(Local) Differential Privacy has NO Disparate Impact on Fairness

Héber Hwang Arcolezi, Karima Makhlouf, Catuscia Palamidessi

(Local) Differential Privacy has NO Disparate Impact on Fairness*

Héber H. Arcolezi, Karima Makhlouf, and Catuscia Palamidessi

Inria and École Polytechnique (IPP), Palaiseau, France
{heber.hwang-arcolezi,karima.makhlouf,catuscia}@lix.polytechnique.fr

Abstract. In recent years, Local Differential Privacy (LDP), a robust privacy-preserving methodology, has gained widespread adoption in real-world applications. With LDP, users can perturb their data on their devices before sending it out for analysis. However, as the collection of multiple sensitive information becomes more prevalent across various industries, collecting a single sensitive attribute under LDP may not be sufficient. Correlated attributes in the data may still lead to inferences about the sensitive attribute. This paper empirically studies the impact of collecting multiple sensitive attributes under LDP on fairness. We propose a novel privacy budget allocation scheme that considers the varying domain size of sensitive attributes. This generally led to a better privacy-utility-fairness trade-off in our experiments than the state-of-art solution. Our results show that LDP leads to slightly improved fairness in learning problems without significantly affecting the performance of the models. We conduct extensive experiments evaluating three benchmark datasets using several group fairness metrics and seven state-of-the-art LDP protocols. Overall, this study challenges the common belief that differential privacy necessarily leads to worsened fairness in machine learning.

Keywords: Fairness · Local Differential Privacy · Machine Learning.

1 Introduction

The advent of the Big Data era has brought many benefits but has also raised significant concerns about privacy and algorithm bias in Machine Learning (ML). On the one hand, with massive amounts of data generated and collected by various entities, protecting individuals' personal information has become increasingly challenging. In this context, research communities have proposed different methods to preserve privacy, with ϵ-differential privacy (ϵ-DP) [16] standing out as a formal definition that allows quantifying the privacy-utility trade-off with the parameter ϵ (the smaller, the more private). At the same time, there have been many efforts to develop methods and metrics to evaluate and promote fairness in ML due to unequal treatments of individuals or groups based on factors such as race, gender, or socio-economic status [5, 29–31].

* Version of Record (DBSec'23): https://doi.org/10.1007/978-3-031-37586-6_1.

This means that privacy and fairness are essential for ML to apply in practice successfully. In real-life scenarios, it is not common anymore for entities to have access to *sensitive* (or *protected*[1]) attributes like race due to legal restrictions and regulations[2] governing their collection. Therefore, it can be difficult for these entities to quantify/assess the fairness of the models they deploy since they cannot access the protected attributes used for the fairness assessment. One way to address this problem [32], ignoring legal feasibility, is to enable users to share their sensitive attributes using protocols satisfying Local Differential Privacy (LDP) [25], and learn a non-discriminatory predictor.

However, while collecting the sensitive attribute in a privacy-preserving manner may seem sufficient, it is worth noting that proxy variables can exist [24] and can still lead to inferences about the sensitive attribute (*e.g.*, by exploiting correlations). It is also important to acknowledge that proxy variables may be considered as personal information under the GDPR, requiring the same level of privacy protection. Thus, as collecting multiple sensitive information (*i.e.*, *multidimensional data*) becomes increasingly prevalent in various industries, protecting this information is a legal obligation and an ethical responsibility.

Therefore, this paper contributes to an in-depth empirical analysis of how pre-processing multidimensional data with ϵ-LDP affects the fairness and utility in ML binary classification tasks. We evaluated several group fairness metrics [5, 30], including disparate impact [9], equal opportunity [21], and overall accuracy [12], on benchmark datasets, namely, Adult [14], ACSCoverage [14], and LSAC [40]. To broaden the scope of our study, we have experimentally assessed seven state-of-the-art LDP protocols, namely, Generalized Randomized Response (GRR) [23], Binary Local Hashing (BLH) [10], Optimal Local Hashing (OLH) [39], RAPPOR [18], Optimal Unary Encoding (OUE) [39], Subset Selection (SS) [38, 41], and Thresholding with Histogram Encoding (THE) [39].

