Data Mining and Knowledge Discovery
https://doi.org/10.1007/s10618-021-00778-0

Differentially Private Distance Learning in Categorical Data

Elena Battaglia[1] · **Simone Celano**[1] · **Ruggero G. Pensa**[1]

Received: 20 November 2020 / Accepted: 26 June 2021

Abstract
Most privacy-preserving machine learning methods are designed around continuous or numeric data, but categorical attributes are common in many application scenarios, including clinical and health records, census and survey data. Distance-based methods, in particular, have limited applicability to categorical data, since they do not capture the complexity of the relationships among different values of a categorical attribute. Although distance learning algorithms exist for categorical data, they may disclose private information about individual records if applied to a secret dataset. To address this problem, we introduce a differentially private family of algorithms for learning distances between any pair of values of a categorical attribute according to the way they are co-distributed with the values of other categorical attributes forming the so-called context. We define different variants of our algorithm and we show empirically that our approach consumes little privacy budget while providing accurate distances, making it suitable in distance-based applications, such as clustering and classification.

Keywords Differential privacy · Metric learning · Categorical attributes · Distance-based methods

1 Introduction

Most machine learning and data analysis methods rely, directly or indirectly, on their ability to compute distances or similarities between data objects. Distance-based clustering (e.g., k-means, hierarchical clustering, k-medoids) is only the most glaring example, but distance-based methods range from classification (e.g., kNN, SVM) and anomaly detection algorithms to proximity search (e.g., nearest neighbor search).

Responsible editor: Annalisa Appice, Sergio Escalera, Jose A. Gamez, Heike Trautmann

✉ Ruggero G. Pensa
ruggero.pensa@unito.it

Elena Battaglia
elena.battaglia@unito.it

[1] Dept. of Computer Science, University of Turin, Turin, Italy

Published online: 13 July 2021

Furthermore, many graph-based methods (spectral clustering (Shi and Malik 1997), semi-supervised label propagation (Yamaguchi et al. 2016), graph convolutional neural networks (Velickovic et al. 2018) leverage distances or similarities among data objects to compute adjacency matrices or k-NN graphs and perform consequent operations on them. Although different definitions of distance/similarity exist, they are relatively easy to compute, provided that data are given in form of numeric vectors. Additionally, for most of the above-mentioned distance-based methods, differentially private counterparts of them have been proposed as well. Differential privacy (Dwork and Roth 2014) is a computational paradigm which guarantees that the output of a statistical query applied to a secret dataset does not allow to understand whether a particular data object is present in the dataset or not. In recent years, many differentially private variants have been proposed for most distance based algorithms, including kNN (Gursoy et al. 2017), SVM (Chaudhuri et al. 2011) and k-means (Su et al. 2017).

When data are described by categorical features/attributes, instead, distances can only account for the match or mismatch of the values of an attribute between two data objects, leading to poorer and less expressive proximity measures (e.g., the Jaccard similarity). And yet, intuitively, a patient whose disease is "gastritis" should be closer to a patient affected by "ulcer" than to one having "migraine"[1]. An efficient solution consists in using some distance learning algorithm to infer the distance between any pair of different values of the same categorical attribute from data. Among all existing methods, DILCA (Ienco et al. 2012) is one of the most effective, although, more recently, other distance learning techniques have been proposed for ordinal data only (Zhang and Cheung 2020). DILCA's objective is to compute the distance between any pair of values of a categorical attribute by taking into account the way the two values are co-distributed with respect to the values of other categorical attributes forming the so-called context. According to DILCA, if two values of a categorical attribute are similarly distributed w.r.t. the values of the context attributes, then their distance is lower than that computed for two values of the same attribute that are divergently distributed w.r.t. the values of the same context attributes. DILCA has been successfully used in different scenarios including clustering (Ienco et al. 2012), semi-supervised learning (Ienco and Pensa 2016) and anomaly detection (Ienco et al. 2017). However, if applied to a secret dataset, it may disclose a lot of private information. To understand this point, let us consider the following example.

Example 1 (motivating example) Let us consider a dataset containing information about people living in a country. For each person, there are only two pieces of information: the city of residence and her or his income. DILCA uses the information carried by attribute "income" (the so-called context) to compute the distances between the values of attribute "city of residence". Thus, we expect close cities (according to this distance) to have a similar average income; on the other side, we expect cities with a different average income level to be far one from each other. Let us suppose now that DILCA returns a very small distance between cities A and B. A is a notoriously rich city, while B is a small village in which we would expect to have an average income much lower than that of A. On the other hand, we know that a very rich person (able

[1] Although semantic similarities could be exploited as well, ontologies or taxonomies of categorical values are not always available.

to significantly raise B's average income level on its own) lives in B. From the result of the distance computation between A and B we are therefore able to understand that the "atypical" person was probably included in the secret dataset, harming his privacy.

In this paper, we address the problem of learning meaningful distances for categorical data in a differentially private way. To this purpose, we first introduce a straightforward extension of DILCA where the values co-occurrence counts of two different categorical attributes are computed using the Laplace mechanism (Dwork and Roth 2014). However, we show that this algorithm consumes too much privacy budget, and propose less expensive alternatives (adopting either the Laplace or the exponential mechanisms). We proof theoretically that our distance-learning algorithms satisfy ε-differential privacy, and show experimentally that they provide accurate distances even with relatively small values of privacy budget ε. Additionally, we show that our family of algorithms (which we call DP-DILCA) is effective in two distance-based learning scenarios, including clustering and k-NN classification.

2 Background and Related Work

In this section, we introduce the necessary background required to understand the theoretical foundations of our method and, contextually, we introduce its related scientific literature.

2.1 Differential Privacy

Differential privacy (Dwork and Roth 2014) is a privacy definition that guarantees the outcome of a calculation to be insensitive to any particular record in the data set. Informally, differential privacy requires the output of a data analysis mechanism to be approximately the same if any single tuple is replaced with a new one. In order to obtain this privacy guarantee, the algorithm employed to compute the result of the analysis must contain some form of randomness: in this way, the probability of obtaining a particular outcome r from dataset D is associated to any pair dataset-outcome (D, r). The intuition behind the definition of differential privacy is that, if the probability that outcome r comes from a particular dataset D is very close to the probability that the same outcome comes from any "similar" dataset D', then it is impossible to exactly identify dataset D just looking at the result r. To protect the identity of any single record, we consider as "similar" (or "adjacent") two datasets that differ for only one record. There are different interpretation in literature of the notion of adjacent datasets. Many papers consider two datasets D and D' adjacent if one of them contains all the records of the other plus a new record (Friedman and Schuster 2010; Su et al. 2017). Other papers consider D and D' adjacent if one can be obtained from the other replacing only one record (Dwork and Roth 2014; Chaudhuri et al. 2011). We opt for this last definition, so we will consider the number of record N fixed. More formally, we report the following definitions (Dwork and Roth 2014):

Definition 1 (Neighboring/adjacent datasets) Let D and D' be two datasets of the same data universe Ω, with N records. We say that D and D' are *neighboring* or

adjacent (in symbols, $D \sim D'$) if there exist two records d in D and d' in D' such that $D' = (D \setminus \{d\}) \cup \{d'\}$.

Definition 2 (ε-differential privacy) Let $\mathcal{M} : \Omega \longrightarrow \mathcal{R}$ be a randomized mechanism (i.e. a stochastic function with values in a generic set $\mathcal{R}$) and consider a real number $\varepsilon > 0$. We say that $\mathcal{M}$ preserves ε-differential privacy if $\forall D, D' \in \Omega$ such that $D \sim D'$ and $\forall r \in \mathcal{R}$

$$\frac{P(\mathcal{M}(D) = r)}{P(\mathcal{M}(D') = r)} \leq e^{\varepsilon}.$$

The parameter ε (also called *privacy budget*) allows us to control the level of privacy of the mechanism. Lower values of ε mean stronger privacy, as for ε near 0 we have $e^{\varepsilon} \approx 1$ and the probability that outcome r comes from dataset D or from dataset D' is almost the same.[2]

Differential privacy satisfies the following properties (Dwork and Roth 2014).

Theorem 1 *(Composition) Let $\mathcal{M}_1 : \Omega \longrightarrow \mathcal{R}$, $\mathcal{M}_2 : \Omega \longrightarrow \mathcal{S}$ be two randomized mechanism and let $g : \mathcal{R} \times \mathcal{S} \longrightarrow \mathcal{T}$ be a function. If $\mathcal{M}_1$ preserves ε_1-differential privacy and $\mathcal{M}_2$ preserves ε_2-differential privacy, then $g(\mathcal{M}_1, \mathcal{M}_2)$ preserves $(\varepsilon_1 + \varepsilon_2)$-differential privacy.*

Theorem 2 *(Post-processing) Let $\mathcal{M} : \Omega \longrightarrow \mathcal{R}$ be a randomized mechanism preserving ε-differential privacy and let f be any function with domain $\mathcal{R}$. Then $f \circ \mathcal{M}$ preserves ε-differential privacy.*

Theorem 1 states that by combining the results of several differentially private mechanisms, the outcome will be differentially private too, and the overall level ε of privacy guaranteed will be the sum of the level of privacy of each mechanism. In this sense, the ε parameter can be interpreted as the total privacy budget, and one can allocate part of it for any computation required to obtain the final outcome. On the other hand, Theorem 2 says that once a quantity r has been computed in a differentially private way, any following transformation of this quantity is still differentially private, with no need to spend part of the privacy budget for it. The two theorems together provide a useful and complete tool that allows one to modify an existing algorithm in order to make it differentially private: any time the algorithm needs to access the original data, some differentially private mechanism can be used, spending part of the overall privacy budget; all the other steps of the algorithm can be left unchanged.

Notice that in the definition of differential privacy there is no reference to the fact that a good mechanism needs to be accurate. Anyway, accuracy is an important property of any good differentially private mechanism: if the goal is to compute a differentially private query q over a dataset D, in addition to making the result private, the mechanisms should also render the same result "realistic", i.e. the result obtained through the application of a differentially private mechanism should be near to the actual result $q(D)$. A formal definition of the accuracy of a mechanism, inspired by Dwork and Roth (2014), can be the following:

[2] When ε is much less than 1, e^{ε} is approximately $1 + \varepsilon$. When $\varepsilon > 1$, e^{ε} grows very fast. For example, when $\varepsilon = 3$ e^{ε} is about 20, and when $\varepsilon = 5$ it is about 148.4.

Definition 3 (Accuracy) Let $q : \Omega \longrightarrow \mathcal{R}$ be a function and $\mathcal{M}$ a differentially private mechanism. $\mathcal{M}$ has accuracy $a \in \mathbb{R}$ with probability $1 - \delta$ if, for any D,

$$\mathcal{P}(d(\mathcal{M}(D), q(D)) > a) \leq \delta$$

where d is a distance defined on $\mathcal{R}$ and $\delta \in (0, 1)$.

Several mechanisms and techniques preserving differential privacy have been proposed in literature. Two of the most famous mechanisms, that we will use in the remainder of the paper, are the Laplace and the Exponential mechanisms (Dwork and Roth 2014; McSherry and Talwar 2007). The first can be applied to compute the result of a numeric function in a differentially private way; the second can be used to choose, within a given set, the element that maximizes a utility function whose result depends on some secret dataset D. Both these mechanism calibrate the amount of random noise they inject in the computation by looking at the sensitivity of the function (or utility function) considered.

Definition 4 (Global sensitivity) Let $q : \Omega \longrightarrow \mathbb{R}^d$ be a numeric function. The global sensitivity $GS(q)$ is a measure of the maximal variation of function q when computed over two adjacent datasets and is defined as

$$GS(q) = \max_{D \sim D'} ||q(D) - q(D')||_1.$$

Definition 5 (Laplace Mechanism) Let $q : \Omega \longrightarrow \mathbb{R}^d$ be a numeric function. The *Laplace mechanism* is $\mathcal{M}(D) = q(D) + (X_1, \ldots, X_d)$, where $X_1, \ldots, X_k$ are random variables extracted from a Laplace distribution with parameters $\left(0, \frac{GS(q)}{\varepsilon}\right)$, where $GS(q)$ is the global sensitivity of q.

Definition 6 (Exponential Mechanism) Let $q : \Omega \longrightarrow \mathcal{R}$ be the function that returns, among all possible values in $\mathcal{R}$, the one that maximizes some utility function $u : \Omega \times \mathcal{R} \longrightarrow \mathbb{R}$. The *Exponential mechanism* $M(D)$ returns a value of $\mathcal{R}$ with probability proportional to $exp\left(\frac{\varepsilon \cdot u(D,r)}{2GS(u)}\right)$, where $GS(u)$ is the global sensitivity of the utility function.

It can be proved that these mechanisms preserve ε-differential privacy (Dwork and Roth 2014). In both the mechanisms, the amount of noise introduced depends on the value of ε: there is a trade-off between the accuracy of the mechanisms and the level of privacy protection they guarantee. If a large value of ε is chosen, the mechanism will return a result that is close to the actual one with high probability. But, as ε gets smaller, the probability of adding a significant amount of noise to the result grows. How to choose a good value for ε is still an open issue. This is evident in the literature, where algorithms have been evaluated with ε ranging from as little as 0.01 to as much as 10 (see Table 1 of (Hsu et al. 2014)). Many academic works tend to prefer low values of ε (less than 1), probably because for small values of ε the quantity e^{ε} can be approximate to $1 + \varepsilon$, which makes it easier to understand the meaning of Definition 2.

In practical applications, however, higher values of ε are usually adopted (Domingo-Ferrer et al. 2021). For a recent discussion on the choice of ε, the reader can refer to Dwork et al. (2019).

