the noise terms $\epsilon_{n,j}$ is $\sigma_\epsilon^2 = \alpha\sigma_j^2$ for some preset α. As a result, values of the jth attribute in T and in T' have the same means and covariances, but not the same variances ($\sigma_j^2 = (\sigma')_j^2/(1+\alpha)$) nor correlation coefficients.

2.2.2. Rank swapping

Rank swapping [10] with parameter p with respect to the jth attribute is defined as follows. Firstly, the records of T are sorted in an increasing order of the values $\mathbf{v}_n(j)$, $1 \leq n \leq N$. To simplify notation, let us assume that the records are already sorted, i.e. that $\mathbf{v}_n(j) \leq \mathbf{v}_m(j)$ for all $1 \leq n < m \leq N$. Then, each value $\mathbf{v}_n(j)$ is swapped with another value $\mathbf{v}_m(j)$, randomly and uniformly chosen from the limited range $n < m \leq n + p \cdot \frac{N}{100}$. When rank swapping is applied to a given dataset, the above described swapping procedure is executed for each of the quasi-identifiers in a sequential manner.

The parameter p is used to control the swapping range; it is usually considered as a percentage of N, the total number of records in T. As noted in [46], the fact that each value is swapped with a value in a fixed, closed (and possibly public) rank makes this basic rank swapping method more prone to re-identification attacks, whence the resulting privacy protection is reduced. To mitigate this drawback, two variants of rank swapping are proposed in [46], where some values (with a small but still non-negligible probability) are swapped with values out of the theoretical rank.

2.2.3. Rank Shuffling

The first idea in the design of rank shuffling [26, 25] is more or less the same as in rank swapping: each attribute is considered independently, and records are ordered according to the values of the considered attribute. But then, the swapping step is replaced with a shuffling step: a random permutation inside some block of (close) records.

The first step of rank shuffling is to sort the records of T in an increasing order of the values $\mathbf{v}_n(j)$, $1 \leq n \leq N$. Then, one selects two parameters — a width q of a sliding window and a sliding parameter $s \in [1, q]$. The initial window consists of the records $\mathbf{v}_1, \ldots, \mathbf{v}_q$. Next, all values $\mathbf{v}_n(j)$ within that window are randomly shuffled, and then the shuffling window is shifted s positions (so that the next window consists of the records with indices $s+1, \ldots, s+q$). The values in attribute j within that window are shuffled in a similar manner. That process repeats until the sliding windows cover all records in the table. The whole procedure is run for each attribute to be protected, in a sequential manner.

2.2.4. IPSO

The Information Preserving Statistical Obfuscation (IPSO) procedure was proposed in [6] and later on was improved in [44]. The idea of this synthetic method is to take the confidential attributes (denoted as T_c) of T as independent variables and the non-confidential attributes (denoted as T_{nc}) as dependent variables. A multiple regression of T_{nc} on T_c is computed; the obtained model is used to generate a new sample T'_{nc} for the non-confidential attributes. Finally, the dataset $T = T'_{nc}||T_c$ is released.

Depending on the way in which the regression is computed, preserving more or less statistical functions of the original data, three different forms of IPSO (denoted A, B and

C) can be considered. The most ambitious one is IPSO-C, which obtains the best results in terms of data utility, because the multivariate multiple regression model that is used therein preserves both the covariance matrix of T_{nc} and the matrix of regression coefficients from T_c to T_{nc}. Therefore, in our experiments we will only consider that form.

2.2.5. Data Shuffling

Data shuffling [43] can be thought of as a combination of synthetic and shuffling techniques. As in IPSO, the first step is to generate a new synthetic dataset $\tilde{T}$ by using a data model, built from multiple regression analysis on T. Then, the values of each attribute are sorted in an increasing order, in both T and $\tilde{T}$. Next, the first value of that attribute in $\tilde{T}$ is replaced with the first value in T, and so on. Finally, the sorting is undone, and the process is repeated for the following attribute to be protected. In this way, the resulting dataset T' contains exactly the same values as T, but shuffled in a controlled way.

2.2.6. Microaggregation + Differential Privacy

Microaggregation is one of the most common methods that are used to obtain k-anonymity [48, 54] for numerical data. That method starts by partitioning the data records into disjoint groups of k or more records that are close to each other. Then, all records within each of these groups are replaced by the group's centroid. In this way, the probability of identifying an original record from T in the modified database T' is limited since all k (or more) records in each group in T' have the same probability to correspond to that original record. To achieve minimum information loss, the goal is to find an optimal microaggregation that minimizes the sum of distances between original records and protected records (centroids). Since finding an optimal solution to this problem is NP-hard [47] (for the general multivariate case), many effective heuristic algorithms have been proposed to provide good quality results, like MDAV [12], CBFS[34], or the more general purpose sequential algorithm [22].

Differential privacy [14] has emerged in the last years as a rigorous theoretical privacy model. The main technique in differential privacy is to add to query results over the underlying database a noise (typically Laplacian) which is calibrated to the global sensitivity of the query (being the maximal amount by which the query result may change if one adds to the database a single record). In the SDC scenario that we consider in this work, the goal is to publish a sanitized table T' having the same structure and size as the input table T, so that the users of that data could apply on it any type of statistical analysis. The standard solution of adding Laplacian noise, in this case, should consider as possible queries all the identity queries that request a specific value of the dataset; this leads to a so large parameter for the Laplace distribution that the resulting perturbed data becomes statistically useless [7, 49]. (See [8] for a thorough discussion of limitations and challenges of the differential privacy paradigm.)

To overcome this problem and to improve the utility of methods offering differential privacy, several recent studies [9, 24, 36, 52, 53, 63] propose a hybrid approach: to combine the Laplacian noise idea with other methods like generalization or microaggregation. In our experiments, we will consider one specific SDC method resulting from this approach, that was recently proposed in [52, 53]. The idea is to first apply on T a microaggregation method; such methods usually have low global sensitivity. Then the Laplacian noise technique is applied on the resulting microaggregated dataset $\bar{T}$, by taking into account, as the queries f, the identity functions for all the centroids in $\bar{T}$. This SDC method has one parameter k for the microaggregation phase, and another parameter ϵ for the differential privacy phase.

2.3. Classical utility and privacy measures

A good SDC protection method must achieve a good trade-off between utility and privacy. In other words, the protected dataset T' must be:

- Similar enough to T so that statistical values computed on T' are very similar to those that would be computed directly on T. In other words, the (statistical) *information loss* that is caused by the transition from T to T' must be small.
- Different enough from T so that an attacker has a small probability either to precisely estimate a sensitive value of T or to correctly link an original record from T (for which he supposedly knows the quasi-identifiers) with its protected image in T'. These probabilities are denoted as the *disclosure risks*.

Information loss (IL) measures evaluate the amount of statistical utility that was lost in the protected dataset T', in comparison to the original dataset T. Several approaches are used in the SDC community to calculate the information loss. In [13] the authors calculate the average divergence of some statistical values when they are computed on both the original and the protected datasets. A probabilistic variation of these measures (PIL) was presented in [40] to ensure that the information loss value is always within the interval [0,1]. The standard PIL takes into account five specific statistical values: mean, variance, covariance, correlation coefficient and quantiles.

Regarding privacy, two types of disclosure risk measures can be considered, depending on the intention and the resources of the adversary. Firstly, if the adversary cannot obtain any information from an external data source, he can still try to get an approximation of the original values. *Interval Disclosure* (ID) is one of the approaches to model this scenario. This measure is very similar to the measure presented in [4], where the disclosure risk is quantified by measuring how closely the original values of a protected attribute can be estimated.

Secondly, an adversary who observes the protected dataset T' may know the values of some of the original attributes of T, that he has obtained from an external data source. The goal of this adversary is to link a protected record in T' with its correct original attributes (*record linkage*). Specifically, we make the usual assumption of privacy-preserving data publishing models such as k-anonymity or ℓ-diversity, that the adversary knows the value of some attributes in the table T (those are called quasi-identifiers, and as stated in Section 2.1 above, we assume that those are attributes $1 \leq m \leq M$). We also assume that the adversary acquired this knowledge for all data subjects in the table T; as stated earlier, such a strong assumption is also very common, e.g. [20, 35, 50, 51, 56, 61, 62]. Hence, the adversary has, on one hand, the projection of T onto its first M attributes, and on the other hand the published table T' in which the first M attributes are perturbed, and it includes additional L attributes that could be sensitive. The percentage of correct links established by the adversary between the original and protected datasets is therefore a measure of disclosure risk, known as the *Linkage Disclosure* (LD) risk. When the attributes are numerical, a natural option for the adversary is to find pairs of original and protected records at small (Euclidean) distance. The resulting measure is known as Distance-Based Record Linkage (DBRL for short) disclosure measure.