Moreover, since proxy variables can still introduce unintended biases and thus lead to unfair decisions [24], we consider the setting in which each proxy (sensitive attribute) is collected independently under LDP guarantees. In other words, applying this independent setting automatically removes the correlation between the proxy attributes. To this end, the privacy budget ϵ should be divided among all sensitive attributes to ensure ϵ-LDP under sequential composition [17]. Let d_s be the total number of sensitive attributes, the LDP literature for multidimensional data [6,37] considers a **uniform** solution that collects each sensitive attribute under $\frac{\epsilon}{d_s}$-LDP. In this paper, we propose a new **k-based** solution that considers the varying domain size k of different sensitive attributes. More precisely, for the j-th sensitive attribute, we allocate $\epsilon_j = \frac{\epsilon \cdot k_j}{\sum_{i=1}^{d_s} k_i}$.

Overall, our study challenges the common belief that using DP necessarily leads to worsened fairness in ML [8, 20]. Our findings show that training a classifier on LDP-based multidimensional data slightly improved fairness results

[1] Throughout this paper, we use the term *sensitive* attribute from a privacy perspective and the term *protected* attribute from a fairness perspective. Note that we always consider *protected* attributes as *sensitive* attributes.

[2] For example, the General Data Protection Regulation (GDPR) [3].

without significantly affecting classifier performance. We hope this work can aid practitioners in collecting multidimensional user data in a privacy-preserving manner by providing insights into which LDP protocol and privacy budget-splitting solutions are best suited to their needs.

In summary, the three main contributions of this paper are:

- We empirically analyze the impact of pre-processing multidimensional data with ϵ-LDP on fairness and utility;
- We compare the impact of seven state-of-the-art LDP protocols under a homogeneous encoding when training ML binary classifiers (see Fig. 1) on fairness and utility;
- We propose a new privacy budget splitting solution named k-based, which generally led to a better privacy-utility-fairness trade-off in our experiments.

All our codes are available in a **GitHub repository** [2].

Outline. The rest of this paper is organized as follows. Section 2 discusses related work. In Section 3, we present the notation, fairness, and LDP protocols used. Next, Section 4 states the problem addressed in this paper and the proposed k-based solution. Section 5 details the experimental setting and main results. Finally, we conclude this work indicating future perspectives in Section 6.

2 Related Work

The recent survey work by Fioretto *et al.* [19] discusses two views about the relationship between central DP and fairness in learning and decision tasks. The first view considers DP and fairness in an aligned space (*e.g.*, [15]), which mainly corresponds to individual fairness metrics. The other view regards DP and fairness as "enemies" (*e.g.*, [8, 20, 34]), which mainly corresponds to group fairness notions. For instance, Pujol *et al.* [34] investigated disparities in decision tasks using ϵ-DP data. Regarding learning tasks, Bagdasaryan, Poursaeed, & Shmatikov [8] studied the impact of training ϵ-DP deep learning (*a.k.a. gradient perturbation*) models on unprivileged groups. By keeping the same hyperparameters as the non-private baseline model, the authors noticed that the accuracy for the unprivileged group dropped more than for the privileged one. Similarly, Ganev et al. [20] have also noticed disparities for the unprivileged group when generating ϵ-DP synthetic data for training ML models by also keeping default hyperparameters of the differentially private generative models. In this paper, we aim to explore to what extent training an ML classifier on ϵ-LDP multidimensional data (*a.k.a. input perturbation*) while fixing the same set of hyperparameters negatively impacts the unprivileged group is valid.

Regarding the local DP setting, the work of Mozannar, Ohannessian, & Srebro [32] was the first one to propose a fair classifier when sanitizing only the protected attribute with ϵ-LDP in both training and testing sets. More recently, the work of Chen *et al.* [13] considers a "semi-private" setting in which a small portion of users share their protected attribute with no sanitization and all other users apply an ϵ-LDP protocol. While the two aforementioned research

works [13, 32] answer interesting questions by collecting a single sensitive attribute using only the GRR [23] protocol, we consider in this work multiple sensitive attributes, which reflects real-world data collections, seven ϵ-LDP protocols, and several fairness and utility metrics. In addition, we also propose a new privacy budget splitting solution named k-based, which generally leads to better fairness and performance in ML binary classification tasks.