2.2 DILCA and Private Categorical Distance Computation

Measuring similarities or distances between two data objects is a crucial step for many machine learning and data mining tasks. While the notion of similarity for continuous data is relatively well-understood and extensively studied, for categorical data the similarity computation is not straightforward. The simplest comparison measure for categorical data is *overlap* (Kasif et al. 1998): given two tuples, it counts the number of attributes whose values in the two tuples are the same. The overlap measure does not distinguish different values of attributes, hence matches and mismatches are treated equally. Boriah et al. (2008) present 14 different categorical measures using different heuristics to weight the mismatch of the values of the same attributes. Alamuri et al. (2014) survey the main approaches to distance computation for categorical data. Zhang et al. (2015) create the co-occurence graph of all the values of all the categorical attributes and then compute the shortest path distance between two values of the same attribute as a proximity measure.

Among all the proposed methods for distance computation, we focus on DILCA (Ienco et al. 2012), a framework to learn context-based distances between each pair of values of a categorical attribute Y. The main idea behind DILCA is that the distribution of the co-occurrences of the values of Y and the values of the other attributes in the dataset may help define a distance between the values of Y (intuitively, two values that are similarly co-distributed w.r.t. all the other values of all the other attributes are similar and so they should be close in the new distance). However, not all the other attributes in the dataset should be taken in consideration, but only those that are more relevant to Y. We call this set of relevant attributes with respect to Y the *context* of Y. The problem of identifying a set of attributes that are relevant (and not redundant) for a target attribute Y is a classic problem in data mining named supervised feature selection.

Let $\mathcal{F}$ be a set of m categorical attributes and let us consider a target attribute Y. DILCA computes the distances between the values of Y in two steps:

– *Context Selection:* it performs supervised feature selection in order to select an informative set of attributes with respect to target attribute Y. The correlation/association between two attributes X and Y is measured through the Symmetric Uncertainty (Yu and Liu 2003), an association based measure inspired by information theory and defined as follows:

$$SU(X, Y) = 2 \cdot \frac{I(X, Y)}{H(X) + H(Y)}$$

where $I(X, Y)$ is the Mutual Information between X and Y and $H(X)$, $H(Y)$ are the entropies. The Symmetric Uncertainty between two categorical attributes of a dataset is computed starting from their contingency table. Ienco et al.

(2012) propose two methods to select a good context for Y: the first, called $DILCA_M$, selects all the attributes X with $SU(X, Y)$ greater than the mean value $\frac{1}{m-1}\sum_{X\in\mathcal{F}\setminus Y} SU(X, Y)$; the second, called $DILCA_{RR}$, selects only those attributes that are *relevant* for Y but not *redundant*. In order to obtain this result, it employs a feature selection algorithm that requires the computation of $SU(X_i, X_j)$ for each pair of attributes $X_i, X_j \in \mathcal{F}$.

– *Distance Computation:* let $y_1, \ldots y_n$ be the values of attribute Y. For each pair y_i, y_j with $i, j = 1, \ldots, n$, the distance between y_i and y_j is computed as:

$$d(y_i, y_j) = \sqrt{\frac{\sum_{X\in context(Y)} \sum_{k=1}^{|X|} (P(y_i|x_k) - P(y_j|x_k))^2}{\sum_{X\in context(Y)} |X|}} \tag{1}$$

where $context(Y)$ is the set of the attributes selected in the previous step, $|X|$ is the number of values attribute X can assume, and $P(y_i|x_k)$ is the conditional probability that Y takes value y_i given that X has value x_k. The conditional probabilities $P(y_i|x_k)$ are estimated from the data: the contingency table between attributes X and Y is constructed and this contingency table can be interpreted as the empirical joint distribution of the two variables.

The distance measure computed by DILCA is a metric, since it is an application of the Euclidean distance. Furthermore, $0 \leq d(y_i, y_j) \leq 1$ for each pair y_i, y_j.

To the best of our knowledge, no differentially private methods for categorical distance learning from data have been proposed so far. However, there are recent solutions to the problem of standard distance computation in a differentially private fashion. Stanojevic et al. (2017), for instance, propose a way to estimate the cardinality of the intersection and the union of two sets, when the sets are represented by two bit vectors previously obfuscated with the randomized response mechanism. This technique can be used to estimate the pairwise Jaccard similarity matrix between the objects of an obfuscated dataset with binary attributes. Similarly, Aumüller et al. (2020) present a method to privately release two sets, in a way that preserves the Jaccard similarity between them. It consists in the private publication of a vector representation of each set, obtained through the application of a fixed number of MinHash functions. Xu et al. (2017), instead, present an algorithm for the differentially private release of high-dimensional data, designed to preserve pairwise L2-distances between records. Although all these techniques can be used to estimate the pairwise similarity between binary tuples (as done by Gao et al. (2020) in the context of recommender systems), these methods are substantially different from the one we propose in this paper. They can be used to compute similarities/distances among records of a secret dataset, while, in this paper, we propose a family of algorithms that privately learn distances among values of a categorial attribute from a secret dataset. The repeated application of our technique on all the categorical attributes describing the data leads to the learning of a metric on the data space. In other words, differential privacy in our work is used to disclose something about the space, without revealing the presence of a particular record in the secret training set, while, in the other methods, it is used to disclose something about the secret data, but the results cannot be generalized to

learn the distance between two generic object from the same data space. Finally, all other methods accounts for value matches and mismatches in the same way, which is exactly what we want to exceed by computing a more expressive metric showing its effectiveness with multivariate categorical attributes as well.

3 DP-DILCA

In this section, we introduce our family of methods whose final goal is to inject some form of randomness in DILCA in order to make the resulting distances among the values of the target attribute Y differentially private. All along this section, we will consider the following illustrative running example in order to show how the proposed methods work.

Example 2 (running example) Let us consider a set of five categorical attributes describing some people living in some territory: 'city', 'has_car','sex', 'income_level' and 'wealth_level'. Both the last two attributes can take three ordinal values (0,1 and 2), and they are strongly correlated. There is a trusted curator that owns different snapshots of the data and wants to publish the distance matrix M among the cities, without releasing the original datasets. Suppose now that the curator owns two secret adjacent snapshots D and D' (according to Definition 1). Figure 1(a) and 1(b) show the contingency tables between the target attribute 'city' and all the other attributes, as well as the matrices of the distances among the values of attribute 'city' computed by DILCA on datasets D and D' respectively. Although the distance matrices of the values of attribute 'city' do not dramatically change when computed starting from dataset D or D', the difference is enough to allow a malicious adversary to understand whether the secret dataset is D or D' and thus whether the atypical record of a person living in city A and having 'income_level' = 2 is present or less in the data. The problem, in terms of privacy, is that the algorithm used to compute the distances is deterministic, so an adversary undecided on which is the true dataset between D and D' is able to identify the correct dataset by simply running the algorithm on both the datasets. Suppose now that the curator uses a differentially private algorithm to compute the distances among the cities, adopting as training set dataset D: even if the results obtained are (hopefully) similar to the actual distance matrix in Figure 1(a), the adversary cannot say whether the fact that A and B are equally far from C depends on the fact that the dataset is D or on the noise added by the algorithm.

3.1 Differentially private distance computation

A naive way to modify DILCA and make it private is to act on the contingency tables computation stage of the algorithm (see Section 2.2), by investigating a way to create all the needed contingency tables privately. Since all the computations following this step only look at the contingency tables (and do not access the original data matrix anymore), the post-processing theorem (reported in Section 2.1) guarantees that, once the contingency tables are computed in a differentially private way, the final result will be differentially private as well.

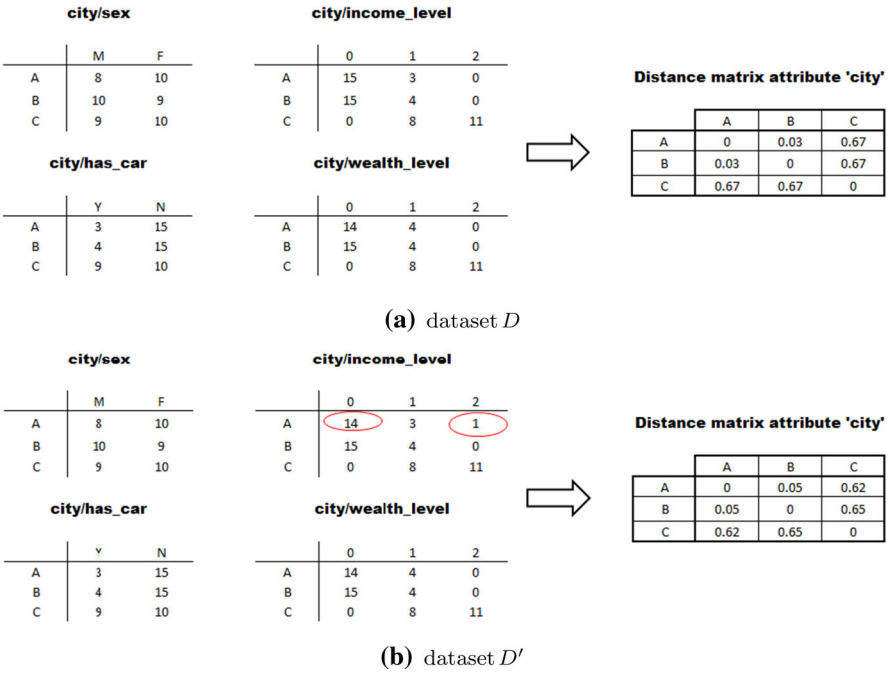


Fig. 1 Data used in Example 2. Datasets D (a) and D' (b) are two adjacent snapshots owned by a trusted curator. We report the contingency tables between attribute 'city' and all other attributes computed from datasets D and D', together with the matrix of the distances among the values of 'city'. The red ovals indicate the values that change because of the substitution of one record from D to D'

Algorithm 1: $BaselineDP-DILCA(D, Y, method, \varepsilon)$

Input: The original dataset D with attributes $F = \{X_1, ..., X_m\}$, the target attribute $Y \in F$, the privacy budget ε

Result: The distance matrix $distMatrix(Y)$

1 $\varepsilon \leftarrow \frac{\varepsilon}{\binom{m}{2}}$;

2 **for** $X, X' \in \{X_1, \ldots, X_m\}$ **do**

3 compute $ContTable(X, X')$;

4 $ContTable(X, X') \leftarrow ContTable(X, X') + Lap(0, 2/\varepsilon)^{|X|\times|X'|}$;

5 **end**

6 Compute $Context(Y)$ using the selected *method* ;

7 Compute $distMatrix(Y)$ using equation (1);

Algorithm 1 gives a sketch of this first implementation of DP-DILCA. The only variation from the original algorithm is that the contingency tables are computed using the Laplace Mechanism, at steps 3-4.

Theorem 3 *Given $\varepsilon > 0$, Algorithm 1 satisfies ε-differential privacy.*

Proof The algorithm computes the contingency tables between each pair of attributes, including the target attribute Y. The total number of pairs of m objects is $\binom{m}{2}$, thus the privacy budget spent for each table should be $\varepsilon' = \frac{\varepsilon}{\binom{m}{2}}$. Consider now the function that

Algorithm 2: $DP - DILCA(D, Y, method, \varepsilon, h)$

Input: The original dataset D with attributes $F = \{X_1, ..., X_m\}$, the target attribute $Y \in F$, the context selection method, the privacy budget ε, the split parameter $h \in (0, 1)$
Result: The distance matrix $distMatrix(Y)$
1 compute $context(Y)$ with the selected method and privacy budget $\varepsilon \cdot h$;
2 **for** $X_i \in context(Y)$ **do**
3 compute $ContTable(Y, X_i)$;
4 $ContTable(Y, X_i) \leftarrow ContTable(Y, X_i) + Lap\left(0, \frac{2 \cdot |context(Y)|}{\varepsilon \cdot (1-h)}\right)^{|Y| \times |X_i|}$;
5 **end**
6 Compute $distMatrix(Y)$ using DILCA;

computes the contingency table between two attributes X and Y. Let t_{ij} be the ij-th entry of the contingency table, corresponding to the number of records in the original dataset having $X = x_i$ and $Y = y_j$. If we change one record of the dataset, having $X = x_i$ and $Y = y_j$, with a new record having $X = x_h$ and $Y = y_k$, only two entries of the contingency table will change: t_{ij} decreases of one unit, while t_{hk} increases of one unit. Thus the global sensitivity of the function that computes the contingency table is 2. We can apply the Laplace mechanism by adding random noise extracted from a Laplace distribution with parameters $(0, \frac{2}{\varepsilon'})$ to each cell of the actual contingency matrix between X and Y, and the obtained contingency table will be ε'-differentially private. □

Although the naive method illustrated in Algorithm 1 respects differential privacy, it may be inaccurate, because it requires the computation of an high number of contingency tables ($\binom{m}{2}$, where m is the number of attributes in the dataset) and, consequently, the injection of a great amount of distortion. This is particularly true for datasets with a large number of attributes m. In Section 4, we will use this first method as a baseline.

An alternative option is to distort only the contingency tables between target attribute Y and the attributes $X \neq Y$ in the context of Y. In fact, in the computation of the distance matrix among the values of Y only those contingency tables are used. If the context of Y contains only few attributes with respect to the total number of remaining $m - 1$ attributes, the amount of noise introduced in the computation will be significantly less. However, the selection of a good context for Y is a sensitive function, because it looks at the original dataset to decide which attributes are more relevant for Y. Hence, it should be performed in a differentially private way and a fraction $h \in (0, 1)$ of the privacy budget should be devoted to it. The sketch of this new procedure is given in Algorithm 2, while private context computation is discussed in Section 3.2.