Let us thus proceed by formally defining the classical way in which the Distance-Based Record Linkage disclosure measure has been considered until now in the SDC literature.

Definition 2.1. *Let $d(\cdot,\cdot)$ be a metric on $\mathbb{R}^M$. For each $1 \leq n \leq N$ let B_n be the subset of records $\mathbf{v}' \in T'$ that minimize the distance $d(\mathbf{v}_n, \mathbf{v}')$, where the distance is evaluated on the projection of the two records on their first M entries. Let χ_n be an indicator variable that equals 1 if and only if $\mathbf{v}'_n$ (the real image of $\mathbf{v}_n$ in T') is in B_n. Then*

$$P(T,T') := \frac{1}{N}\sum_{n=1}^{N}\frac{\chi_n}{|B_n|}$$

is the DBRL disclosure risk measure of T' [13].

In other words, $P(T, T')$ is the average success probability of the adversary, if he tries to link $\mathbf{v} \in T$ to one of its closest records in T', where ties are broken arbitrarily.

In the next section we define and discuss some alternative DBRL disclosure risk measures that are based on a stronger adversarial assumption.

3. A global DBRL disclosure risk measure

As discussed in Section 2.3, the usual DBRL disclosure measure is based on the assumption that the adversary knows only the quasi-identifier attributes in a single record in T, say $(\mathbf{v}(1), \ldots, \mathbf{v}(M))$. When such an adversary looks for the image of $\mathbf{v}$ in T', he looks for a record $\mathbf{v}' \in T'$ that is closest to $\mathbf{v}$ in its quasi-identifier attributes (that is the strategy that underlies Definition 2.1). However, such an assumption is rather limited. For the sake of achieving a better privacy guarantee, a more widely accepted adversarial assumption (see, e.g. [20, 35, 50, 51, 56, 61, 62]) is a stronger one: the adversary knows the set of individuals who contributed their information to the database and was able to extract their quasi-identifier information from publicly available databases. Such an adversary knows the projection of T on its quasi-identifier attributes. He then wishes to trace the image of his target individual in T' in order to infer from it the corresponding sensitive information of that individual.

An adversary who knows the quasi-identifier attributes in all records in T can rule out many links that the weaker adversary may consider as possible links. Specifically, since the adversary knows that the true linkage between the records in T and T' is a bijection, he may rule out links that are not part of any such bijection. Moreover, instead of looking for a record $\mathbf{v}' \in T'$ which is closest to the original record $\mathbf{v} \in T$ that he targets, he may look for a full linkage between the records in T and those in T' where the sum of all distances between linked pairs is minimal. We proceed to formalize those adversarial strategies.

Let

$$\delta := \max_{1 \leq n \leq N} d(\mathbf{v}_n, \mathbf{v}'_n) \tag{1}$$

be the maximal distortion when releasing T' instead of T, and let Δ be some upper bound on δ. (Recall that, as discussed in Section 2.1, the record $\mathbf{v}'_n$ in T' is the distorted image of the record $\mathbf{v}_n$ in T, $1 \leq n \leq N$.) There are several scenarios that we may consider: (a) The data owner publishes the sanitized table T' together with some value $\Delta \geq \delta$, possibly even $\Delta = \delta$; such a scenario may occur in order to provide for the data miner a guarantee on the precision of the sanitized data, and, consequently, on the reliability of the analysis which is going to be performed on the sanitized data. (b) When publishing T', the data owner usually publishes information about the method of protection that has been used and the corresponding security parameters (in accord with Kerckhoffs' principle [29]); the adversary may use this information in order to derive an upper bound Δ on δ. (c) No information is released by the data owner, or the adversary is unable to derive a meaningful upper bound Δ. In that case we can take $\Delta = \infty$. In any of these cases, we will denote the release of the original table T as $\langle T', \Delta \rangle$.

Definition 3.1. *Let $\langle T', \Delta \rangle$ be a release of the table T. The corresponding bipartite graph $G = G_{T, \langle T', \Delta \rangle}$ is a graph on the set of nodes $V := T \cup T'$, where the set of edges E consists of all pairs $(\mathbf{v}, \mathbf{v}')$ where $\mathbf{v} \in T$, $\mathbf{v}' \in T'$, and $d(\mathbf{v}, \mathbf{v}') \leq \Delta$.*

Note that if the value of Δ which is released with T' is greater than or equal to $\max_{1 \leq n,m \leq N} d(\mathbf{v}_n, \mathbf{v}'_m)$ (the maximal distance between any record in T and any record in T') then Δ is redundant. In that case, G is a complete bipartite graph, containing all N^2 edges between a node in T and a node in T'.

A perfect matching in G is a set of N edges that cover all $2N$ nodes of G. Namely, it is a set of edges of the form $\{(\mathbf{v}_n, \mathbf{v}'_{\pi(n)}) : 1 \leq n \leq N\}$, where π is a permutation on $\{1, \ldots, N\}$. Clearly, G has at least one perfect matching, which is the one that corresponds to the identity permutation and describes the true linkage of all records: $\{(\mathbf{v}_n, \mathbf{v}'_n) : 1 \leq n \leq N\}$. Indeed, since $d(\mathbf{v}_n, \mathbf{v}'_n) \leq \Delta$ for all $1 \leq n \leq N$, all of those N pairs of nodes are edges in G (as implied by Definition 3.1), and that set of edges is a perfect matching in G.

There are several attack strategies that the strong adversary may attempt to launch, given the bipartite graph $G = G_{T,\langle T',\Delta\rangle}$. The basic strategy is to apply the Hungarian method (also known as the Kuhn-Munkers algorithm [31, 42]), which finds a perfect matching of minimal-cost in a weighted bipartite graph. Namely, it finds a permutation π over $\{1, \ldots, N\}$ that minimizes the sum of distances

$$\sum_{n=1}^{N} d(\mathbf{v}_n, \mathbf{v}'_{\pi(n)})$$

over all such permutations. Then, if the adversary targets a record $\mathbf{v}_n \in T$, he will link it to $\mathbf{v}'_{\pi(n)} \in T'$. In that case, the average adversarial success probability is the percentage of true links in π. This motivates the next definition.

Definition 3.2. *Let $\langle T', \Delta\rangle$ be a data release for T and let π be a minimal-cost perfect matching in the corresponding bipartite graph $G = G_{T,\langle T',\Delta\rangle}$. Then*

$$P_G(T, \langle T', \Delta\rangle) = \frac{|1 \leq n \leq N : \pi(n) = n\}|}{N} \tag{2}$$

is the Global DBRL (or GDBRL) disclosure risk measure of $\langle T', \Delta\rangle$.

3.1. On the time complexity of computing the global DBRL measure

The runtime of the Hungarian algorithm is $O(N \cdot |E|)$, where E is the set of edges in the bipartite graph $G = G_{T,\langle T',\Delta\rangle}$. If Δ is not published (or, equivalently, if $\Delta \geq \max_{1\leq n,m\leq N} d(\mathbf{v}_n, \mathbf{v}'_m)$), then G is a complete bipartite graph, containing all N^2 edges between any node in T and any node in T', whence the runtime of the algorithm is $O(N^3)$. The cubic dependence on N renders the algorithm practical only for moderately sized datasets (say, no more than $N = 100,000$), as we discuss in Section 4.3. If, however, the data owner publishes $\Delta = \delta$, where δ is as in Equation (1), then that reduces the number of edges in the bipartite graph G significantly and then it is possible to run the Hungarian algorithm on G for much larger datasets.

We proceed to describe in the next two sub-sections manners for accelerating the runtime of the Hungarian algorithm.