3 Preliminaries and Background

This section briefly reviews the group fairness metrics, LDP, and LDP protocols. The notation used throughout this paper is summarized in Table 1.

Symbol	Description
n	Number of users
$[n]$	Set of integers, $\{1, 2, \ldots, n\}$
$\mathbf{x}_i$	i-th coordinate of vector $\mathbf{x}$
$z = \mathcal{M}(v)$	Protocol $\mathcal{M}$ perturbs v into z under ϵ-LDP
X	Set of "non-sensitive" attributes
A_s	Set of sensitive attributes (**privacy viewpoint**)
A_p	Protected attribute (**fairness viewpoint**), $A_p \in A_s$
Z_s	Set of locally differentially private sensitive attributes, $Z_s = \mathcal{M}(A_s)$
k_j	Domain size of the j-th attribute
d_s	Number of sensitive attributes, $d_s = \lvert A_s \rvert$
Y	Set of target values, $Y = \{0, 1\}$
D	Original dataset, $D = (X, A_s, Y)$
D_z	Dataset with sanitized sensitive attributes, $D_z = (X, Z_s, Y)$

Table 1: Notations

Note that in this work, we always consider a single protected attribute and assess fairness w.r.t. that attribute. For LDP, we consider a set of sensitive attributes instead. Moreover, the protected attribute is always considered sensitive, but the opposite is untrue.

3.1 Group Fairness Metrics

In this paper, we focus on group fairness metrics, which assess the fairness of ML models for different demographic groups that differ by the protected attribute (*e.g.*, race, gender, age, ...). Let A_p be the protected attribute, $\hat{Y}$ be a predictor of a binary target $Y \in \{0, 1\}$. The metrics we use to evaluate fairness are:

- **Disparate Impact (DI)** [9]**.** DI is defined as the ratio of the proportion of positive predictions ($\hat{Y} = 1$) for the *unprivileged* group ($A_p = 0$) over the ratio of the proportion of positive predictions for the *privileged* group ($A_p = 1$). The formula for DI is:

$$\mathrm{DI} = \frac{\Pr[\hat{Y} = 1 | A_p = 0]}{\Pr[\hat{Y} = 1 | A_p = 1]}. \tag{1}$$

Note that a perfect DI value is equal to 1.

- **Statistical Parity Difference (SPD) [4].** Instead of the ratio, SDP computes the difference in the proportion of positive predictions for *unprivileged* and *privileged* groups and is defined as:

$$\mathrm{SPD} = \Pr[\hat{Y} = 1 | A_p = 1] - \Pr[\hat{Y} = 1 | A_p = 0]. \tag{2}$$

A perfect SPD value is equal to 0.

- **Equal Opportunity Difference (EOD) [21].** EOD measures the difference between the true positive rates (*i.e.*, recall) of the *unprivileged* group and the *privileged* groups. Formally, EOD is defined as:

$$\mathrm{EOD} = \Pr[\hat{Y} = 1 | Y = 1, A_p = 1] - \Pr[\hat{Y} = 1 | Y = 1, A_p = 0]. \tag{3}$$

A perfect EOD value is equal to 0.

- **Overall Accuracy Difference (OAD) [12].** OAD measures the difference between the overall accuracy rates between the *privileged* group and the *unprivileged* group. Formally, OAD is represented as:

$$\mathrm{OAD} = \Pr[\hat{Y} = Y | A_p = 1] - \Pr[\hat{Y} = Y | A_p = 0]. \tag{4}$$

A perfect OAD value is equal to 0.