Theorem 4 *Given $\varepsilon > 0$ and $h \in (0, 1)$, if $context(Y)$ is computed in a differentially private way, then Algorithm 2 preserves ε-differential privacy.*

Proof By hypothesis, the computation of $context(Y)$ preserves $(\varepsilon \cdot h)$-differential privacy. In step 4 of the algorithm we apply the Laplace Mechanism to the computation of the contingency tables, with parameter of the Laplace distribution equal to $\frac{2 \cdot |context(Y)|}{\varepsilon \cdot (h-1)}$. As noticed before, the global sensitivity of the function that computes

the contingency matrix between two variables is 2. Thus the Laplace mechanism preserves $\frac{\varepsilon \cdot (1-h)}{|context(Y)|}$-differential privacy. The computation is repeated $|context(Y)|$ times. Finally, the procedure that computes the distance matrix does not access the original dataset anymore, so, according to Theorem 2, it does not require further privacy budget. The overall algorithm is then $\varepsilon \cdot h + |context(Y)| \cdot \frac{\varepsilon \cdot (1-h)}{|context(Y)|} = \varepsilon$-differentially private. □

To clarify the difference between the two methods proposed in Algorithm 1 and Algorithm 2, let us refer to Example 2. There are five attributes in dataset D (Figure 1(a)), then Baseline DP-DILCA needs to compute $\binom{5}{2} = 10$ distorted contingency matrices, using a privacy budget $\frac{\varepsilon}{10}$ for each one. Instead, DP-DILCA devotes $h \cdot \varepsilon$ privacy budget to the computation of the context and then $\frac{(1-h)\cdot\varepsilon}{k}$ privacy budget for the computation of each contingency table, where k is the number of attributes in the context of the target attribute. By looking at the contingency tables between the target attribute 'city' and all the other attributes in Figure 1(a), we can see that attributes 'sex' and 'has_car' are not useful to discriminate between the different values of attribute 'city', while attributes 'income_level' and 'wealth_level' are more informative. However, the co-distributions of these two attributes w.r.t. 'city' are very similar and we can conclude that the presence of both the attributes in the context of 'city' would be redundant: a suitable context for attribute 'city' could be {'income_level'} (this is exactly the context identified by $DILCA_{RR}$). To quantify the difference between the privacy budget used by Baseline DP-DILCA and DP-DILCA, let us set the overall privacy budget equal to 1 and the parameter h equal to 0.3. The privacy budget spent by the two algorithms to the computation of each contingency table involved in the final computation of the distances is 0.1 and 0.7 respectively.

3.2 Differentially private context selection

The context selection procedure used by DILCA is an application of a filter method for supervised feature selection. Indeed, some work has been done on differentially private feature selection. For instance, Yang and Li (2014) and Li et al. (2016) present two alternative differentially private implementations of a feature selection method that preserves nearest-neighbor classification capability. They differ for the step of the algorithm where they apply the randomized mechanism: Yang and Li (2014) adopt output perturbation, while Li et al. (2016) perturb the objective function. However, both these methods are designed for continuous data. Anandan and Clifton (2018) study the sensitivity of several association measures used for feature selection (such as Chi-Squared Statistic, Bray-Curtis dissimilarity, Information Gain) and integrate the noised version of these measures in two differentially private classifiers.

In this section, we propose three different methods to perform differentially private context selection. The first method is a differentially private version of $DILCA_M$, obtained through the multiple application of the Laplace mechanism; the second and third ones use the exponential mechanism to extract an highly informative subset of attributes with respect to the target attribute Y. The last two methods differ in the definition of optimal context they consider.

3.2.1 Differentially private *DILCA*$_M$

In its original formulation, DILCA uses the Symmetric Uncertainty $SU(X, Y)$ as a measure of the association of two attributes X and Y in order to decide whether X should be in the context of Y or not. It would be convenient to compute $SU(X, Y)$ privately, for instance with the Laplace mechanism. The main building block for the application of such a mechanism would be the estimation of an upper bound of the global sensitivity of $SU(X, Y)$: unfortunately, it is not easy to analytically compute the variation of SU when changing a record in the original dataset. However, using a well known property of the Mutual Information (Cover and Thomas 2001), it can be noticed that

$$SU(X, Y) = 2 \cdot \frac{I(X, Y)}{H(X) + H(Y)} = 2 \cdot \frac{H(X) + H(Y) - H(X, Y)}{H(X) + H(Y)}$$

and, thanks to Theorem 1, a distorted version $\widetilde{SU}(X, Y)$ of the desired quantity can be obtained as the composition of distorted entropies, computed through the Laplace mechanism:

$$\widetilde{SU}(X, Y) = 2 \cdot \frac{\widetilde{H}(X) + \widetilde{H}(Y) - \widetilde{H}(X, Y)}{\widetilde{H}(X) + \widetilde{H}(Y)},$$

where $\widetilde{H}(\cdot) = H(\cdot) + Lap\left(0, \frac{GS(H)}{\varepsilon}\right)$. The following theorem gives an upper bound of the global sensitivity of the entropy $GS(H)$.

Theorem 5 *(Sensitivity of entropy) Let D be a dataset with N records and a categorical attribute having values $\{x_1, \ldots, x_k\}$. Let n_i be the number of records of D having value x_i and let X be a random variable with probability distribution $p(X = x_i) = \frac{n_i}{N}$. The global sensitivity of $H(X)$ is lower than $\frac{1}{N}\left(\frac{1}{ln(2)} + log(N)\right)$.*

Proof Let us expand the formula of entropy $H(X)$:

$$\begin{aligned} H(X) &= -\sum_{i=1}^{k} p(x_i) log(p(x_i)) = -\sum_{i=1}^{k} \frac{n_i}{N} log\left(\frac{n_i}{N}\right) \\ &= -\frac{1}{N}\sum_{i=1}^{k} n_i (log(n_i) - log(N)) = -\frac{1}{N}\sum_{i=1}^{k} n_i \cdot log(n_i) + log(N) \end{aligned}$$

Suppose we change a record of D having value x_a with another having value x_b. Only two counts will change: the number of records with value x_a will become $n_a - 1$ and, similarly, the number of records with value x_b will become $n_b + 1$. All the other counts n_i will remain untouched. Let X' be the random variable associated to the new

probability distribution. Thus we have:

$$
\begin{aligned}
&|H(X) - H(X')| \\
&= \frac{1}{N} |-n_a log(n_a) - n_b log(n_b) + (n_a - 1) log(n_a - 1) + (n_b + 1) log(n_b + 1)| \\
&= \frac{1}{N} \left| (n_a - 1) log \left(\frac{n_a}{n_a - 1} \right) + (n_b + 1) log \left(\frac{n_b + 1}{n_b} \right) \right. \\
&\quad \left. - log(n_a - 1) + log(n_b + 1) \right| \\
&\leq \frac{1}{N} \left| (n_a - 1) log \left(\frac{(n_a - 1) + 1}{n_a - 1} \right) - (n_b + 1) log \left(\frac{n_b + 1}{n_b} \right) \right| \\
&\quad + \frac{1}{N} |log(n_a - 1) - log(n_b + 1)|
\end{aligned}
$$

We recall that all the logarithms $log()$ are in base 2 and that, for $a > 0$,

$$
\left| a \cdot log \left(\frac{a + 1}{a} \right) \right| \leq \frac{1}{ln(2)}. \tag{2}
$$

Moreover, $n_a - 1$ and $n_b + 1$ are both between 0 and N, then

$$
|H(X) - H(X')| \leq \frac{1}{N} \left(\frac{1}{ln(2)} + log(N) \right).
$$

□

Once we are able to compute $SU(Y, X)$ for each attribute, we can apply the same selection method of $DILCA_M$ using the distorted values of the Symmetric Uncertainties instead of the actual ones. Algorithm 3 summarizes the related procedure: it computes the distorted values of $SU(Y, X_i)$ and then selects, in the context of Y, all those attributes that have Symmetric Uncertainty with Y greater than the mean value $M = \frac{1}{m-1} \sum_{X_i \neq Y} (SU(Y, X_i))$. The algorithm needs the computation of $m - 1$ different values of Symmetric Uncertainty. Furthermore, each $SU(Y, X)$ is computed as the composition of three entropy functions, $H(X)$, $H(Y)$ and $H(X, Y)$. Thus, the total number of entropies to be computed for the context selection is $2 \cdot m - 1$ ($m - 1$ different $H(X)$, $m - 1$ different $H(X, Y)$ and one $H(Y)$): since at steps 7-8 each entropy is obtained through the Laplace mechanism with privacy budget $\frac{\varepsilon}{(2 \cdot m - 1)}$, Algorithm 3 preserves ε-differential privacy.

We conclude this section by showing how the context selection strategy just presented works on the toy dataset D introduced in Example 2. The value of the Symmetric Uncertainty between the target attribute 'city' and the other attributes 'sex', 'has_car','income_level' and 'wealth_level' are, respectively, 0.001, 0.035, 0.217 and 0.204. Since the last two attributes have Symmetric Uncertainty greater than the mean value 0.11, they are both selected in the context of 'city' by algorithm $DILCA_M$ (note that, differently from $DILCA_{RR}$, $DILCA_M$ is not able to discard the redundant attribute 'wealth_level'). When, instead of $DILCA_M$, its differentially private variant DP-MeanSU is applied, a certain amount of noise is added to the computation

Algorithm 3: $DP - MeanSU(D, Y, \varepsilon)$

Input: The original dataset D with N records and attributes $F = \{X_1, ..., X_m\}$, the target attribute $Y \in F$, the privacy budget ε
Result: The set $context(Y)$

$gs \leftarrow \frac{1}{N}\left(\frac{1}{ln(2)} + log(N)\right)$;
$\varepsilon_H \leftarrow \frac{\varepsilon}{2 \cdot m - 1}$;
Compute $H(Y)$;
$H(Y) \leftarrow H(Y) + Lap\left(0, \frac{gs}{\varepsilon_H}\right)$;
for $X \in \{X_1, \ldots, X_m\}$, $X \neq Y$: **do**
    Compute $H(X)$ and $H(X, Y)$;
    $H(X) \leftarrow H(X) + Lap\left(0, \frac{gs}{\varepsilon_H}\right)$;
    $H(X, Y) \leftarrow H(X, Y) + Lap\left(0, \frac{gs}{\varepsilon_H}\right)$;
    $SU(Y, X) \leftarrow 2\frac{H(X)+HX(Y)-H(X,Y)}{H(X)+H(Y)}$;
end
$M \leftarrow \frac{1}{m-1}\sum_{X_i \neq Y}(SU(Y, X_i))$;
$context(Y) \leftarrow \{X_i | SU(Y, X_i) \geq M\}$

of the Symmetric Uncertainties: for instance, we could obtain 0.12, 0.0, 0.26 and 0.35 . Although the values of SU are rather different from the correct values, again the only two attributes having Symmetric Uncertainty greater then the mean value are the last two, so in this example the DP-MeanSU selects the same context of $DILCA_M$. The more the values of Symmetric Uncertainties are far from one another, the higher the probability that DP-MeanSU extracts the correct context.

3.2.2 Differentially private Maximum Relevance

The main drawback of the selection method illustrated in Algorithm 3 is the wasteful use of the privacy budget when the Symmetrical Uncertainty should be evaluated separately for each attribute X_i. The exponential mechanism offers a better approach: rather than evaluating each attribute separately, we can evaluate all the attributes simultaneously in one query whose outcome is the attribute X_i that maximizes some utility function. If this utility function measures the relevance of X_i for target attribute Y, the exponential mechanism will return (with high probability) one attribute that is very relevant for Y. Repeating the procedure k times, we will obtain a set of k attributes that, with high probability, are the k most relevant ones for target attribute Y: for this reason, following Peng et al. (2005), we refer to this method as *MaxRelevance*.

If we want to remain stick to DILCA's strategy, the utility function used to measure the relevance of attribute X for target attribute Y should be the Symmetric Uncertainty. Unfortunately, as pointed out before, we are not able to compute the sensitivity of SU and so we cannot apply the exponential mechanism to this utility function. Thus, we propose a differentially private selection method that measures the connection of two attributes by looking at the (distorted) Mutual Information between them and then extracts the k most relevant attributes. Mutual Information is a widely used measure of association in the supervised feature selection problem (see, for instance, Peng et al.

(2005). As already pointed out, it can be computed as $I(X, Y) = H(X) + H(Y) - H(X, Y)$. Thus, finding the X which maximizes $I(X, Y)$ is equivalent to finding the X which maximizes

$$I'(X, Y) = H(X) - H(X, Y). \tag{3}$$

Theorem 6 *(Sensitivity of $I'(X, Y)$) Given a dataset D with N records and two attributes X and Y, an upper bound of the sensitivity of $I'(X, Y)$ is*

$$\frac{2}{N}\left(\frac{1}{ln(2)} + log(N)\right).$$

Proof We know from Theorem 5 that the sensitivity of function $H(\cdot)$ is $\frac{1}{N}\left(\frac{1}{ln(2)} + log(N)\right)$. Let X' and Y' be the variables obtained by changing one record of the original dataset D. Then

$$\begin{aligned}|I'(X, Y) - I'(X', Y')| &= |H(X) - H(X, Y) - H(X') + H(X', Y')| \\ &\leq |H(X) - H(X')| + |H(X, Y) - H(X', Y')| \leq \frac{2}{N}\left(\frac{1}{ln(2)} + log(N)\right).\end{aligned}$$

□

Algorithm 4 describes the differentially private implementation of *MaxRelevance* for context selection. It requires the specification, as input parameter, of the desired number k of attributes in the context of the target attribute. When setting the value of parameter k, one must consider that lower values of k are preferable, from a differentially private point of view. In step 5 of Algorithm 4, the exponential mechanism is applied k times, in order to extract the top k attributes: each application of the exponential mechanism requires part of the overall privacy budget; thus, the smaller k is, the higher the accuracy of the selected context. Furthermore, Algorithm 2 computes and perturbs k contingency tables: again, lower valus of k mean less noise injected in the computation of the final distance matrix.

Consider, once again, the situation described in Example 2. The correct values of $I'(X, Y)$ for dataset D, where Y is the target attribute 'city' and X are attributes 'sex', 'has_car','income_level' and 'wealth_level', once at time, are -1.58, -1.49, -0.93 and -0.96 respectively. According to MaxRelevance, the context of 'city' is {'income_level'} when $k = 1$ and {'income_level', 'wealth_level'} when $k = 2$, because only the k attributes with highest $I'(X, Y)$ are selected. Instead, according to DP-MaxRelevance, the k attributes to be inserted in the context are selected with probability proportional to $\frac{\varepsilon \cdot I'(X,Y)}{2 \cdot GS(I')}$. For instance, when $\varepsilon = 1$, the probability of selecting attributes 'sex', 'has_car', 'income_level' and 'wealth_level' are 0.08, 0.1, 0.43 and 0.39 respectively. Thus, when $k = 1$, we will obtain a context containing 'income_level' or 'wealth_level' with high probability. Attributes 'sex' or 'has_car', instead, are not associated with 'city' at all: hence, they have a very low probability to be extracted.