3.1.1. Accelerating the runtime of the Hungarian algorithm when Δ is non-trivial

If the value of Δ in the data publication $\langle T', \Delta\rangle$ is non-trivial, in the sense that it enables eliminating at least one edge from the complete bipartite graph, it is possible to further reduce the edge set E before applying the Hungarian algorithm. In order to describe that second reduction phase, we first introduce the following definition.

Definition 3.3. *An edge in G is called perfectly-matchable if there exists a perfect matching in G that includes it. The subgraph of G that consists of all perfectly-matchable edges in G is denoted G^{pm}.*

The graph $G = G_{T,\langle T',\Delta\rangle}$ with $\Delta \geq \delta$ has at least one perfect matching that consists of all true links $\{(\mathbf{v}_n, \mathbf{v}'_n) : 1 \leq n \leq N\}$, since, by the definition of δ, all those N pairs of nodes

are edges in G. All the true links are clearly perfectly-matchable, since they constitute a perfect matching. Therefore, they are all included also in G^{pm}.

It is clear that every edge of G that is not an edge in G^{pm} can be ruled out, since it cannot stand for a true link, because a true link must be perfectly-matchable. On the other hand, the edge set of G^{pm} cannot be further reduced since every such edge belongs to some perfect matching in G, and each such perfect matching describes a "possible world", namely, a possible linkage between the records of T and those of T'.

Given a bipartite graph $G = (V, E)$ that has at least one perfect matching, then with the knowledge of that perfect matching it is possible to find all perfectly-matchable edges in G in time that is linear in $|V| + |E|$ [55]. If no such perfect matching is known upfront, it is needed first to find one and then proceed to find all perfectly-matchable edges in additional linear time. Therefore, the data owner (who knows the true perfect matching $\{(\mathbf{v}_n, \mathbf{v}'_n) : 1 \leq n \leq N\}$) can execute that reduction phase in linear time; that is a reasonable thing to do, towards computing the global DBRL disclosure risk, because of the high cost of the Hungarian algorithm — $O(|V||E|)$. The adversary, on the other hand, needs to find first a perfect matching in the bipartite graph, a procedure that has runtime of $O(|V|^{1/2}|E|)$ [27].

3.1.2. Approximating the GDBRL measure

Here we propose an approach that the data owner may take in order to compute a very good approximation of the GDBRL measure. Our experiments show that this approach may significantly reduce the number of edges in the resulting bipartite graph, and, consequently, reduce the runtime of the Hungarian algorithm.

Definition 3.4. *For each $1 \leq n \leq N$, let $h(n)$ be the number of records $\mathbf{v}'_m \in T'$ such that $d(\mathbf{v}_n, \mathbf{v}'_m) < d(\mathbf{v}_n, \mathbf{v}'_n)$. Namely, $h(n)$ indicates the number of perturbed records $\mathbf{v}'_m \in T'$ that are closer to $\mathbf{v}_n$ than its true perturbed image $\mathbf{v}'_n$. In addition, $h := \max_{1 \leq n \leq N} h(n)$.*

Definition 3.5. *Let $G^1 = G^1_{T,T'}$ (resp. $G^2 = G^2_{T,T'}$) be the bipartite graph with the set of nodes $V := T \cup T'$, where an edge connects $\mathbf{v}_n \in T$ and $\mathbf{v}'_m \in T'$ if there exist at most h (resp. $h(n)$) records $\mathbf{v}'' \in T'$ such that $d(\mathbf{v}_n, \mathbf{v}'') < d(\mathbf{v}_n, \mathbf{v}'_m)$.*

Note that G^1 (resp. G^2) would have represented all possible links between records in T and T' if the data owner had published h (resp. $h(n)$, $1 \leq n \leq N$) alongside T'. However, since we do not suggest to publish that information, then G^1 and G^2 are known only to the data owner. Assuming that π^t is the minimal-cost perfect matching that was found in G^t, $t = 1, 2$, then the corresponding approximated Global DBRL disclosure risk measure of T' (denoted AGDBRL^t) is defined as the percentage of true links in π^t:

$$\text{AGDBRL}^t = \frac{|1 \leq n \leq N : \pi^t(n) = n\}|}{N}\,, \quad t = 1, 2\,. \tag{3}$$

Our experiments show that both AGDBRL^1 and AGDBRL^2 are very good approximations of GDBRL and that their computation is significantly faster than that of GDBRL; see Section 4.4 for more details.

3.2. Examples

Here, we provide two examples of a toy dataset, T, and a sanitized version of it, T', and how the classical DBRL measure $P(T, T')$ and the proposed global DBRL measure $P_G(T, \langle T', \infty \rangle)$ assess the risk in releasing those sanitized versions. In the first example, the value of GDBRL is larger than that of the DBRL measure; the second example is of a case in which the opposite occurs.

Example 1. The first dataset consists of four records $T = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4\}$ in $\mathbb{R}^2$:

$$\mathbf{v}_1 = (1,0), \mathbf{v}_2 = (0,1), \mathbf{v}_3 = (-1,0), \mathbf{v}_4 = (0,-1)\,.$$

The corresponding perturbed dataset is $T' = \{\mathbf{v}'_1, \mathbf{v}'_2, \mathbf{v}'_3, \mathbf{v}'_4\}$ where

$$\mathbf{v}'_1 = (0,0), \mathbf{v}'_2 = (0,2.1), \mathbf{v}'_3 = (-2.1,0), \mathbf{v}'_4 = (0,-2.1)\,.$$

Here, $P(T,T') = 1/4$, since the closest record to $\mathbf{v}_n$ is $\mathbf{v}'_n$ only in one of the cases ($n = 1$). However, as the minimal-cost perfect matching between the records in the two datasets is the true linkage, we get that $P_G(T, \langle T', \infty\rangle) = 1$.

Example 2. Here too we have four records in both T and T', but this time they are of dimensionality 1:

$$\mathbf{v}_1 = 1, \mathbf{v}_2 = 2, \mathbf{v}_3 = 3, \mathbf{v}_4 = 4\,.$$
$$\mathbf{v}'_1 = 2, \mathbf{v}'_2 = 3, \mathbf{v}'_3 = 4, \mathbf{v}'_4 = -0.1\,.$$

As in the previous example, $P(T,T') = 1/4$ since the closest record to $\mathbf{v}_n$ is $\mathbf{v}'_n$ only in one of the cases ($n = 1$). However, the minimal-cost perfect matching is $\{(\mathbf{v}_1, \mathbf{v}'_4), (\mathbf{v}_2, \mathbf{v}'_1), (\mathbf{v}_3, \mathbf{v}'_2), (\mathbf{v}_4, \mathbf{v}'_3)\}$, and since it has an empty intersection with the true matching $\{(\mathbf{v}_1, \mathbf{v}'_1), (\mathbf{v}_2, \mathbf{v}'_2), (\mathbf{v}_3, \mathbf{v}'_3), (\mathbf{v}_4, \mathbf{v}'_4)\}$, we get in this case $P_G(T, \langle T', \infty\rangle) = 0$.

As shown later in Section 4, our experiments found cases where the GDBRL measure (with either $\Delta = \infty$ or $\Delta = \delta$) issued risk assessments that were greater than or equal to the risk assessments of the classical DBRL measure, but also cases where the opposite happened. Therefore, both the theoretical examples in this section and the experimental results support the following recommendation: both measures — DBRL and GDBRL — should be used alongside each other in order to better assess the risk in publishing a sanitized version of a given dataset. (Moreover, since the computation of the GDBRL measure requires to compute all pairwise distances between records in T and in T', the marginal computational cost to compute also the DBRL measure is negligible.)

3.3. Computing a perturbed version of the database and evaluating its privacy preservation

Until now, the Statistical Disclosure Control community has considered the DBRL measure as the only distance-based measure of the Linkage Disclosure risk. We introduced the GDBRL measure and explained why it is essential to use that measure in order to assess the privacy level which is offered by some data release T' for the data subjects in T. Motivated by the examples in Section 3.2 and our experimental results, we propose to use the two measures together. Here we summarize the main steps in computing a perturbed version T' of T and evaluating its privacy preservation using those two measures.