Algorithm 4: $DP - MaxRelevance(D, Y, \varepsilon, k)$

Input: The original dataset D with N records and attributes $F = \{X_1, ..., X_m\}$, the target attribute $Y \in F$, the privacy budget ε, the number k of attributes in the context

Result: The set $context(Y)$

$gs \leftarrow \frac{2}{N}\left(\frac{1}{ln(2)} + log(N)\right)$;

$\mathcal{F} \leftarrow \{X_1, \ldots, X_m\} \setminus \{Y\}$;

$context(Y) \leftarrow \emptyset$;

for $t = 1$ *to* k **do**

  Select an object $X \in \mathcal{F}$ with probability proportional to $exp\left(\frac{\varepsilon \cdot MI(Y,X)}{2 \cdot k \cdot gs}\right)$;

  $context(Y) \leftarrow context(Y) \cup \{X\}$;

  $\mathcal{F} \leftarrow \mathcal{F} \setminus \{X\}$;

end

3.2.3 Differentially private Maximum Dependency

Both previous selection context methods insert the most relevant attributes into the context of the target attribute Y by evaluating the association of each attribute X_i with the target attribute Y individually. However, they may select two or more attributes giving the "same information" about Y. This happens, for instance, when two attributes X_i and X_j in $context(Y)$ are highly correlated, thus they give the same description of Y. On the other hand, an attribute X that is less associated individually with Y and for this reason is not included in $context(Y)$, could add a piece of information about Y that is not captured by any attribute X_i in the context of Y. In this sense, a preferable context selection method is one that looks for the set of attributes that globally has the maximal association with target attribute Y. We can do this by choosing the subset $S \subset \{X_1, \ldots, X_m\} \setminus \{Y\}$ of cardinality k that maximizes the mutual information between Y and the set S. Let us assume, for simplicity, that $S = \{X_1, \ldots, X_k\}$, the mutual information between Y and S can be written as

$$I(Y, S) = H(Y) + H(X_1, \ldots, X_k) - H(Y, X_1, \ldots, X_k). \tag{4}$$

Then, maximizing $I(Y, S)$ is equivalent to maximizing

$$I'(Y, S) = H(X_1, \ldots, X_k) - H(Y, X_1, \ldots, X_k). \tag{5}$$

Peng et al. (2005) note that this feature selection scheme, called *MaxDependency*, is hard to implement, unless for low values of k, because of two issues in the high dimensional space: 1) the number of samples is often insufficient and 2) the slow computational speed. In facts, the number of joint states of k categorical variables increases very quickly with k and gets comparable to the number of records N. When this happens, the joint probabilities of this attributes cannot be estimated correctly from the data. However, the *MaxDependency* scheme can be very useful to select a small number of attributes when N is high. This is exactly the scenario in which we are working: as said before, we want to keep the number of attributes in $context(Y)$ low; furthermore, differentially private algorithms usually work better when the number of

Algorithm 5: $DP - MaxDependency(D, Y, \varepsilon, k)$

Input: The original dataset D with N records and attributes $F = \{X_1, ..., X_m\}$, the target attribute $Y \in F$, the privacy budget ε, the number k of attributes in the context

Result: The set $context(Y)$

$gs \leftarrow \frac{2}{N}\left(\frac{1}{ln(2)} + log(N)\right)$;

$\mathcal{F} \leftarrow \{S \subset \{X_1, \ldots, X_m\} \setminus \{Y\} s.t. |S| = k\}$;

Select a subset $S \in \mathcal{F}$ with probability proportional to $exp\left(\frac{\varepsilon \cdot I(Y,S)}{2 \cdot gs}\right)$;

$context(Y) \leftarrow S$

samples N in a dataset is large as masking the presence of a particular record is easier when there are a lot of other variegated samples (and the significance of the statistical analysis performed is higher).

A differentially private context selection method based on the Maximum Dependency criterion has another advantage: we can apply the exponential mechanism to the function that, among all possible subsets of $\{X_1, \ldots, X_m\}$ of cardinality k, gives as outcome the subset that maximizes Equation 5. In this way, we are applying the exponential mechanism only once (instead of k times as in Algorithm 4) and we can use all the privacy budget for this unique application.

The differentially private *MaxDependency* context selection method is illustrated in Algorithm 5. It consists in the application of the exponential mechanism to the function that, among all subset S of cardinality k, extracts the one that maximizes $I'(Y, S)$. The global sensitivity of the utility function $I'(Y, S)$ is given by the following theorem.

Theorem 7 *(Sensitivity of $I'(Y, S)$) Given a dataset D with N records, a categorical attribute Y and a set of categorical attributes S, an upper bound of the sensitivity of $I'(Y, S)$ is*

$$\frac{2}{N}\left(\frac{1}{ln(2)} + log(N)\right).$$

The proof is analogous to that of Theorem 6.

Let us refer to Example 2 for the last time. Let suppose that we want to use the selection strategy just described to select the context of attribute 'city', with $k = 2$. There are six possible contexts: {'sex', 'income_level'}, {'sex', 'wealth_level'}, {'sex','has_car'}, {'income_level', 'wealth_level'}, {'income_level', 'has_car'} and {'wealth_level', 'has_car'}. The context that maximizes the value of the objective function I' is {'income_level', 'has_car'}: differently from the previous selection strategies, MaxDependency does not select the two attributes that, individually, are more associated with the target attribute but prefers a set of attributes that are not redundant. The private algorithm computes then the probabilities associated to the possible contexts, which are, respectively, 2×10^{-11}, 4×10^{-14}, 2×10^{-45}, 5×10^{-12}, 0.89 and 0.11. It means that with probability 0.89 the algorithm will return the correct context {'income_level', 'has_car'}, while with probability that is about 1 it will return one of the two similar contexts {'income_level', 'has_car'} or {'wealth_level', 'has_car'}.

Table 1 Dataset characteristics

Dataset	# Instances	# Attributes	# Values	# Classes
Dermatology	366	34	131	6
Soybean	683	35	100	19
Cmc	1473	9	32	3
Mushroom	8124	22	117	2
NLTCS	21574	16	32	-
IPUMS-BR	37334	13	84	-
IPUMS-ME	42301	13	85	-
Adult	48842	14	120	2

4 Experiments

In this section, we describe the experiments conducted to evaluate the performance of our differentially private distance learning approach. For this evaluation, we use eight real-world datasets. Five (dermatology, soybean, mushroom, adult and cmc) are well known benchmark datasets available at the UCI Machine Learning Repository[3]. NLTCS[4] contains records of individuals participating in the National Long Term Care Survey. IPUMS-BR and IPUMS-ME[5] contain census records collected, respectively, from Brazil and Mexico in 2000. The characteristics of the datasets are summarized in Table 1. Some datasets contain numerical attributes: we discretize these attributes into five bins using k-means discretization.

4.1 Assessment of context selection

In the first experiment, we run all the variants of DP-DILCA on the real-world datasets in order to assess the quality of the context they select. For each dataset, we consider one attribute at a time as target attribute and we compute its differentially private context for increasing levels of privacy budget ε. Then we compare the context selected by DP-DILCA with the context obtained with the corresponding non-private method: *Baseline-DP-DILCA* (Algorithm 1), where the context selection strategy used is the same of $DILCA_{RR}$ (see Section refsubsec:DILCA), is compared with that of $DILCA_{RR}$, *DP-MeanSU* (Algorithm 3) with $DILCA_M$, *DP-MaxRelevance* (Algorithm 4) with MaxRelevance and *DP-MaxDependency* (Algorithm 5) with MaxDependency. MaxRelevance and MaxDependency require the specification of the number k of desired attributes in the context as input parameter. In all the experiments we set $k = 3$.

To evaluate the similarity between the private and non-private context for each target attribute, we use three popular measures in Information Retrieval: *recall*, *precision* and *F-score*. Called C and C_{DP} respectively the context selected by a non-private algo-

[3] https://archive.ics.uci.edu/

[4] http://lib.stat.cmu.edu/

[5] https://international.ipums.org/

rithm and the context selected by the correspondent differentially private algorithm, the recall is computed as $recall(C, C_{DP}) = |C \cap C_{DP}|/|C|$, while the precision is $precision(C, C_{DP}) = |C \cap C_{DP}|/|C_{DP}|$. The F-score is the harmonic mean of precision and recall:

$$F(C, C_{DP}) = 2 \cdot \frac{precision(C, C_{DP}) * recall(C, C_{DP})}{precision(C, C_{DP}) + recall(C, C_{DP})}.$$

For each ε, we repeat the experiments 30 times and we compute the mean value of all scores. In the experiments, we expect to see results that are at least equal to those we would obtain for random context selection. The expected value of the F-score when the context is selected uniformly at random depends on the number m of attributes in the dataset and on the number of possible contexts (for DP-MaxRelevance and DP-MaxDependency only the contexts containing $k = 3$ attributes are possible outcomes, while DP-MeanSU could select a context of any size). Thus, a comparison of the mean F-scores among different datasets and different context selection methods would not be fair. For this reason, similarly as done by Hubert and Arabie (1985), we adjust the mean F-score by computing $\frac{mean(F)-E}{1-E}$, where E is the expected value of the random selection context and is different for any dataset and context selection method. Figure 2 shows the results of our comparison: for each ε we report the average value of the normalized F-score over all the attributes of each dataset. In all the datasets, the results achieved by DP-MeanSU, DP-MaxRelevance and DP-MaxDependency increase with respect to ε and, especially for high levels of ε, they outperform the results of the baseline method. This is in line with what we expected, because the amount of noise introduced with the baseline method is much higher than the amount of noise introduced with the other three approaches. The shape of the curve and the level of accuracy reached by each context selection approach heavily depend on the data: one can notice that the best scores are reached in the datasets with more records. This is not surprising: the effort needed to mask the presence of a particular record is higher when the original dataset contains only few records. More formally, the amount of noise introduced in the context selection is proportional to the sensitivity of the entropy $\frac{1}{N}\left(\frac{1}{ln(2)} + log(N)\right)$: this quantity decreases as the dataset size N increases, thus it is reasonable to expect higher accuracy level for bigger datasets, at the same level of ε. In general, DP-MaxRelevance and DP-MaxDependency show better results than DP-MeanSU. In the smallest datasets (dermatology and soybean), the context selection procedure is more unstable. In particular, for DP-MeanSU no relevant growth in the value of the F-score wrt ε can be appreciated and the results are similar to those one would obtain for random context selection. In Appendix A.1, we also investigate the behavior of the context selection approach in controlled scenarios with synthetic data.

4.2 Assessment of the distance matrices

In this section we repeat the same experiments on the real-world data presented in Section 4.1, but we focus on the final output of DP-DILCA: the distances between the values of the target attribute. As before, for each dataset we consider one attribute at a

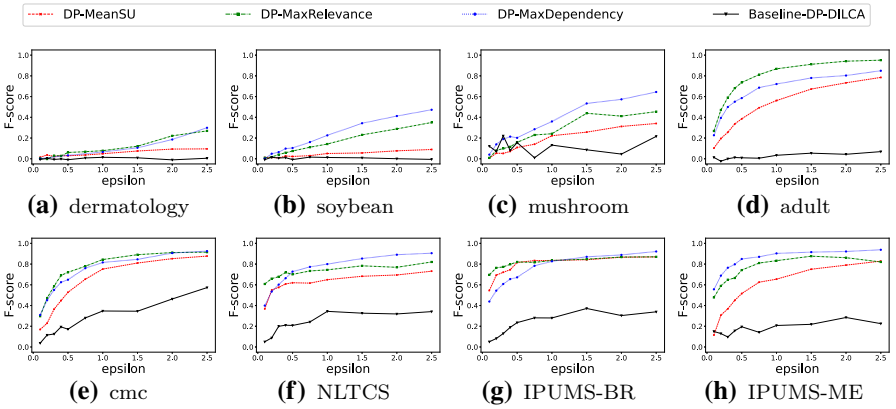


Fig. 2 Average adjusted F-score of the differentially private context

time as target and we compute the differentially private distance matrix associated to its values, for increasing levels of privacy budget ε. Then we compare the distances obtained with DP-DILCA with those obtained with the corresponding non-private method. Again, we set the parameter k equal to 3. In this experiment we need to set another parameter: the portion h of privacy budget we want to allocate to the context computation. We set $h = 0.3$: in this way we are giving more importance to the final step (the distance computation step) then to the context selection phase.

We quantify the linear correlation between the private distance matrix M', with shape $n \times n$, and its non-private counterpart M through the sample Pearson's correlation coefficient[6], defined as

$$\rho(M, M') = \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} (M_{ij} - \bar{M}) - (M'_{ij} - \bar{M}')}{\sqrt{\sum_{i=1}^{n} \sum_{j=i+1}^{n} (M_{ij} - \bar{M})^2} \sqrt{\sum_{i=1}^{n} \sum_{j=i+1}^{n} (M'_{ij} - \bar{M}')^2}}$$

where $\bar{M}$ and $\bar{M}'$ are the mean values of matrices M and M' respectively. The ρ coefficient takes values between -1 (perfect negative correlation) and 1 (perfect positive correlation). If the two matrices are not correlated we will have $\rho \cong 0$.

For each ε, we repeat the experiments 30 times and we compute the mean value of the sample Pearson correlation coefficient. Figure 3 shows the results of our computations: for each ε we report the average value of the measure over all the attributes of each dataset. Notice that the Pearson coefficient is always 1 when the target attribute has only two values. Considering these attributes in the computation would distort the resulting average Pearson coefficient, particularly favoring those datasets with many binary attributes. For this reason we exclude from the computation of the average Pearson coefficient the binary attributes. For the same reason, for NLTCS (Figure 3(f)), which consists of binary attributes only, the Person's correlation is always maximum.

The results show that there is positive correlation between private and non-private distances. The Pearson coefficient increases as ε grows. In line with what has been

[6] In Appendix A.3 we also compute the difference in terms of L1 distance.

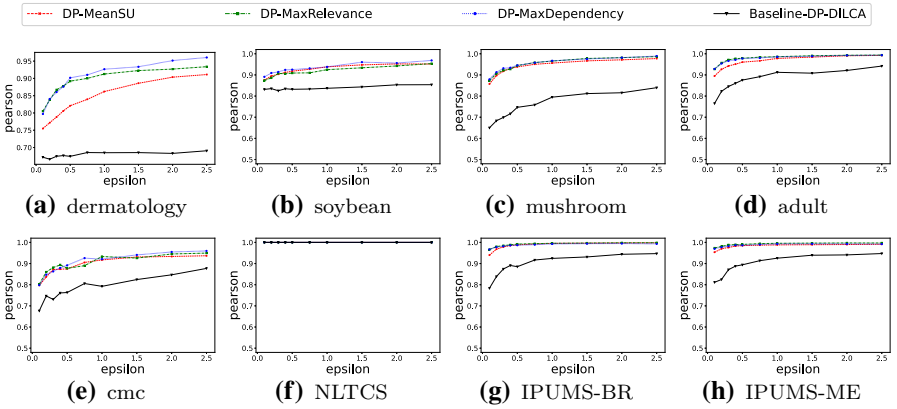


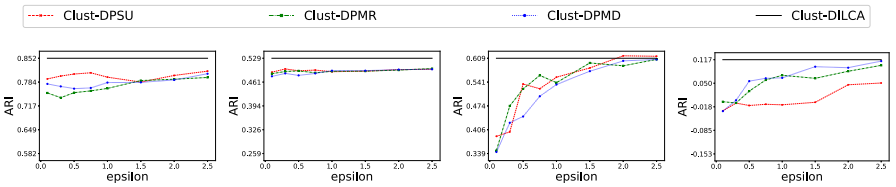
Fig. 3 Average Pearson correlation between the differentially private distance matrices and the correspondent non private ones

said previously, the datasets with the highest values of Pearson coefficient are those with more records. Surprisingly, soybean and dermatology obtain good results too. A possible explanation of this behavior is connected to the high inter-correlation among the attributes of these two datasets: thus, even if the context selection phase fails in identifying the most relevant attributes, the final distance computation is not that affected, as the selected context is still relevant.

4.3 Statistical validation of the results

In order to have a statistical validation of the results, we conduct three different types of tests. All the details about the tests and the complete results are reported in Appendix A.4.

1. In the previous sections we have used DP-DILCA to compute the context and the distance matrix for each target attribute of each dataset, with different levels of privacy budget ε and with three different variants of the algorithm, for a total of 4410 experiments. Each experiment has been repeated 30 times. For each set of experiments, we want to understand whether the distortion introduced by the private algorithm is too high, making the results statistically similar to those we would obtain with an algorithm that randomly selects the output (context or distance matrix). Thus, we compare each one of the 4410 sets of results with those obtained selecting the contexts and the distance matrices uniformly at random, and perform a Mann-Whitney U test to test the null hypothesis that the two sets of results belong to the same distribution. The results of the tests lead to slightly different conclusions for the F-score and the Pearson coefficient. As regards the former, we observe that, for low values of ε and for some target attributes, we cannot reject the null hypothesis. This is particularly true in the smallest datasets and for variant DP-MeanSU of the algorithm. In these cases, then, the context selection performed by the private algorithm is not significantly better than the random context selection. As ε increases, the number of experiments for which



(a) dermatology **(b)** soybean **(c)** mushroom **(d)** adult

Fig. 4 ARI of the clustering results

the null hypothesis can be rejected grows; for $\varepsilon \geq 1.5$ almost all the experiments show contexts that are statistically better than the random selected ones. As regards the Pearson coefficient, instead, for each one of the 4410 sets of results we can always reject the null hypothesis: for each level of privacy budget and regardless of the shape of the dataset, all the versions of our algorithm find distance matrices that are significantly more correlated to the correct distance matrices computed by non-private DILCA than if we would have randomly generated them. Thus, we can exclude that the high levels of Pearson coefficient are reached by chance.

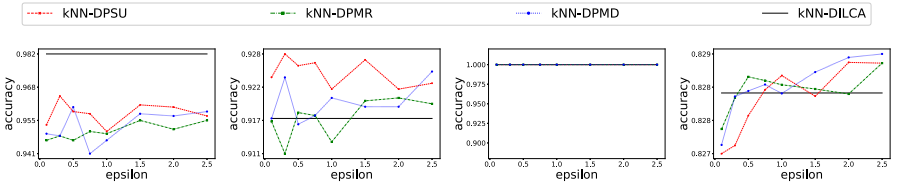
2. Figures 2 and 3 show an increasing trend of the results as ε increases, for both the F-score of the context and the Pearson coefficient of the distance matrices. We use a Page's trend test (Page 1963) at confidence level $\alpha = 0.01$ to test the null hypothesis that $m_{0.1} = m_{0.2} = \cdots = m_2 = m_{2.5}$ against the alternative hypothesis that $m_{0.1} \leq m_{0.2} \leq \cdots \leq m_2 \leq m_{2.5}$, with at least one strict inequality, where m_ε is the mean of the considered measure (F-score or Pearson coefficient) on the experiments with privacy budget ε. We conduct the test for each variant of the algorithm separately. We can reject the null hypothesis for all the variants of the algorithm and for both the measures (the maximum p-value in the six tests is 1.21×10^{-71}). Thus, we conclude that the quality of the outcomes of DP-DILCA significantly grows as the privacy budget ε increases.
3. Finally, we conduct a Friedman statistical test followed by a Nemenyi post-hoc test (Demsar 2006) in order to asses whether the differences among the three variants of DP-DILCA are statistically significant. For this test, we are interested in comparing the "quality" of the three variants on the final output of the algorithm, the distance matrix, thus we test the null hypothesis that the differences among the Pearson coefficients associated to the matrices computed by the three variants are not statistically significant. At confidence level $\alpha = 0.01$, the null hypothesis of the Friedman test can be easily rejected (p-value: 2.5×10^{-43}); we then proceed with the Nemenyi post-hoc test. The results show that the difference between DP-MaxRelevance and DP-MeanSU is higher than the critical difference, and the same applies to the difference between DP-MaxDependency and DP-MeanSU (the p-values are, respectively, 1.16×10^{-36} and 2.85×10^{-36}). The difference between DP-MaxRelevance and DP-MaxDependency, instead, is not significant. We can conclude that DP-MaxRelevance and DP-MaxDependency are statistically better than DP-MeanSU.

4.4 Experiments on clustering and classification

In this section, we assess the effectiveness and utility of the distances computed by our differentially private algorithms. To this purpose, we embed DP-DILCA into two distance-based learning algorithms: the Ward's hierarchical clustering algorithm and the kNN classifier. Both the algorithms take as input the matrix of the pairwise distances between the data objects. DP-DILCA's output is the distance between values of a categorical attribute; if it is applied to all attributes in F, then the distance between any pair of objects o_i, o_j, both described by F can be computed as $objDist(o_i, o_j) = \sqrt{\sum_{X \in F} distMatrix_X[o_i.X, o_j.X]^2}$, where $distMatrix_X$ is the distance matrix returned by DP-DILCA for attribute X and $o_i.X$ and $o_j.X$ are the values of attribute X on objects o_i and o_j (Ienco et al. 2012). We will refer to this metric as $objDist_{DPSU}$, $objDist_{DPMR}$, $objDist_{DPMD}$, depending on the variant of DP-DILCA (DP-meanSU, DP-MaxRelevance, DP-MaxDependency respectively) used to compute the distances among the categorical values of each attribute. Similarly, we will call $objDist_{DILCA}$ the metric obtained by the non-private DILCA algorithm. Given a dataset D with m categorical attributes, the algorithm that returns one of the perturbed version of the metric is ε-differentially private if, for each $X \in \mathcal{F}$, $distMatrix_X$ is computed with privacy budget ε/m. We assess the accuracy of this metric in Appendix A.5.

We run the experiment about clustering as follows: for each real-world dataset, we compute the object distance matrix using the different private and non private metrics, then we run Ward's hierarchical clustering with these matrices as input. Since the hierarchical algorithm returns a dendrogram which, at each level, contains a different number of clusters, we consider the level corresponding to the number of clusters equal to the number of classes. We call the overall clustering models $Clust_{DPSU}$, $Clust_{DPMR}$, $Clust_{DPMD}$ and $Clust_{DILCA}$, depending on the distance metric adopted. We evaluate the quality of the results through the adjusted rand index (ARI) computed w.r.t. the actual classes (Hubert and Arabie 1985). For this reason we do not run this experiment on datasets IPUMS-BR, IPUMS-ME and NLTCS, for whom the classes are not given. We also exclude from the experiment dataset cmc, because the given classes do not match at all the results obtained through the clustering algorithm in the non-private setting (ARI is around 0.01, as for the expected index computed for a random clustering).

Figure 4 shows the mean ARI results over 30 experiments. The value of ε on the x axis of the plot is the overall privacy budget used for the learning of the metric, while the privacy budget spent for computing the distances among values of a single attribute is $\frac{\varepsilon}{m}$. For all the datasets, the ARI values of the clustering models with private distance computation grow with respect to the privacy budget, but the growth is more pronounced in the two largest datasets, adult and mushroom. Here, for high values of ε, they get results close to those of the clustering with non-private distances. The three private distance computation methods have similar performances in terms of ARI in



(a) dermatology **(b)** soybean **(c)** mushroom **(d)** adult

Fig. 5 Accuracy of the results of the kNN classification with $k = 5$

dermatology, soybean and mushroom, while in adult $Clust_{DPMR}$ and $Clust_{DPMD}$ outperform $Clust_{DPSU}$[7].

As last experiment, we run the kNN classification algorithm, with $k = 5$. We perform a 4-fold cross-validation: one fold is retained as test set, then the metrics $objDist_{DPSU}$, $objDist_{DPMR}$, $objDist_{DPMD}$ and $objDist_{DILCA}$ are learned on the remaining 3 folds and the classification model is trained on the same set. We call the overall models kNN_{DPSU}, kNN_{DPMR}, kNN_{DPMD} and kNN_{DILCA}, depending on the distance learning algorithm used. For each dataset, we apply the four kNN models 30 times and compute the mean accuracy of the classification on the the test set. The process is repeated four times and the results are further averaged on the four test sets.

In Figure 5 we report the mean accuracy of all the models for increasing levels of privacy budget ε. The results of kNN_{DPSU}, kNN_{DPMR} and kNN_{DPMD} are always very close to those of kNN_{DILCA}, even for very low levels of ε. On dataset mushroom, the results of the private and non-private models are perfect; on dataset soybean and adult, kNN_{DPSU}, kNN_{DPMR} and kNN_{DPMD} even outperform their non-private competitor. The variation of the privacy budget ε seems to have no impact on the accuracy of the model, except for the largest dataset, adult, for which a slight increase of the curves w.r.t. ε can be appreciated. In conclusion, we can say that the noise introduced in the distance computation phase does not affect the classification results too much: this is due to the fact that the distances among objects obtained with DP-DILCA are very similar to those obtained with non-private DILCA (see Appendix A.5).

5 Conclusion

We have introduced a new family of differentially private algorithms for the data-driven computation of meaningful and expressive distances between any two values of a categorical attribute. Our approach is built upon an effective context-based distance learning framework whose output, however, may reveal private information if applied to a secret dataset. For this reason, we have proposed several randomized procedures, based on the Laplace and exponential mechanisms, that satisfy ε-differential privacy and return accurate distance measures even with relatively small privacy budget consumption. Additionally, the metric learnt by our approach can be used profitably in

[7] In Appendix A.6 we show that using DPDILCA with a differentially private k-means clustering algorithm outperforms the same algorithm combined with the Euclidean distance.

distance-based machine learning algorithms, such as hierarchical clustering and kNN classification.

The possible limitations of some of our algorithms concern the choice of a correct context size and the applicability in "hard" scenarios (e.g., small and/or high-dimensional datasets). As regards the first point, note that it is not possible to test different values of the context size parameter k, since this would waste a large part of the privacy budget. As future work, we will investigate a method to identify an optimal value of k. Moreover, when k is high, DP-MaxDependency may require too much computational time with very high-dimensional datasets, since it computes the probability associated to any possible context for each attribute of the dataset. We plan to address this issue by investigating more intelligent ways to explore the context search space. As for the second point, the results has shown that our method achieves the best performances on sufficiently large datasets, and that the quality of the results deteriorates when the number of attributes increases. However, the experiments show that the algorithm is able to find accurate distances in datasets with up to 35 attributes.

As further future work, we will optimize our metric for improving its computation with ordinal attributes, as well as in datasets where numerical and categorical variables coexist. Moreover, we will also leverage semantic relationships among categorical values to estimate better and more explainable distances. Finally, we will design specific unsupervised and (semi)supervised machine learning algorithms adopting our distance learning framework and satisfying differential privacy.

Acknowledgements The authors are thankful to Shuyi Yang for stimulating discussions.

Funding Open Access funding provided by Università degli Studi di Torino within the CRUI-CARE Agreement. This work is supported by Fondazione CRT (grant number 2019-0450).

Availability of data and material. All data are available online and accessible to everyone.

Declarations

Conflicts of interest/Competing interests. Not applicable.

Code availability Source code and scripts used in our experiments are available at https://github.com/elenabattaglia/dpdilca.