1. The data owner chooses a protection method and corresponding security parameters, and uses them to compute T'.
2. He computes the maximal distortion δ of that perturbed version (see (1)), and then decides on an upper bound $\Delta \geq \delta$ to be published alongside T'.
3. He computes the usual DBRL disclosure risk measure $P(T,T')$ (Definition 2.1). If it is too large, he returns to Step 1 to recompute T', until the DBRL disclosure risk measure is sufficiently small. Otherwise, he continues to the next step.
4. He computes the GDBRL measure (or AGDBRL^1 or AGDBRL^2).
5. If the computed measure of record linkage is too large, he returns to Step 1 to recompute T'. Otherwise, he releases $\langle T', \Delta\rangle$.

We do not specify here the exact steps that are taken when either of the two disclosure risk measures is too large. If that happens (in Step 3 or Step 5 above), the data owner needs to decide how to regenerate T' towards the end of finding T' that complies with the privacy requirement. One thing that the data owner may do is to regenerate T' using the same parameters that govern the magnitude of perturbations. As the data protection mechanism is random, it is possible that by such regeneration, he may find T' that meets the required privacy goal. If several such attempts fail, the data owner may proceed to increase the perturbation governing parameters. If even such actions do not result in the sought-after privacy compliance, the data owner may switch to another SDC protection method, or even decide to withhold the release of that dataset. In the summary above we concentrate solely on the principles of the protection mechanism and, in particular, how to use the two (or more) privacy measures when computing a perturbed view of a given dataset using a given SDC method.

3.4. Concluding remarks

(a) We assumed that the adversary knows the quasi-identifier attributes in all of the records of the original dataset T. Therefore, he has two datasets — T (projected on its quasi-identifier attributes) and T' (in which the quasi-identifier attributes were subjected to some perturbation, but the sensitive attributes remained intact) — both having the same number of records, and he looks for an optimal perfect matching between them. Such an adversarial assumption is rather strong, but as we clarified earlier it is a very common one in privacy-preserving data publishing in particular and in cryptography in general, in order to ensure a higher level of privacy or secrecy. In particular, it means that if the data owner has decided to use only a subset of the records in T (for example, by removing from it outlier records) then our adversary knows what is the final subset T that was used for producing the published T'. Of-course, it is possible that the actual adversary is not that well-informed as that theoretical adversary. Namely, the actual adversary may have $|T| > |T'|$ (if he does not know what subset of T was used by the data owner for producing T') or that $|T| < |T'|$ (if he was not able to get information on the quasi-identifiers of all records in T); also a combination of those two scenarios is possible. However, even such an adversary may practice the same global attack as we described here: he could add to either T or T' dummy records until they become of equal length, construct the corresponding bipartite graph where all edges to the dummy nodes have the same weight, and then look for an optimal perfect matching in the resulting graph. The success of such an adversary is expected to be no larger than that of the theoretical adversary, because usually less information translates to less accuracy in discovering links. This is why it is customary to assume the theoretical adversary in trying to make risk assessments in this context. In addition, in trying to make risk assessments one has to model one adversary; as it is impossible to model all such partial-knowledge adversaries, the usual choice is to assume the strong adversary that we described here.

(b) There are many settings in which the sensitive information is some one-to-one mapping between two known sets of items. Two such settings are described in [15] and [33]. [15] discusses an anonymous communication system in which the linkage between sent messages and received messages is to be obfuscated. [33] discusses a release of transaction databases in which the identity of items in each transaction needs to be protected. Both studies suggest measures to assess the risk of revealing correct links. Their measures differ significantly from ours, as they are based on the number of all possible perfect matchings in the underlying bipartite graph (and not on the number of correct links in an optimal perfect matching). Computing that number is computationally intractable; moreover, those measures assess the probability of the adversary finding the true perfect matching which reveals all of the

true links, but they do not assign a succ ess probability to finding other perfect matchings that reveal some (or even most) of the true links.

(c) Both DBRL and GDBRL can be easily extended to the case of generalization, since there is a very natural definition of distance between the original records $\mathbf{v} \in T$ and the generalized records $\mathbf{v}' \in T'$. Specifically, $d(\mathbf{v}, \mathbf{v}') = \infty$ if $\mathbf{v}'$ does not generalize $\mathbf{v}$, and $d(\mathbf{v}, \mathbf{v}') = IL(\mathbf{v}')$ otherwise, where $IL(\mathbf{v}')$ is any measure of information loss in generalized records (e.g. [2, 21, 22, 45]). However, we concentrate here on perturbation-based methods for protecting statistical datasets for the following reason. When the data is being generalized, then it is always generalized until some syntactic privacy criterion is met, e.g. k-anonymity [48, 54] (every generalized record has at least $k-1$ additional generalized records that have the same generalized quasi-identifiers), or ℓ-diversity [38] (the data is generalized until the distribution of sensitive values in each block of indistinguishable records becomes sufficiently diverse). Hence, when generalization is applied, it is always applied towards such a syntactic privacy criterion. In such cases DBRL is never used, and so we do not propose to use GDBRL in such cases either. The DBRL measure (and so is the GDBRL) are intended to be used in case of perturbation, swapping or shuffling techniques. If such techniques are applied, then none of the syntactic definitions of privacy are applicable and then DBRL (or GDBRL) can be used.

4. Experimental evaluation

In this section we describe the experiments that we ran in order to demonstrate the necessity in our new global disclosure risk measures, alongside the classical DBRL measure. In our extensive experimental evaluation, we compared the disclosure risks as measured by the following measures:

- DBRL: That is the usual disclosure risk measure, Definition 2.1.
- GDBRL: That is the measure $P_G(T, \langle T', \infty \rangle)$, Definition 3.2.
- GDBRL-δ: That is the measure $P_G(T, \langle T', \delta \rangle)$, with δ as in Equation (1).

4.1. Experimental setup

4.1.1. The datasets

We consider three census datasets, called *Census*, *Large Census* and *Very Large Census*. They were extracted from the US Census Bureau using the Data Extraction System [60]. The data that was used to create this dataset was extracted from the file-group "March Questionnaire Supplement - Person Data Files" of the data source "Current Population Survey of the year 1995".

The smallest dataset *Census* contains 1080 records, and it consists of all records that are complete (in the sense that they have information for all 13 attributes). *Large Census* contains all 13518 records in which at most one attribute value is missing. (Those two datasets have already been considered in the literature before, e.g. [11].) The largest dataset *Very Large Census* contains all 27753 records in which at most two attribute values are missing.

All three datasets contain 13 attributes; a complete description of these attributes can be found in Table 1

Name	Description
AFNLWGT	Final weight (2 implied decimal places)
AGI	Adjusted gross income
EMCONTRB	Employer contribution for health insurance
ERNVAL	Business or farm net earnings in 19
FEDTAX	Federal income tax liability
FICA	Social security retirement payroll deduction
INTVAL	Amount of interest income
PEARNVAL	Total person earnings
POTHVAL	Total other persons income
PTOTVAL	Total person income
STATETAX	State income tax liability
TAXINC	Taxable income amount
WSALVAL	Amount: Total wage & salary

Table 1: Attributes Description.

4.1.2. The SDC methods

We protected the original census datasets with different parameterizations of the SDC methods that were described in Section 2.2:

1. Noise addition methods (`NA`) [30], using the following values of the standard deviation parameter: $\sigma \in \{15, 20, 25, 30\}$.

2. Rank swapping (`RSwp`), using the p-distribution variant that was proposed in [46]. The chosen values for the parameter p were $p \in \{2, 5, 10\}$.

3. Rank shuffling (`RShf`), as proposed in [25, 26], where the width of the sliding window was $q \in \{2, 3, 4\}$ and the sliding parameter was $s = q$.

4. For the IPSO synthetic method (`IPSO`) we implemented the IPSO-C model in [44]. We partitioned the attributes into two blocks (attributes 1-6 and attributes 7-13) in order to build the data model, based on the multiple regression coefficients between these two blocks. That is, the data model for attributes 1-6 is built by taking into account the regression coefficients between attributes 1-6 and attributes 7-13, and the same for the data model for attributes 7-13.