Appendix A Additional experiments and results

A.1 Context selection in synthetic datasets

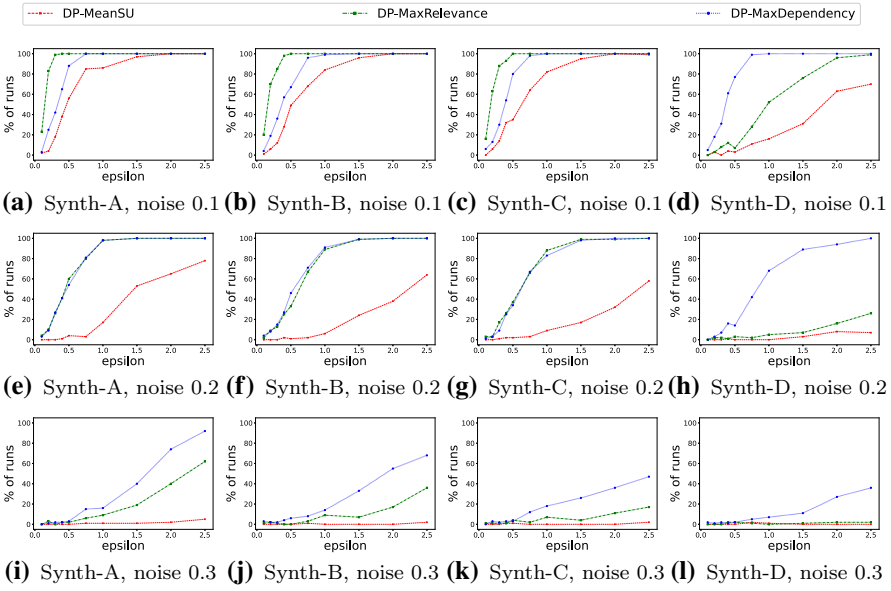
Here we present another set of experiments with the goal of showing which context selection approach works better and in which cases. To do this, we test our algorithms on synthetic datasets, in order to have some controlled scenarios. Each dataset is a 10000×11 boolean matrix, the first column of each being considered as the target attribute. We consider four types of matrices:

- *Synth-A*: the target values contains 5000 ones and 5000 zeros. Three columns are generated starting from the target attribute and changing the values of some randomly selected entries (swapping 1 with 0 and vice versa). The amount of noise introduced by this swapping procedure is controlled by a parameter n, which represents the portion of entries that are swapped. These are considered as context attributes. The remaining seven columns are created uniformly at random.
- *Synth-B*: the procedure we use to generate this matrix is the same used for Synth-A, but the target attribute is unbalanced, as it contains 2000 ones and 8000 zeros.
- *Synth-C*: with the same procedure used for Synth-A and Synth-B, we generate three columns with a fixed level of noise n (the context attributes). Then we generate other two columns with a higher level of noise ($n = 0.35$). These columns can be interpreted as redundant w.r.t. the first three columns. The final six columns are created uniformly at random.
- *Synth-D*: here we create a perfect 10000×5 block matrix with five blocks of ones. Then we add a certain amount of noise with the swapping procedure described above. We will consider the first column of the matrix as the target attribute, while the other four columns form its context. The remaining six columns are created uniformly at random.

For each type, we create three different matrices, with level of noise $n \in \{0.1, 0.2, 0.3\}$, for a total of twelve synthetic datasets. The main characteristics of the synthetic datasets are summarized in Table 2. We run the three variants of DP-DILCA on each synthetic dataset 100 times and we count the number of times the correct context is selected. The number k of desired attributes in the context, when required, is set equal to the number of attributes in the correct context. The results of the experiments are reported in Figure 6. The less challenging scenarios are Synth-A and Synth-B: here, when the level of noise is 0.1, the algorithm that works better is DP-MaxRelevance, which for $\varepsilon > 0.3$ stably identifies the correct context. To get the same stability the other methods need higher levels of privacy budget ($\varepsilon = 0.75$ for DP-MaxDependency and $\varepsilon = 2$ for DP-MeanSU). As the noise increases, the results of all the methods degrade; the algorithm showing less sensitivity to noise is DP-MaxDependency. The most sensitive, instead, is DP-MeanSU, that with noise $n = 0.3$ never identifies the perfect context. To be fair, we have to consider that DP-MeanSU is disadvantaged compared to other methods because it does not know in advance the number of elements in the context (it generally puts also irrelevant elements in it). The results obtained on dataset Synth-A and Synth-B are very similar, so we can conclude that the oddity in the representation of the values of the target attribute does not affect very much the quality of the results. Comparable

Table 2 Synthetic dataset characteristics

Dataset	portion of ones in Y	|context(Y)|	# weakly correlated attributes
Synth-A	0.5	3	0
Synth-B	0.2	3	0
Synth-C	0.2	4	0
Synth-D	0.3	2	2



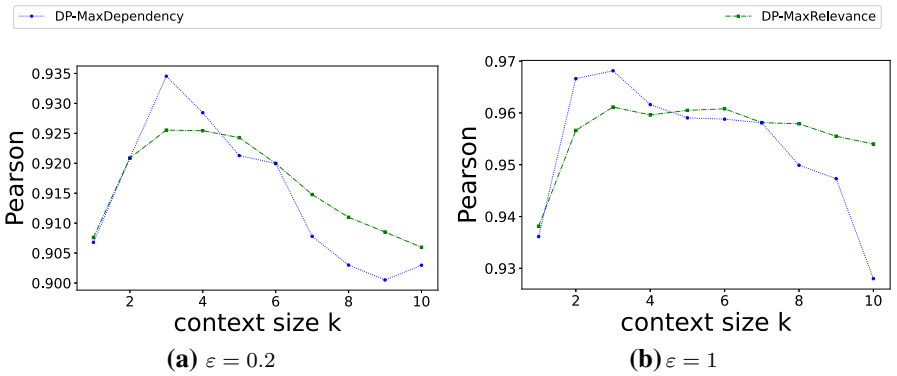
(a) Synth-A, noise 0.1 **(b)** Synth-B, noise 0.1 **(c)** Synth-C, noise 0.1 **(d)** Synth-D, noise 0.1

(e) Synth-A, noise 0.2 **(f)** Synth-B, noise 0.2 **(g)** Synth-C, noise 0.2 **(h)** Synth-D, noise 0.2

(i) Synth-A, noise 0.3 **(j)** Synth-B, noise 0.3 **(k)** Synth-C, noise 0.3 **(l)** Synth-D, noise 0.3

Fig. 6 Pecentage of runs in which DP-DILCA selects the correct context on synthetic datasets

results are achieved also in Synth-C, except for noise equal to 0.3 (Fig 6(k)) where the results are worst for each variant of DP-DILCA. In this case, it must be considered that the columns in the context are generated with noise $n = 0.3$, while other columns outside the context are generated with a very similar level of noise, $n = 0.35$. In other words, there is only a subtle distinction between the attributes that should be selected and the attributes that should be discarded. Despite this, DP-MaxDependency is able to identify the correct context about half of the times, for sufficiently high levels of ε. The most challenging scenario is Synth-D: here DP-MaxDependency outperforms the other algorithms for all levels of noise. This is not a surprise: the attributes in the correct context of Synth-D give a good description of the target attribute when considered all together, while the single contribution in terms of Mutual Information of each attribute is not very high. Thus, in this scenario a context selection method as DP-MaxDependency, which considers the global association of a set of attribute with the target attribute, is preferable.



(a) $\varepsilon = 0.2$ **(b)** $\varepsilon = 1$

Fig. 7 Average Pearson coefficient between distance matrices computed by DP-DILCA and DILCA, for increasing values of k of elements in context, and two values of privacy budget ε

A.2 Sensitivity analysis of parameter *k*

In this section we assess the impact of the choice of parameter k on the results of algorithms DP-MaxRelevance and DP-MaxDependency. According to our theoretical analysis, the amount of injected noise depends on the number of contingency tables that the algorithm has to compute: the amount of noise injected by the Laplace mechanism to each cell of the contingency matrices between the target attribute and the attributes in its context is inversely proportional to the privacy budget and the privacy budget spent for each contingency matrix is $\frac{\varepsilon}{k}$, where k is the number of attributes in the context. Consequently, one may think that the best value for k should always be the lowest one ($k = 1$). However, the quality of the final distances depends not only on the the amount of noise added in the computation of the contingency tables, but also on the choice of a good context: it is true that if we have to compute only one contingency table we will end in a final distance matrix that should be more similar to the one computed without noise injection. But, if a unique attribute is not able to fully capture the differences among the values of the target attribute, the final distances will be worst than those we could obtain by increasing the number of attributes allowed in the context. Furthermore, setting a "wrong" k can affect also the stability of the differentially private context selection phase. In our experiments, the average number of elements selected in the contexts by non-private $DILCA_M$, considering as target each attribute of each dataset, is three and for this reason we set $k = 3$. This turns out to be a good choice. Figure 7 shows the average Pearson correlation index between the private and non-private distance matrices on all our experiments, for two different values of ε: the best results are those obtained setting $k = 3$. It is worth noting, however, that the overall variation of the Person's coefficient (in particular, for DP-MaxRelevance) is not that wide.

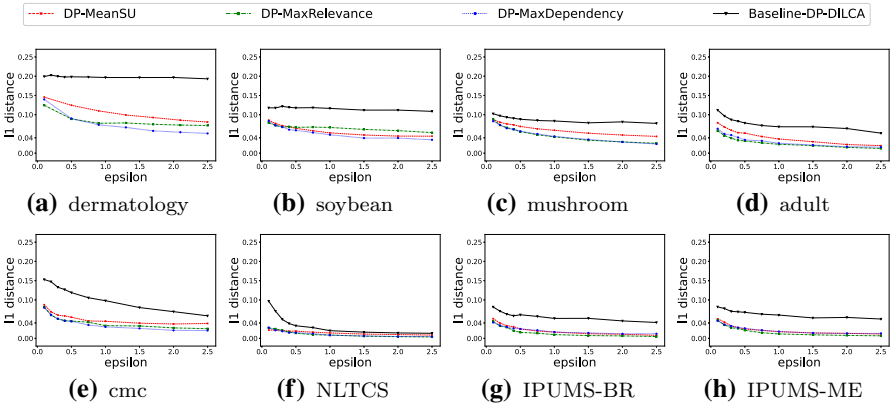


Fig. 8 Average distance in l1 norm between the DP distance matrix and the correspondent non private distance matrix

A.3 More on the assessment of the distance matrices

We assess the similarity between the private and non-private distance matrix of each target attribute through two different measures: the L1 distance and the Pearson's correlation coefficient. The L1 distance quantifies how far the distances computed in a differentially private way are from the non-private distances; given two matrices M and M' with the same shape $n \times m$, it is defined as

$$d_{L1}(M, M') = \sum_{i=1}^{n} \sum_{j=1}^{m} |M_{ij} - M'_{ij}|.$$

The magnitude of $d_{L1}(M, M')$ depends on the size of matrix M. In order to compare the results of DP-DILCA obtained for different target values (with different shapes, though) we normalize the L1-distance over the shape of the distance matrix: for instance, if the target attribute has t different values, we divide the L1 distance by t^2. For each ε, we repeat the experiments 30 times and we compute the mean value of both the normalized L1 distance (Fig. 8) and the sample Pearson correlation coefficient (Fig. 3).

Figure 8 shows the results of our computations: for each ε we reported the average value of the measure over all the attributes of each dataset. The results show that the distance between private and non-private distance matrices decreases as ε grows. In line with what has been said previously, the datasets with the lowest values of normalized L1 distance are those with more records. Surprisingly, soybean and dermatology obtain good results too. A possible explanation of this behavior is connected to the high inter-correlation among the attributes of these two datasets, as already observed in Section 4.1. Thus, even if the context selection phase fails in identifying the most relevant attributes, the final distance computation is not that affected, as the selected context is still relevant.

A.4 More on the statistical validation of the results

We conduct three series of statistical tests to validate the results of the experiments described in Section 4.

A.4.1 Statistical validation of the contexts

To better understand the validity of the results, for each value of ε and each target attribute, we compare the mean F-score obtained with each variant of DP-DILCA with the mean F-score we would have obtained if the context selection had been performed uniformly at random. In more detail, we compare the results of DP-MaxRelevance and DP-MaxDependency with those obtained considering all the contexts with three elements equally probable. The results of DP-MeanSU, instead, are compared with those obtained by randomly selecting a context among all the possible contexts (with all possible sizes). For each set of results, we conduct a Mann-Whitney U test to verify the null hypothesis that the two sets of results, the first from DP-DILCA and the second from random context selection, belong to the same distribution. In this way, we evaluate whether the private algorithms introduce too much noise (an amount of noise such that the results become similar to those one would obtain by chance) or not. We opt for a non-parametric test because the distribution of the F-score does not follow a normal distribution, neither when the context is randomly selected nor when it is selected by DP-DILCA.

We have a total of 4410 set of experiments, so we conduct 4410 tests. To cope with the problem of multiple comparisons, we use the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995) to control the False Discovery Rate at level $\alpha = 0.01$.

In all the largest datasets, the results of the tests are similar: for the vast majority of the target attributes, the null hypothesis can be rejected at any level of privacy budget ε. There are only few attributes for which the test is not passed when the value of ε is very small. For higher values of ε, all the experiments become significant. It is worth noticing that these "problematic" attributes are those that are weakly correlated with all the other attributes in the same dataset. In these cases, the non-private version of DILCA decides whether to include an attribute in the context or not based on a variation of few thousandths or less in the association measure that captures the correlations between attributes (MI or SU, depending on the strategy). We obtain different results for the smallest datasets (dermatology and soybean): here, for about half of the target attributes, the test is not passed independently of the value of ε, when the algorithm used is DP-MeanSU. Differently, for DP-MaxRelevance and DP-MaxDependency, the number of target attributes for which the results are significant increases w.r.t. ε, and almost all the results become significant when $\varepsilon > 1.5$. Figure 9 shows the percentage of attributes in each dataset for which the three algorithms have F-score that are significantly better than those obtained with random context selection, at different levels of privacy budget. Finally, Table 3 reports the maximum p-values of the tests, grouped by dataset and ε.