5. Data shuffling (`DShf`), as proposed in [43], where we build the data model in the same way as in IPSO-C.

6. For the method described in [52, 53], combining microaggregation and differential privacy (`MicDP`), we used $k \in \{15, 25\}$ and $\epsilon \in \{1, 10\}$. We also ran experiments with smaller values of ϵ, like $\epsilon = 0.1$, but in this case the perturbation method `MicDP` became both very secure (the resulting linkage disclosure was zero) and utterly useless (the added noise completely destroyed the utility of the data). For this reason, we decided not to include such unrealistic parameterizations in our experiments.

We did not consider k-anonymity methods (like plain microaggregation) in our experimentation for the simple reason that the DBRL measure is not relevant for such methods, and, consequently, neither are the new GDBRL measures. Indeed, since in the output of such methods each record has at least $k-1$ other records that coincide with it when projected on the quasi-identifier attributes, the DBRL measure, as well as the GDBRL measures, will

always issue values that are at most $1/k$, when applied on the output of such protection methods. Namely, in k-anonymity methods the privacy level is set upfront (by setting a value for k) and thus there is no need to evaluate the level of privacy a-posteriori. Our contribution (in the form of introducing the GDBRL measures as an enhancement of the existing DBRL measure) is relevant only for SDC methods in which there is no a-priori setting of the privacy level, and, consequently, it is necessary to evaluate the privacy a-posteriori on the resulting output T'. For such methods, the classical DBRL measure was so far the only distance-based mean for such privacy assessments, and here we suggest new so-called GDBRL measures that offer better assessments, as demonstrated by our experiments herein.

4.1.3. The experiments

For each of the SDC methods and for each setting of the relevant parameters, we evaluated and compared the three privacy measures — the classical DBRL one and the two GDBRL measures that we presented herein, all of them by taking into account the Euclidean (ℓ_2) metric,

$$d(\mathbf{v}, \mathbf{v}') = \left(\sum_{1 \leq j \leq M} |\mathbf{v}(j) - \mathbf{v}'(j)|^2 \right)^{1/2} . \tag{4}$$

(See Section 4.5 for some experiments with different distances.) The results are shown in Tables 2, 3 and 4 and are discussed below. Note that we did not evaluate the utility of those SDC methods since the contribution of the present study is not in introducing a new SDC method or in suggesting a new approach to measuring utility, but in presenting new measures to estimate the privacy that is offered by SDC methods.

All experiments were performed in an Intel Xeon CPU E5-2630L 2.00GHz with 64 GB of RAM. The operating system installed is Ubuntu 11.10, with kernel Linux 3.0.0 x86_64. We used C++ compiled with gcc 4.6.1 (with optimization flag -O3) to implement the computation of the different disclosure risk measures.

4.2. Discussion of the results

Tables 2, 3 and 4 show the disclosure risk results as obtained by the DBRL, GDBRL, and GDBRL-δ measures, for the different datasets — *Census*, *Large Census* and *Very Large Census*, respectively. Each row in any of these three tables refers to one selection of a protection method and a corresponding security parameter. For each selection of a protection method and a security parameter, we executed five different runs. The second, third and fourth columns in each table refer to the three disclosure risk measures (DBRL, GDBRL, and GDBRL-δ) and they show the average number of correctly linked records, and the average value of the disclosure risk measure (which equals the percentage of the correctly linked records).

As emerge from our results, in many cases the proposed GDBRL measures issue risk assessments that are larger (or equal) to the corresponding risk assessment of the classical DBRL measure. More precisely, from the 48 cases included in our experiment (16 protection methods and parameter settings for each of the 3 datasets), the GDBRL measures gives larger values in 28 cases, while the DBRL measure gives larger values in the other 20 cases. However, out of these latter 20 cases, there are 10 cases where the DBRL value (and hence also the GDBRL value) was less than 0.5%, which means that the difference is quite meaningless. More importantly, for those parameterizations of the corresponding perturbation methods, the risk is very low and so the utility of the methods becomes also very low. Most of these cases correspond to the perturbation method `MicDP`.

As illustrative examples, if we look at the results in Table 3 for the protected table that was obtained by additive noise with $\sigma = 25$, we observe a risk increase of 7% (for GDBRL)

or 6% (for GDBRL-δ), with respect to the classical DBRL risk. The protection method where the increase in the risk evaluation is most prominent is IPSO-C; for that method, the difference in the risk assessments are 17%, 55% and 61% for *Census*, *Large Census* and *Very Large Census*, respectively. These particular results for IPSO-C illustrate the message of this work perfectly: some methods that could be considered as *quite secure* by the SDC community may turn out to be *quite insecure* once the new GDBRL measures are taken into account.

SDC method		DBRL Links	DBRL % DR	GDBRL Links	GDBRL % DR	GDBRL-δ Links	GDBRL-δ % DR
NA	$\sigma = 15$	971.6	0.90	1060.6	0.98	974.0	0.90
	$\sigma = 20$	840.4	0.78	1022.2	0.95	863.0	0.80
	$\sigma = 25$	680.0	0.63	912.2	0.84	842.8	0.78
	$\sigma = 30$	556.2	0.51	786.2	0.73	786.2	0.73
RSwp	$p = 2$	957.8	0.89	1030.8	0.95	968.2	0.90
	$p = 5$	411.6	0.38	583.2	0.54	554.2	0.51
	$p = 10$	74.4	0.07	137.2	0.13	131.8	0.12
RShf	$q = 2$	89.8	0.08	59.4	0.06	59.4	0.06
	$q = 3$	41.2	0.04	27.2	0.03	27.2	0.03
	$q = 4$	21.2	0.02	19	0.02	19	0.02
IPSO		415.0	0.38	1074.0	0.99	1074 .0	0.99
DShf		27.7	0.03	22.0	0.02	22.0	0.02
MicDP	$k = 15\ \epsilon = 1$	1.4	0.00	1.6	0.00	1.6	0.00
	$k = 15\ \epsilon = 10$	1.4	0.00	0.8	0.00	0.8	0.00
	$k = 25\ \epsilon = 1$	0.8	0.00	2.4	0.00	2.4	0.00
	$k = 25\ \epsilon = 10$	3.4	0.00	1.8	0.00	1.8	0.00

Table 2: Disclosure risk results for the census dataset

An interesting, albeit not surprising observation is that, according to all measures, the risk decreases when the size of the database increases. This effect is very evident when we move from the *Census* dataset to the *Large Census* dataset (see, for instance, the GDBRL values for additive noise in Tables 2 and 3). Basically, when the size of the dataset increases, it is more difficult for the adversary to find the correct links (with both the classical and the new approaches) because the number of close records in T' for a given record $\mathbf{v}_n \in T$ increases.

4.3. Scalability of the Hungarian method

The main problem when applying the Hungarian method on very large graphs (in our case, when the datasets are very large) is the amount of information that must be stored in memory during the optimization process. The Hungarian method stores the weights of all edges in the bipartite graph. In our case, if the bipartite graph is complete, then the distances $d(\mathbf{v}, \mathbf{v}')$ for all $\mathbf{v} \in T$ and $\mathbf{v}' \in T'$ must be stored, which means a memory cost of $O(N^2)$, where N is the number of records in the dataset. Furthermore, if we want to optimize the runtime of the Hungarian method by simplifying some data access operations, the real memory cost can be even worse.

The only solution to manage this amount of data in cases where N is very large is to use hard disk memory. Alas, this solution increases the time cost of each access to data. This may have a serious impact on the global runtime (which is $O(N^3)$ when the bipartite graph is complete), rendering the algorithm impractical for values $N \geq 100,000$, at least for standard computers like the ones that we used in our experiments.

SDC method		DBRL Links	DBRL % DR	GDBRL Links	GDBRL % DR	GDBRL-δ Links	GDBRL-δ % DR
NA	$\sigma = 15$	4485.8	0.33	5810.6	0.43	5451.0	0.40
	$\sigma = 20$	2830.4	0.21	3943.8	0.29	3643.6	0.27
	$\sigma = 25$	1899.4	0.14	2825.6	0.21	2658.6	0.20
	$\sigma = 30$	1355.0	0.10	2107.0	0.16	2107.0	0.16
RSwp	$p = 2$	4660.2	0.34	5838.0	0.43	5782.6	0.43
	$p = 5$	451.6	0.03	779.4	0.06	774.4	0.06
	$p = 10$	48.0	0.00	103.6	0.01	105.2	0.01
RShf	$q = 2$	760.8	0.06	254.4	0.02	254.4	0.02
	$q = 3$	233.2	0.02	64.4	0.00	64.4	0.00
	$q = 4$	98.4	0.01	32.4	0.00	32.4	0.00
IPSO		1496.0	0.11	8963.0	0.66	8963.0	0.66
DShf		51.0	0.00	43.7	0.00	43.7	0.00
MicDP	$k = 15\ \epsilon = 1$	1.2	0.00	1.4	0.00	0.6	0.00
	$k = 15\ \epsilon = 10$	1.0	0.00	0.8	0.00	0.5	0.00
	$k = 25\ \epsilon = 1$	0.8	0.00	0.6	0.00	0.8	0.00
	$k = 25\ \epsilon = 10$	1.8	0.00	1.0	0.00	1.4	0.00

Table 3: Disclosure risk results for the large census dataset

Several solutions to this problem may be considered if more sophisticated resources are available. For instance, the usage of persistent memories (such as flash memory, which is currently the only available possibility) should decrease the time cost of each access to data stored in rotational disks.

A different solution could consist in distributing the data (the input for the Hungarian algorithm) across many computing nodes, and providing the involved applications with a single namespace, which can be done by a distributed file system (as in MapReduce Distributed File System [5, 19]) or by using key/value pairs (as in most NoSQL databases [32, 41] and systems like the one in [16]).

In the near future, such solutions might be combined with other technological advances, like remote direct memory access (RDMA) or new persistent memories like Solid State Disks (SSD), so that the scalability problem of the Hungarian algorithm is overcome, and then the risk measures that we introduced in this work could be computed even for much larger datasets.

4.4. Approximating the GDBRL measure

Here we report experiments with the approximations of GDBRL, AGDBRL^1 and AGDBRL^2, as defined in Section 3.1.2. As these measures are suggested for use only when the dataset is large, we focused on the largest dataset *Very Large Census* and evaluated them in each of our experiments on that dataset. The results are shown in the last two columns in Table 4. We can see there that AGDBRL^1 is an extremely good approximation for GDBRL. The maximal deviation between the two is of 0.0036% (in MicDP with $k = 25$ and $\epsilon = 1$), while the average deviation over all 16 experiments is 0.000495%. AGDBRL^2 is somewhat less accurate approximation but it is still a very good one: the maximal deviation for that approximate measure is 0.97% (for RSwp with $p = 5$) and the average deviation over all 16 experiments is of 0.0817%. Table 5 shows the runtimes of computing GDBRL, AGDBRL^1, and AGDBRL^2 on the *Very Large Census* dataset.

As can be seen in the table, there is a small number of cases in which using either of the approximate measures does not reduce the computation runtime. This happens because the

SDC method		DBRL Links	DBRL % DR	GDBRL Links	GDBRL % DR	GDBRL-δ Links	GDBRL-δ % DR	AGDBRL^1 Links	AGDBRL^1 % DR	AGDBRL^2 Links	AGDBRL^2 % DR
NA	$\sigma = 15$	8181	0.29	10651	0.38	10651	0.38	10651	0.38	10651.0	0.38
	$\sigma = 20$	5314	0.19	7366	0.27	7366	0.27	7366	0.27	7366.0	0.27
	$\sigma = 25$	3603	0.13	5384	0.19	5383	0.19	5384	0.19	5383.8	0.19
	$\sigma = 30$	2570	0.09	4050	0.15	4049	0.15	4050	0.15	4049.2	0.15
RSwp	$p = 2$	8108	0.29	10114	0.36	10114	0.36	10114	0.36	10114.2	0.36
	$p = 5$	730	0.03	1294	0.05	1115	0.04	1294	0.05	1025.4	0.04
	$p = 10$	65	0.00	161	0.01	58	0.00	161	0.01	56.2	0.00
RShf	$q = 2$	1342	0.05	344.4	0.01	344.4	0.01	344.4	0.01	356.5	0.01
	$q = 3$	337.4	0.01	77.4	0.00	77.4	0.00	77.4	0.00	77.4	0.00
	$q = 4$	152.6	0.01	39.4	0.00	39.4	0.00	39.4	0.00	40	0.00
IPSO		2965	0.11	7658	0.28	7658	0.28	7658	0.28	7658.0	0.28
DShf		81.7	0.00	72.0	0.00	72.0	0.00	72.0	0.00	71.7	0.00
MicDP	$k = 15\ \epsilon = 1$	1.2	0.00	0.8	0.00	0.8	0.00	0.8	0.00	0.8	0.00
	$k = 15\ \epsilon = 10$	0.8	0.00	0.8	0.00	0.4	0.00	0.4	0.00	0.4	0.00
	$k = 25\ \epsilon = 1$	1.2	0.00	0.4	0.00	1.4	0.00	1.4	0.00	1.4	0.00
	$k = 25\ \epsilon = 10$	1.2	0.00	2.2	0.00	1.4	0.00	1.4	0.00	1.4	0.00

Table 4: Disclosure risk results for the very large census dataset

perturbation that was applied in those cases was very strong, whence the distance between an original record and its protected version became large. Such large distances gave rise to large values of $h(n)$ in Definition 3.4 and, as a result, the number of edges that were removed from the bipartite graph was much smaller. Moreover, such edges are of such high weight that they are not visited by the Hungarian algorithm. Therefore, the benefit in terms of runtime from removing those edges is negligible. Since the runtime for computing the AGDBRL measures also includes the computation of the values $h(n)$, $1 \leq n \leq N$, and the pruning of the bipartite graph, the overall runtime for computing the AGDBRL measures becomes slightly larger than that of computing of GDBRL measure. However, in most of the cases the runtime for computing AGDBRL^1 or AGDBRL^2 is lower than the runtime for computing GDBRL, sometimes even substantially. Given the high accuracy of those approximate AGDBRL measures, we suggest adopting them for larger datasets because of their potential to reduce the computation time significantly.

4.5. Using other metrics

The definition of the DBRL and GDBRL measures depend on a definition of a metric on $\mathbb{R}^M$, $d(\cdot, \cdot)$. In all of our experiments so far we used the Euclidean (ℓ_2) metric, see Equation (4). Here we consider the possibility of using other metrics when running the different record linkage processes based on distances between original and protected records. In particular, we consider the absolute (or ℓ_1) distance,

$$d(\mathbf{v}, \mathbf{v}') = \sum_{1 \leq j \leq M} |\mathbf{v}(j) - \mathbf{v}'(j)|\,,$$

and the Mahalanobis distance [39], which has been previously considered by the SDC community for the purpose of record linkage [59].

We have repeated the DBRL, GDBRL and GDBRL-δ experiments for the smallest dataset, *Census*, in order to compare the results obtained with these three different distances. The average number of correct links can be found in Table 6. For the noise addition

SDC method		GDBRL	AGDBRL[1]	AGDBRL[2]
NA	$\sigma = 15$	515.00	81.48	42.85
	$\sigma = 20$	550.67	151.91	72.16
	$\sigma = 25$	676.09	241.21	103.36
	$\sigma = 30$	279.07	345.57	162.36
RSwp	$p = 2$	3986.45	1106.39	853.60
	$p = 5$	2919.99	2058.45	1537.81
	$p = 10$	13666.98	4286.49	285.13
RShf	$q = 2$	11609.15	11273.10	1045.08
	$q = 3$	15428.51	14062.25	853.60
	$q = 4$	20489.06	17535.62	688.85
IPSO		5098.69	3960.37	2743.87
DShf		2278.34	2424.80	1806.22
MicDP	$k = 15\ \epsilon = 1$	34001.49	28811.26	28811.26
	$k = 15\ \epsilon = 10$	29812.22	28529.26	28529.26
	$k = 25\ \epsilon = 1$	31056.93	27154.10	27154.10
	$k = 25\ \epsilon = 10$	24335.04	25503.12	25503.12

Table 5: Runtimes (in seconds) for computing the global measures for the very large census dataset

and rank swapping protection methods, the best linkage results are obtained by considering the Euclidean distance. For rank shuffling, IPSO and data shuffling, considering the absolute distance leads to more correct links. Furthemore, in some of these cases (rank and data shuffling), the combination of standard DBRL and the absolute distance is the one which leads to the best linkage results (outperforming GDBRL). For IPSO, once again, the GDBRL measure widely outperforms the classical DBRL measure.