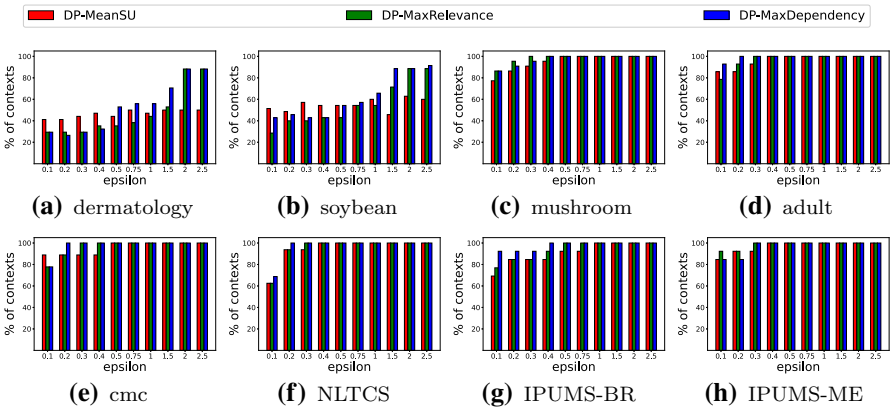


Fig. 9 Number of target attributes for which the context selection performed by DP-DILCA is significantly better than the random context selection, for different level of ε

Table 3 Maximum p-values of the Mann Whitney U tests for the F-score, for each dataset and for different amounts of privacy budget ε. The p-values above the acceptable significance level are in bold

	Privacy Budget					
Dataset	**0.1**	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**
dermatology	$\mathbf{4.8 \times 10^{-1}}$	$\mathbf{4.9 \times 10^{-1}}$	$\mathbf{4.7 \times 10^{-1}}$	$\mathbf{4.6 \times 10^{-1}}$	$\mathbf{4.7 \times 10^{-1}}$	$\mathbf{4.7 \times 10^{-1}}$
soybean	$\mathbf{4.7 \times 10^{-1}}$	$\mathbf{4.8 \times 10^{-1}}$	$\mathbf{4.7 \times 10^{-1}}$	$\mathbf{4.6 \times 10^{-1}}$	$\mathbf{4.0 \times 10^{-1}}$	$\mathbf{4.1 \times 10^{-1}}$
cmc	$\mathbf{9.5 \times 10^{-2}}$	5.7×10^{-3}	1.5×10^{-3}	3.4×10^{-8}	3.2×10^{-7}	3.1×10^{-7}
mushroom	$\mathbf{3.1 \times 10^{-2}}$	4.6×10^{-3}	8.5×10^{-9}	1.3×10^{-6}	5.0×10^{-11}	4.3×10^{-12}
NLTCS	$\mathbf{9.1 \times 10^{-2}}$	4.1×10^{-7}	9.9×10^{-10}	5.1×10^{-11}	4.2×10^{-11}	3.6×10^{-11}
IPUMS_BR	$\mathbf{4.1 \times 10^{-1}}$	$\mathbf{9.7 \times 10^{-2}}$	4.7×10^{-3}	9.0×10^{-4}	2.7×10^{-4}	3.7×10^{-4}
IPUMS_ME	$\mathbf{3.0 \times 10^{-1}}$	3.3×10^{-3}	6.0×10^{-8}	1.3×10^{-8}	5.5×10^{-9}	9.1×10^{-10}
adult	$\mathbf{2.4 \times 10^{-2}}$	1.1×10^{-3}	4.4×10^{-7}	1.2×10^{-8}	4.9×10^{-10}	8.2×10^{-11}

A.4.2 Statistical validation of the distance matrices

We repeat the same experiment also for assessing the statistical validity of the Pearson coefficient scores of the experiments in Section 4.2. Thus, for each dataset, target attribute, ε and variant of the algorithm, we compare the set of 30 Pearson coefficients obtained by DP-DILCA with those obtained randomly selecting the distance matrix among all the possible distance matrices with the same shape. We conduct a Mann-Whitney U test to reject the null hypothesis that the two sets of results, the first from DP-DILCA and the second from random distance computation, belong to the same distribution. Again, we use the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995) to control the False Discovery Rate at level $\alpha = 0.01$. In all cases, we can always reject the null hypothesis. We conclude that all the variants of DP-DILCA compute distance matrices that are significantly correlated with the non-private

Table 4 Maximum p-values of the Mann Whitney U tests for the Pearson coefficient, for each dataset and for different amounts of privacy budget ε

	Privacy Budget					
Dataset	**0.1**	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**
dermatology	7.1×10^{-3}	1.8×10^{-3}	7.7×10^{-4}	4.5×10^{-4}	5.6×10^{-5}	1.8×10^{-5}
soybean	4.5×10^{-3}	3.9×10^{-4}	6.1×10^{-4}	2.1×10^{-4}	5.4×10^{-5}	5.0×10^{-5}
cmc	3.1×10^{-3}	2.7×10^{-3}	2.2×10^{-4}	2.0×10^{-4}	1.9×10^{-4}	2.7×10^{-5}
mushroom	2.6×10^{-7}	1.2×10^{-7}	1.0×10^{-7}	3.7×10^{-8}	3.2×10^{-8}	3.2×10^{-8}
IPUMS_BR	2.0×10^{-7}	3.2×10^{-8}	3.2×10^{-8}	3.2×10^{-8}	3.2×10^{-8}	3.2×10^{-8}
IPUMS_ME	4.7×10^{-7}	1.3×10^{-7}	4.2×10^{-8}	3.2×10^{-8}	4.2×10^{-8}	3.2×10^{-8}
adult	3.8×10^{-7}	3.2×10^{-8}	3.2×10^{-8}	3.2×10^{-8}	3.2×10^{-8}	3.2×10^{-8}

matrices computed by DILCA. Table 4 reports the maximum p-value of the tests, grouped by dataset and ε.

A.4.3 Growth of the results w.r.t. the privacy budget

Figures 2 and 3 show an increasing trend of the results w.r.t the privacy budget ε, for both the F-score of the context and the Pearson coefficient of the distance matrices. We use a Page's trend test (Page 1963) at confidence level $\alpha = 0.01$ to test the null hypothesis

$$\mathcal{H}_0 : m_{0.1} = m_{0.2} = m_{0.3} = m_{0.4} = m_{0.5} = m_{0.75} = m_1 = m_{1.5} = m_2 = m_{2.5}$$

against the alternative hypothesis

$$\mathcal{H}_{\mathcal{A}} : m_{0.1} \leq m_{0.2} \leq m_{0.3} \leq m_{0.4} \leq m_{0.5} \leq m_{0.75} \leq m_1 \leq m_{1.5} \leq m_2 \leq m_{2.5}$$

with at least one strict inequality, where m_ε is the mean of the considered measure (F-score or Pearson coefficient) among the experiments with privacy budget ε. We conduct the test for each variant of the algorithm separately. The values of the statistics, the critical values and the p-values associated to each statistic are reported in Table 5. We can reject the null hypothesis for all the variants of the algorithm and for both measures. Thus, we conclude that both the F-score and the Pearson coefficient significantly grow with the privacy budget ε.

A.4.4 Comparison of the variants of DP-DILCA

Finally, we conduct a Friedman statistical test followed by a Nemenyi post-hoc test (Demsar 2006) to asses whether the differences among the three variants of DP-DILCA are statistically significant, at confidence level $\alpha = 0.01$. The value of the test statistic is $Q = 210$ and it is higher than the critical value $CV = 9.21$, with a p-value of 2.5×10^{-43}, thus the null hypothesis of the Friedman test can be rejected for the

Table 5 The values of the statistics of the Page's trend test, for each variant of DP-DILCA, and their correspondent p-values. The critical value for the test statistic, at significance level $\alpha = 0.01$ is 5.412

	F-score		**Pearson**	
	Test Statistic	**p-value**	**Test Statistic**	**p-value**
DP-MeanSU	327.70	1.52×10^{-73}	351.41	1.04×10^{-78}
DP-MaxRelevance	369.49	1.21×10^{-82}	318.98	1.21×10^{-71}
DP-MaxDependency	487.15	2.97×10^{-108}	344.85	2.81×10^{-77}

Pearson coefficient values; we then proceed with the Nemenyi post-hoc test. The results show that the difference between DP-MaxRelevance and DP-MeanSU is $D_1 = 0.6$ and is greater than the critical difference $CD = 0.1490$, and the same applies to the difference between DP-MaxDependency and DP-MeanSU that is $D_2 = 0.59$ (the p-values associated to the test statistics D_1 and D_2 are, respectively, 1.16×10^{-36} and 2.85×10^{-36}). The difference between DP-MaxRelevance and DP-MaxDependency, instead, is $D_3 = 0.003$, lower than the critical difference and not significant (p-value: 0.53). We can conclude that DP-MaxRelevance and DP-MaxDependency are statistically better than DP-MeanSU.

A.5 Assessment of the object distance matrices

We recall that $objDist_{DPSU}$ (or $objDist_{DPMR}$ or $objDist_{DPMD}$) is a distance metric defined as follows: for each pair of objects o_i, o_j described by the set of categorical attributes F,

$$objDist_{DPSU}(o_i, o_j) = \sqrt{\sum_{X \in F} distMatrix_X[o_i.X, o_j.X]^2}$$

where $distMatrix_X$ is the matrix containing the distances among values of attribute X obtained through the application of DP-MeanSU (DP-MaxRelevance or DP-MaxDependency respectively).

Given a dataset x with m categorical attributes, the algorithm that returns one of $objDist_{DPSU}$, $objDist_{DPMR}$, $objDist_{DPMD}$ is ε-differentially private if, for each $X \in \mathcal{F}$, $distMatrix_X$ is computed with privacy budget ε/m.

In this section we assess the quality of the distance metrics as follows: we learn the three metrics $objDist_{DPSU}$, $objDist_{DPMR}$ and $objDist_{DPMD}$ and also the non-private distance function $objDist_{DILCA}$ on the real-world data used for all the experiments, for increasing values of ε. Then we apply the learned functions on the same datasets, obtaining the pairwise distance matrix between the objects. The results are reported in Figure 10 and Figure 11. The value of ε on the x axis of the plot is the overall privacy budget used for the learning of the metric, while the privacy budget spent for computing the distances among values of a single attribute is $\frac{\varepsilon}{m}$. Even for very low values of privacy budget, the distance matrices obtained with the three variants of DP-DILCA are all very close to the non-private objects distance matrices,

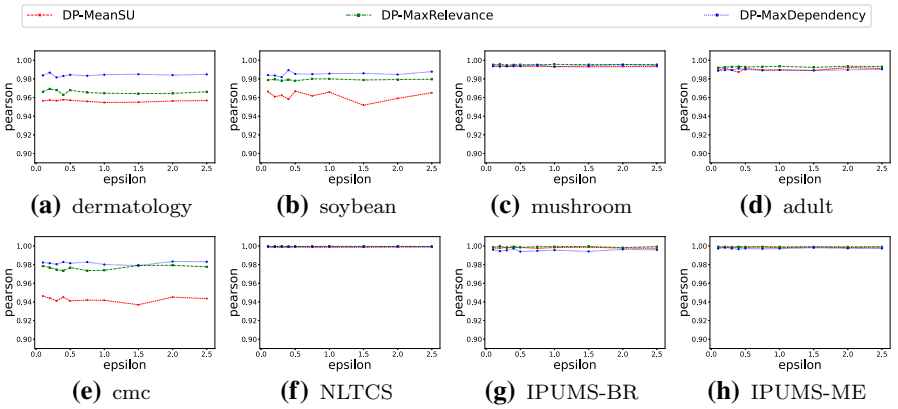


Fig. 10 Average Pearson correlation coefficient between the DP object distance matrix and the correspondent non private object distance matrix

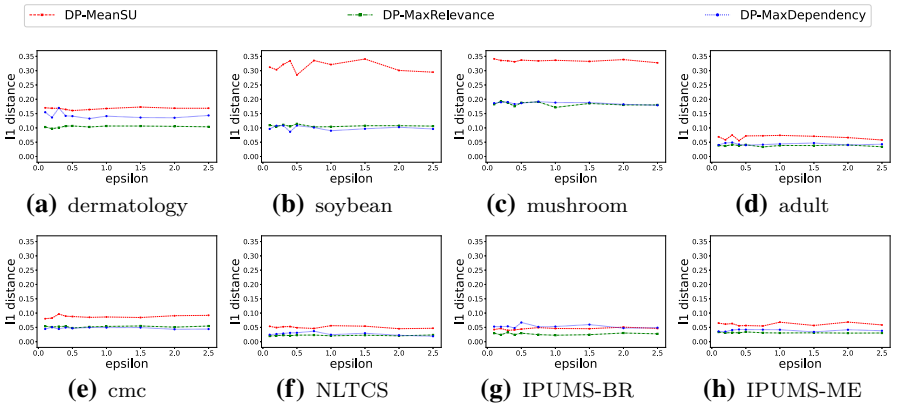


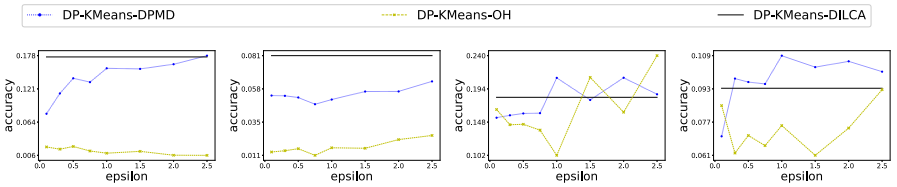
Fig. 11 Average distance in l1 norm between the DP object distance matrix and the correspondent non private object distance matrix

especially in the bigger datasets. Finally, in these experiments it clearly emerges that algorithms DP-MaxRelevance and DP-MaxDependency outperform DP-MeanSU in all the datasets.