The conclusions from this experiment reinforce the main message of this study: in order to have a more precise analysis of the privacy offered by a perturbation method when applied to a particular dataset, it is preferable to consider several privacy measures, for instance, to consider both DBRL and GDBRL measures, with different metrics.

4.6. Categorical attributes

Even if distance-based record linkage seems more suitable for the case of numerical attributes, the ideas presented in this study can be extended also to the case of categorical attributes. A first problem is that many protection methods in this setting are based on suppression or generalization, towards meeting some syntactic privacy condition such as k-anonymity or ℓ-diversity; as explained in Section 4.1.2 or in item (c) in Section 3.4, distance-based measures are less suitable in the presence of such a syntactic privacy condition. Hence, we consider here a perturbative method for categorical data which does not apply suppressions or generalizations; the method that we chose is PRAM [23]. The idea of PRAM is that the values of a categorical variable in the original dataset are changed into other categories, taking into account pre-defined probabilities, which are stored in a transition (Markov) matrix. These probabilities can depend on a parameter α. It is also possible to apply the perturbation to sub-groups of records independently, if a stratification mode is chosen.

If we assume that the categorical attributes in the dataset do not have a hierarchical/sorted relation, then the natural distance to be considered, for the comparison between original and protected records, is the Hamming distance, $d(\mathbf{v}, \mathbf{v}') = |\{j : \mathbf{v}(j) \neq \mathbf{v}'(j)\}|$. Of course, if there is some (semantic) structure or order or hierarchy that can be defined on the set of values of some of the categorical attributes, then other distances can be considered.

SDC method		DBRL			GDBRL			GDBRL-δ		
		Eucl.	Abs.	Mah.	Eucl.	Abs.	Mah.	Eucl.	Abs.	Mah.
NA	$\sigma = 15$	971.6	960.4	524.6	1059.0	1046.8	571.8	974.6	963.4	526.2
	$\sigma = 20$	840.4	824.6	419.8	1021.9	1002.7	510.5	863.1	846.9	431.1
	$\sigma = 25$	680.0	663.2	308.4	911.9	889.4	413.6	841.8	821.0	381.8
	$\sigma = 30$	556.2	536.2	249.2	784.2	756.0	351.4	784.2	756.0	351.4
RSwp	$p = 2$	957.8	978.6	519.4	1030.6	1053.0	558.9	968.7	989.8	525.3
	$p = 5$	411.6	383.8	163.4	583.6	544.2	231.7	554.8	517.3	220.2
	$p = 10$	74.4	59.0	31.6	137.7	109.2	58.5	131.7	104.4	55.9
RShf	$q = 2$	89.8	167.4	34.2	59.4	110.8	22.6	59.4	110.8	22.6
	$q = 3$	41.2	82.6	20.8	27.1	54.4	13.7	27.1	54.4	13.7
	$q = 4$	21.2	44.2	10.4	19.2	40.0	9.4	19.2	40.0	9.4
IPSO		415.0	676.0	676.0	1074.0	1074.0	1074.0	1074.0	1074.0	1074.0
DShf		27.7	32.7	26.3	22.4	26.5	21.3	22.4	26.5	21.3
MicDP	$k = 15\ \epsilon = 1$	1.4	1.0	1.4	1.6	1.1	1.6	1.6	1.1	1.6
	$k = 15\ \epsilon = 10$	1.4	2.0	1.4	0.8	1.1	0.8	0.8	1.1	0.8
	$k = 25\ \epsilon = 1$	0.8	1.0	1.4	2.4	3.0	4.2	2.4	3.0	4.2
	$k = 25\ \epsilon = 10$	3.4	4.8	4.0	1.8	2.5	2.1	1.8	2.5	2.1

Table 6: Average number of correct links for the census dataset (1080 records), with 3 different distances

We ran experiments on two datasets with categorical attributes: the Household dataset is obtained by selecting the 493 different records (repeated records are deleted) in the dataset testdata in [57], which comes from the International Household Survey Network [28]. Each record contains values for 8 categorical attributes. The Adult dataset is obtained by first selecting the 8 categorical atributes in the Adult Data Set available at the UCI repository [1], extracted from the US Census Bureau, and then deleting the repeated records, that is, those with identical values for these 8 attributes; this results in a categorical dataset with 8688 different records. We protected these two datasets with the version of PRAM available at the R-sdcMicro package [57], with $\alpha = 0.5$ and with no stratification. Then we computed the DBRL, GDBRL and GDBRL-δ measures by taking into account the original and protected datasets, and the Hamming distance. The obtained results are shown in Table 7. Once again, in these examples, the number of correct links obtained with the GDBRL strategy is higher than with the DBRL strategy. Hence, also in scenarios that involve categorical attributes, other distance-based measures like GDBRL, and not only DBRL, should be considered to analyze the privacy level.

Dataset (#records)	DBRL		GDBRL		GDBRL-δ	
	Links	% DR	Links	% DR	Links	% DR
Household (493)	237	0.48	280	0.56	270	0.54
Adult (8688)	4712	0.54	5564	0.64	5376	0.62

Table 7: Disclosure risk results for the Housing and Adult dataset, protected with PRAM

5. Conclusion and future work

In this work we have revisited the classical measure of distance-based record linkage (DBRL) risk, that is widely considered by the Statistical Disclosure Control (SDC) community when analyzing the privacy offered by specific dataset protection methods, in the

scenario of privacy-preserving data publishing. We have proposed another measure, the global distance-based record linkage (GDBRL) risk, which takes into account the fact that for any SDC protection method there exists a bijection between the original records in the dataset T and the protected records in the sanitized (and released) dataset T'. We ran experiments with different datasets and well-known protection methods, in order to compare the results obtained by the classical DBRL measure and the new GDBRL measure. The results of the experiments support our initial intuition: the new measure issues in many cases higher risk values than the ones issued by the classical measure. In some cases, those differences are quite significant. Hence, relying solely on the classical DBRL measure in order to assess the privacy risks for a given sanitized version might be quite misleading: the true risk can be much higher than the risk as estimated by that measure.

Therefore, we believe that there is an urgent need to add the new risk measure, GDBRL, to the set of measures that should be taken into account when analyzing the privacy offered by SDC protection methods. We stress that, in our opinion, measures like DBRL or GDBRL must co-exist with other approaches like differential privacy, in this setting of privacy-preserving data publishing.

As we have discussed in Section 4.3, the computation of the new GDBRL measure may be prohibitive if the datasets are very large, due to the memory and time complexities of the Hungarian algorithm. As future work, we would like to consider possible solutions to this scalability problem, which might involve the usage of more complex computing architectures.

A different line of future work is the study of disclosure risks that emerge from combining record linkage and interval disclosure techniques. For instance, an interval disclosure mechanism could be run first in order to select records with very similar (expected) confidential values, and then the record linkage process could be run on subsets of similar records only.

Bibliography

[1] Adult Data Set, https://archive.ics.uci.edu/ml/datasets/Adult.

[2] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Anonymizing tables. In *International Conference on Database Theory (ICDT)*, volume 3363 of LNCS, pages 246–258, 2005.

[3] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *SIGMOD Conference*, pages 439–450, 2000.

[4] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *Proc. of the ACM SIGMOD Int. Conf. on Management of Data*, pages 439–450, 2000.

[5] Apache Software Foundation. Hadoop Distributed File System (HDFS) Architecture.

[6] J. Burridge. Information preserving statistical obfuscation. *Statistics and Computing*, 13:321–327, 2003.

[7] R. Chen, N. Mohammed, B. Fung, B. Desai, and L. Xiong. Publishing set-valued data via differential privacy. *Proceedings of the VLDB Endowment*, 4(11):1087–1098, 2011.

[8] C. Clifton and T. Tassa. On syntactic anonymity and differential privacy. *Trans. on Data Privacy*, 6:161–183, 2013.