A.6 Differentially private k-means clustering

In this section we give a practical and fully private application of DP-DILCA. Suppose one wants to apply a differentially private version of k-means clustering algorithm (we will call it DP-Kmeans) to a secret dataset with categorical attributes. DP-KMeans[8] (Su et al. 2017), as well as non-private k-means, only applies to numerical datasets. An easy way to apply DP-KMeans to a categorical dataset is to transform it in a numerical dataset by encoding each categorical attribute X in a bit vector of length $|X|$, where

[8] We use the implementation provided in https://github.com/IBM/differential-privacy-library



(a) dermatology **(b)** soybean **(c)** mushroom **(d)** adult

Fig. 12 ARI of the results of DP-Kmeans

each entry represents a possible value for attribute X (this transformation is known as One-Hot encoding). We call the overall algorithm DP-KMeans-OH. The Euclidean distance applied to such vectors is not able to distinguish between different values of the same attributes and it treats matches and mismatches all with the same weight. A more sophisticated way to project the categorical dataset in an Euclidean space is to exploit the distances among values computed by DP-DILCA: we can represent each value of each categorical attribute X as a point in a $\mathbb{R}^d$, where $d \leq |X|$, following the method proposed by Crippen and Havel (1978). The pairwise Euclidean distances between the objects of the transformed dataset coincides with the pairwise distances of $objDist_{DPMD}$ over the original dataset (for this experiment, we use only DP-MaxDependency, since, according to the results reported in the previous sections, it is the variant of DP-DILCA that works better). We call the overall algorithm DP-KMeans-DPMD.

In the experiment, we keep the privacy budget spent by DP-Kmeans fixed to $\varepsilon_{KMeans} = 2$, while we compute $objDist_{DPMD}$ with different levels of privacy budget ε_{DPMD} between 0.1 and 2.5. The overall privacy budget of DP-KMeans-DPMD is $\varepsilon = \varepsilon_{KMeans} + \varepsilon_{DPMD}$. We compare the results of DP-KMeans-DPMD with those of DP-KMeans-OH, where the privacy budget ε is entirely devoted to the clustering phase. We also apply DP-KMeans to the distances learned by non-private DILCA: we call this algorithm DP-KMeans-DILCA. Note that DP-KMeans-DILCA does not respect ε_{KMeans}-differential privacy even if the clustering is done in a differentially private manner, with $\varepsilon_{KMeans} = 2$, because the preprocessing phase does not guarantee differential privacy. In Figure 12 we report the results in terms of average ARI over 100 experiments. The value of ε on the x-axis refers to the privacy budget devoted to the distance learning: the overall privacy budget of the clustering algorithms is $\varepsilon + 2$. Algorithm DP-KMeans-DPMD outperforms DP-KMeans-OH in almost all cases.

A.7 Execution time analysis

In this section, we analyze the execution time of the three variants of DP-DILCA. The three methods differ on the way in which they compute the context of the target attribute: DP-MeanSU computes $2m-1$ entropies, DP-MaxRelevance $2m-2$ entropies and DP-MaxDependency $2 \cdot \binom{m-1}{k}$ entropies, where m is the number of attributes and k is the number of attributes in the context. Thus, we expect DP-MaxDependency to be by far the slowest method, especially when the number of attributes m is large.

Table 6 Execution time in seconds of different methods for different datasets

Dataset	DP-MeanSU	DP-MaxRelevance	DP-MaxDependency
cmc	0.01	0.01	0.58
IPUMS_BR	0.07	0.05	7.80
IPUMS_ME	0.08	0.06	8.61
adult	0.11	0.08	13.42
NLTCS	0.05	0.04	6.19
mushroom	0.07	0.03	12.46
dermatology	0.11	0.03	41.82
soybean	0.10	0.03	44.37

In Table 6, we report the execution time of the three variants of DP-DILCA on the real word datsets, expressed in seconds. For each dataset, we consider every attribute as target, one at time, and compute the distances among its values. The context size parameter k is set equal to 3. The results are in line with our expectations: the fastest method is DP-MaxRelevance. DP-MeanSU has comparable performances; the slowest one is DP-MaxDependency and the difference between this method and the others is particularly pronounced in datasets soybean and dermatology, which have highest number of attributes.

A.8 Qualitative evaluation of the results

Here, we provide some insights about the quality of the context selected and the distances computed by DP-DILCA. To this purpose, we choose two different target attributes from dataset "adult" and analyze their contexts and distances.

Let us consider as target the attribute 'age'. Table 7 shows the contexts selected by the non-private context selection methods. $DILCA_M$ and MaxRelevance select the same context: the three attributes more associated with 'age' are 'marital-status' (married, unmarried, divorced, widowed...), 'relationship' (wife, husband, unmarried...) and 'hours-per-week' (the number of hours the person works per week). Intuitively, all these attributes are individually related with the target attribute 'age', but marital-status and relationship bring the same pieces of information: the presence of both the attributes in the context seems to be redundant. Indeed, MaxDependency, which selects the set of attributes that are globally most related to the target, selects another context, choosing attributes that describe different aspects related to attribute 'age' ('marital-status', 'education' and 'occupation'). The fact that $DILCA_M$ and MaxRelevance select the same context means not only that they agree on which values are maximally related to 'age', but also that the other attributes are weakly correlated with the target (the association between 'age' and any other attribute, quantified by the Symmetric Uncertainty, is less then the average).

Consider now the same context selection strategies but in their private versions, with a medium level of privacy budget $\epsilon = 0.5$:

Table 7 Contexts selected by non-private algorithms for attribute 'age'

Context selection method	Attributes in context
$DILCA_M$	'marital-status' , 'relationship', 'hours-per-week'
MaxRelevance	'marital-status' , 'relationship', 'hours-per-week'
MaxDependency	'marital-status' , 'education', 'occupation'

Table 8 Contexts selected by non-private algorithms for attribute 'race'

Context selection method	Attributes in context
DILCA_M	'native-country'
MaxRelevance	'native-country', 'marital-status' , 'relationship'
MaxDependency	'native-country', 'relationship', 'occupation'

- DP-MeanSU tends to select the same context of its non-private counterpart, but sometimes it adds some extra-attribute: for instance, a frequently selected context is {'marital-status' , 'relationship', 'hours-per-week', 'education'} or {'marital-status' , 'relationship', 'hours-per-week', 'occupation'}. It is worth noting that, in our experiments, for this level of ε, we have never observed that attributes clearly non-correlated with 'age', such as 'sex' or 'native-country', have been selected.
- DP-MaxRelevance and DP-MaxDependency always identify the same context of their non-private counterparts.

As another example, consider now target attribute 'race'. There are no attributes that are clearly correlated with this target. However, the only attribute in the dataset that has some connection with 'race' is 'native-country'. Indeed, this is the only attribute selected by $DILCA_M$ in the context of 'race', as shown in Table 8. The other two methods, MaxRelevance and MaxDependency, are forced by design to select other two attributes in the context and choose attributes that, intuitively, should not be related to the race. Notice that, again, MaxRelevance selects two attributes that are highly inter-related ('marital-status' and 'relationship'): since these two attributes are similarly distributed, also their co-distributions with the target attribute 'race' are similar; consequently, when one attribute is selected in the context also the other one is selected. This does not happen for MaxDependency, that is explicitly designed to avoid redundancy.

When we move to the private versions of the algorithms,

- DP-MeanSU selects contexts with many attributes (four, on average). The context always contains 'native-country', but the other attributes seem to be randomly chosen. For instance, selected contexts are {'native-country , 'working-class', 'sex'} and {'native-country , 'working-class', 'education', 'marital-status', 'relationship'}.
- Usually, DP-MaxRelevance selects the same context of its non-private version. However, sometimes, one between 'marital-status' and 'relationship' is substituted

by another attribute (for instance, {'native-country', 'hours-per-week' , 'relationship'} is a relatively frequent context).
– DP-MaxDependency has only two different outcomes: the correct context {'native-country', 'relationship', 'occupation'}, with higher frequency, or {'native-country', 'relationship', 'education'}. This suggests that DP-MaxDependency is more stable than the other two methods.

We now analyze the distance matrix computed by DP-DILCA, in comparison with non-private DILCA. To do this, we consider the distances among the values of attribute 'age'. This attribute has 5 different values (intervals of age) and we choose it because it is easy to interpret the results, since the attribute is ordinal. The distances computed by $DILCA_M$ are coherent with the meaning of the values: in particular, given three values $a < b < c$, the distance d computed by $DILCA_M$ is such that $d(a, b) < d(a, c)$. This is not a trivial property and suggests that, in this case, the algorithm has been able to capture the correct relationships among the data. In order to compare the distances obtained by $DILCA_M$ with those of DP-MeanSU, for each method we rank the distances in decreasing order and we check whether the rankings are the same: in most repetitions of the experiment, the ranking remains the same. However, sometimes two consecutive distances are swapped, but we register a maximum of two swaps from the original non-private ranking. Similar results are obtained with the other two methods, DP-MaxRelevance and DP-MaxDependency.

References

Alamuri M, Raju SB, Negi A (2014) A survey of distance/similarity measures for categorical data. In: Proceedings of IJCNN 2014. IEEE, pp 1907–1914

Anandan B, Clifton C (2018) Differentially private feature selection for data mining. In: Proceedings of ACM IWSPA@CODASPY 2018, pp 43–53

Aumüller M, Bourgeat A, Schmurr J (2020) Differentially private sketches for jaccard similarity estimation. In: Proceedings of SISAP 2020, Springer, pp 18–32

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) 57(1):289–300

Boriah S, Chandola V, Kumar V (2008) Similarity measures for categorical data: A comparative evaluation. Proceedings of SIAM SDM 2008:243–254

Chaudhuri K, Monteleoni C, Sarwate AD (2011) Differentially private empirical risk minimization. J Mach Learn Res 12:1069–1109

Cover TM, Thomas JA (2001) Elements of Information Theory. Wiley

Crippen GM, Havel TF (1978) Stable calculation of coordinates for distance information. Acta Crystallogr. A 34:282–284

Demsar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30

Domingo-Ferrer J, Sánchez D, Blanco-Justicia A (2021) The limits of differential privacy (and its misuse in data release and machine learning). Communications of the ACM (to appear)

Dwork C, Roth A (2014) The algorithmic foundations of differential privacy. Found Trends Theor Comput Sci 9(3–4):211–407

Dwork C, Kohli N, Mulligan D (2019) Differential privacy in practice: Expose your epsilons! Journal of Privacy and Confidentiality 9(2)

Friedman A, Schuster A (2010) Data mining with differential privacy. In: Proceedings of ACM SIGKDD 2010, ACM, pp 493–502

Gao C, Huang C, Lin D, Jin D, Li Y (2020) DPLCF: differentially private local collaborative filtering. In: Proceedings of ACM SIGIR 2020, ACM, pp 961–970

Gursoy ME, Inan A, Nergiz ME, Saygin Y (2017) Differentially private nearest neighbor classification. Data Min Knowl Discov 31(5):1544–1575

Hsu J, Gaboardi M, Haeberlen A, Khanna S, Narayan A, Pierce BC, Roth A (2014) Differential privacy: An economic method for choosing epsilon. In: Proceedings of IEEE CSF 2014, IEEE Computer Society, pp 398–410

Hubert L, Arabie P (1985) Comparing partitions. Journal of Classification 2(1):193–218

Ienco D, Pensa RG (2016) Positive and unlabeled learning in categorical data. Neurocomputing 196:113–124

Ienco D, Pensa RG, Meo R (2012) From context to distance: Learning dissimilarity for categorical data clustering. ACM Trans Knowl Discov Data 6(1):1:1-1:25

Ienco D, Pensa RG, Meo R (2017) A semisupervised approach to the detection and characterization of outliers in categorical data. IEEE Trans Neural Networks Learn Syst 28(5):1017–1029

Kasif S, Salzberg S, Waltz DL, Rachlin J, Aha DW (1998) A probabilistic framework for memory-based reasoning. Artif Intell 104(1–2):287–311

Li Y, Yang J, Ji W (2016) Local learning-based feature weighting with privacy preservation. Neurocomputing 174:1107–1115

McSherry F, Talwar K (2007) Mechanism design via differential privacy. In: Proceedings of IEEE FOCS 2007, IEEE Computer Society, pp 94–103

Page EB (1963) Ordered hypotheses for multiple treatments: A significance test for linear ranks. Journal of the American Statistical Association 58(301):216–230

Peng H, Long F, Ding CHQ (2005) Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238

Shi J, Malik J (1997) Normalized cuts and image segmentation. In: Proceedings of IEEE CVPR 1997, IEEE Computer Society, pp 731–737

Stanojevic R, Nabeel M, Yu T (2017) Distributed cardinality estimation of set operations with differential privacy. In: Proceedings of IEEE PAC 2017, IEEE, pp 37–48

Su D, Cao J, Li N, Bertino E, Lyu M, Jin H (2017) Differentially private k-means clustering and a hybrid approach to private optimization. ACM Trans Priv Secur 20(4):16:1–16:33

Velickovic P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y (2018) Graph attention networks. In: Proceedings of ICLR 2018, OpenReview.net

Xu C, Ren J, Zhang Y, Qin Z, Ren K (2017) Dppro: Differentially private high-dimensional data release via random projection. IEEE Trans Inf Forensics Secur 12(12):3081–3093

Yamaguchi Y, Faloutsos C, Kitagawa H (2016) CAMLP: confidence-aware modulated label propagation. In: Proceedings of SIAM SDM 2016, SIAM, pp 513–521

Yang J, Li Y (2014) Differentially private feature selection. In: Proceedings of IJCNN 2014, IEEE, pp 4182–4189

Yu L, Liu H (2003) Feature selection for high-dimensional data: A fast correlation-based filter solution. Proceedings of ICML 2003:856–863

Zhang K, Wang Q, Chen Z, Marsic I, Kumar V, Jiang G, Zhang J (2015) From categorical to numerical: Multiple transitive distance learning and embedding. In: Proceedings of SIAM SDM 2015, SIAM, pp 46–54

Zhang Y, Cheung Y (2020) An ordinal data clustering algorithm with automated distance learning. In: Proceedings of AAAI 2020, AAAI Press, pp 6869–6876