[9] G. Cormode, C. M. Procopiuc, D. Srivastava, E. Shen, and T. Yu. Differentially private spatial decompositions. In *ICDE*, pages 20–31, 2012.

[10] T. Dalenius and S. Reiss. Data-swapping: a technique for disclosure control. *Journal of Statistical Planning and Inference*, 6:73–85, 1982.

[11] J. Domingo-Ferrer, A. Martínez-Ballesté, J. Mateo-Sanz, and F. Sebé. Efficient multivariate data-oriented microaggregation. *The Very Large Database Journal*, 15:355–369, 2006.

[12] J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. on Knowledge and Data Engineering*, 14(1):189–201, 2002.

[13] J. Domingo-Ferrer and V. Torra. *Disclosure control methods and information loss for microdata*, pages 91–110. Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, 2001.

[14] C. Dwork. Differential privacy. In *ICALP (2)*, pages 1–12, 2006.

[15] M. Edman, F. Sivrikaya, and B. Yener. A combinatorial approach to measuring anonymity. In *IEEE International Conference on Intelligence and Security Informatics, ISI 2007, New Brunswick, New Jersey, USA, May 23-24, 2007, Proceedings*, pages 356–363, 2007.

[16] B. Fitzpatrick. Distributed caching with memcached. *Linux J.*, 2004(124):5–, Aug. 2004.

[17] B. Fung, K. Wang, R. Chen, and P. Yu. Privacy-preserving data publishing: a survey of recent developments. *ACM Computing Surveys (CSUR)*, 42(4):1–53, 2010.

[18] L. Getoor and A. Machanavajjhala. Entity resolution for big data. In *KDD*, page 1527, 2013.

[19] S. Ghemawat, H. Gobioff, and S.-T. Leung. The google file system. *SIGOPS Oper. Syst. Rev.*, 37(5):29–43, 2003.

[20] A. Gionis, A. Mazza, and T. Tassa. k-Anonymization revisited. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 744–753, 2008.

[21] A. Gionis and T. Tassa. k-Anonymization with minimal loss of information. *IEEE Transactions on Knowledge and Data Engineering*, 21:206–219, 2009.

[22] J. Goldberger and T. Tassa. Efficient anonymizations with enhanced utility. *Transactions on Data Privacy*, 3:149–175, 2010.

[23] J. Gouweleeuw, P. Kooiman, L. Willenborg, and P.-P. D. Wolf. Post-randomisation for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, 14(4):463–478, 1998.

[24] M. Hay, V. Rastogi, G. Miklau, and D. Suciu. Boosting the accuracy of differentially private histograms through consistency. *PVLDB*, 3:1021–1032, 2010.

[25] J. Herranz and J. Nin. Secure and efficient anonymization of distributed confidential databases. *International Journal of Information Security*, 13(6):497–512, 2014.

[26] J. Herranz, J. Nin, and V. Torra. Distributed privacy-preserving methods for statistical disclosure control. In *Proceedings of the Workshop on Data Privacy Management (DPM/SETOP)*, volume 5939 of *Lecture Notes in Computer Science*, pages 33–47, 2009.

[27] J. Hopcroft and R. Karp. An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. *SIAM Journal on Computing*, 2:225–231, 1973.

[28] International Household Survey Network, http://www.ihsn.org/.

[29] A. Kerckhoffs. La cryptographie militaire. *Journal des sciences militaires*, 9:161–191, 1883.

[30] J. Kim. A method for limiting disclosure in microdata based on random noise and transformation. In *Proceedings of the ASA Section on Survey Research Methodology*, pages 303–308, 1986.

[31] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

[32] A. Lakshman and P. Malik. Cassandra: a structured storage system on a p2p network. In *Proceedings of the twenty-first annual symposium on Parallelism in algorithms and architectures*, SPAA '09, pages 47–47, New York, NY, USA, 2009. ACM.

[33] L. V. S. Lakshmanan, R. T. Ng, and G. Ramesh. On disclosure risk analysis of anonymized itemsets in the presence of prior knowledge. *TKDD*, 2, 2008.

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

[35] R. Lenz. A graph theoretical approach to record linkage. *Joint ECE/Eurostat work session on statistical data confidentiality*, Working paper no. 35, 2003.

[36] N. Li, W. Qardaji, and D. Su. On sampling, anonymization, and differential privacy: Or, k-anonymization meets differential privacy. In *7th ACM Symposium on Information, Computer and Communications Security (ASIACCS'2012)*, pages 32–33, 2012.

[37] Y. Lindell and B. Pinkas. Privacy preserving data mining. In *CRYPTO*, pages 36–54, 2000.

[38] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *IEEE Int. Conf. on Data Engineering*, 2006.

[39] P. Mahalanobis. On the generalised distance in statistics. In *Proceedings of the National Institute of Science of India*, volume 12, pages 49–55, 1936.

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

[41] MongoBD, http://www.mongodb.org.

[42] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5:32–38, 1957.

[43] K. Muralidhar and R. Sarathy. Data shuffling: A new masking approach for numerical data. *Management Science*, 52(5):658–670, 2006.

[44] K. Muralidhar and R. Sarathy. Generating sufficiency-based non-synthetic perturbed data. *Transactions on Data Privacy*, 1(1):17–33, 2008.

[45] M. Nergiz and C. Clifton. Thoughts on k-anonymization. *Data and Knowledge Engineering*, 63(3):622–645, 2007.

[46] J. Nin, J. Herranz, and V. Torra. Rethinking rank swapping to decrease disclosure risk. *Data and Knowledge Engineering*, 64(1):346–364, 2008.

[47] A. Oganian and J. Domingo-Ferrer. On the complexity of optimal microaggregation for statistical disclosure control. *Statistical Journal United Nations Economic Commission for Europe*, 18(4):345–354, 2000.

[48] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information (abstract). In *PODS '98: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, page 188, 1998.

[49] R. Sarathy and K. Muralidhar. Evaluating laplace noise addition to satisfy differential privacy for numeric data. *Transactions on Data Privacy*, 4(1):1–17, 2011.

[50] E. Shmueli and T. Tassa. Privacy by diversity in sequential releases of databases. *Inf. Sci.*, 298:344–372, 2015.

[51] E. Shmueli, T. Tassa, R. Wasserstein, B. Shapira, and L. Rokach. Limiting disclosure of sensitive data in sequential releases of databases. *Information Sciences*, 191:98–127, 2012.

[52] J. Soria-Comas. Improving data utility in differential privacy and k-anonymity. Phd. thesis, Universitat Rovira i Virgili, 2013.

[53] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez. Enhancing data utility in differential privacy via microaggregation-based k-anonymity. *VLDB J.*, 23(5):771–794, 2014.

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

[55] T. Tassa. Finding all maximally-matchable edges in a bipartite graph. *Theoretical Computer Science*, 423:50–58, 2012.

[56] T. Tassa, A. Mazza, and A. Gionis. k-Concealment: an alternative model of k-type anonymity. *Transactions on Data Privacy*, 5:189–222, 2012.

[57] M. Templ, A. Kowarik, and B. Meindl. sdcMicro: Statistical Disclosure Control methods for anonymization of microdata and risk estimation, Available at http://cran.r-project.org/web/packages/sdcMicro/.

[58] V. Torra. *Handbook of Data Mining*, chapter Privacy in Data Mining. Human Factor and Ergonomics, 2009.

[59] V. Torra, J. Abowd, and J. Domingo-Ferrer. Using mahalanobis distance-based record linkage for disclosure risk assessment. In *Proceedings of PSD 2006*, volume 4302 of *Lecture Notes in Computer Science*, pages 233–242, 2006.

[60] U.S. Census Bureau, Data Extraction System, http://www.census.gov/.

[61] W. K. Wong, N. Mamoulis, and D. W.-L. Cheung. Non-homogeneous generalization in privacy preserving data publishing. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*, pages 747–758, 2010.

[62] X. Xiao and Y. Tao. M-invariance: towards privacy preserving re-publication of dynamic datasets. In *Proceedings of the ACM SIGMOD International Conference on Management of data (SIGMOD)*, pages 689–700, 2007.

[63] J. Xu, Z. Zhang, X. Xiao, Y. Yang, and G. Yu. Differentially private histogram publication. In *ICDE*, pages 32–43, 2012.