3.2 Local Differential Privacy

In this article, we use LDP [25] as the privacy model, which is formalized as:

Definition 1 (ϵ-Local Differential Privacy). *A randomized algorithm $\mathcal{M}$ satisfies ϵ-local-differential-privacy (ϵ-LDP), where $\epsilon > 0$, if for any pair of input values $v_1, v_2 \in Domain(\mathcal{M})$ and any possible output z of $\mathcal{M}$:*

$$\Pr[\mathcal{M}(v_1) = z] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(v_2) = z].$$

Proposition 1 (Post-Processing [17]). *If $\mathcal{M}$ is ϵ-LDP, then for any function f, the composition of $\mathcal{M}$ and f,* i.e., *$f(\mathcal{M})$ satisfies ϵ-LDP.*

Proposition 2 (Sequential Composition [17]). *Let $\mathcal{M}_1$ be an ϵ_1-LDP protocol and $\mathcal{M}_2$ be an ϵ_2-LDP protocol. Then, the protocol $\mathcal{M}_{1,2}(v) = (\mathcal{M}_1(v), \mathcal{M}_2(v))$ is $(\epsilon_1 + \epsilon_2)$-LDP.*

3.3 LDP Protocols

Let $A_s = \{v_1, \ldots, v_k\}$ be a sensitive attribute with a discrete domain of size $k = |A_s|$, in this subsection, we briefly review seven state-of-the-art LDP protocols.

Generalized Randomized Response (GRR) GRR [23] uses no particular encoding. Given a value $v \in A_s$, $GRR(v)$ outputs the true value v with probability p, and any other value $v' \in A_s \setminus \{v\}$, otherwise. More formally:

$$\forall z \in A_s: \quad \Pr[z = a] = \begin{cases} p = \frac{e^\epsilon}{e^\epsilon + k - 1} \text{ if } z = a \\ q = \frac{1}{e^\epsilon + k - 1} \text{ otherwise,} \end{cases}$$

in which z is the perturbed value sent to the server.

Binary Local Hashing (BLH) Local Hashing (LH) protocols [10, 39] can handle a large domain size k by first using hash functions to map an input value to a smaller domain of size g (typically $2 \leq g \ll k$), and then applying GRR to the hashed value. Let $\mathscr{H}$ be a universal hash function family such that each hash function $H \in \mathscr{H}$ hashes a value in A_s into $[g]$, *i.e.*, $H : A_s \to [g]$. With BLH, $[g] = \{0, 1\}$, each user selects at random one hash function H, calculates $b = H(v)$, and perturbs b to z as:

$$\Pr[z = 1] = \begin{cases} p = \frac{e^\epsilon}{e^\epsilon + 1} \text{ if } b = 1 \\ q = \frac{1}{e^\epsilon + 1} \text{ if } b = 0. \end{cases}$$

The user sends the tuple $\langle H, z \rangle$, *i.e.*, the hash function and the perturbed value. Thus, for each user, the server can calculate $S(\langle H, z \rangle) = \{v | H(v) = z\}$.

Optimal LH (OLH) To improve the utility of LH protocols, Wang *et al.* [39] proposed OLH in which the output space of the hash functions in family $\mathscr{H}$ is no longer binary as in BLH. Thus, with OLH, $g = \lfloor e^\epsilon + 1 \rceil$, each user selects at random one hash function H, calculates $b = H(v)$, and perturbs b to z as:

$$\forall i \in [g]: \quad \Pr[z = i] = \begin{cases} p = \frac{e^\epsilon}{e^\epsilon + g - 1} \text{ if } b = i \\ q = \frac{1}{e^\epsilon + g - 1} \text{ if } b \neq i. \end{cases}$$

Similar to BLH, the user sends the tuple $\langle H, z \rangle$ and, for each user, the server can calculate $S(\langle H, z \rangle) = \{v | H(v) = z\}$.

RAPPOR The RAPPOR [18] protocol uses One-Hot Encoding (OHE) to interpret the user's input $v \in A_s$ as a one-hot k-dimensional vector. More precisely, $\mathbf{v} = OHE(v)$ is a binary vector with only the bit at position v set to 1 and the other bits set to 0. Then, RAPPOR randomizes the bits from $\mathbf{v}$ independently to generate $\mathbf{z}$ as follows:

$$\forall i \in [k]: \quad \Pr[\mathbf{z}_i = 1] = \begin{cases} p = \frac{e^{\epsilon/2}}{e^{\epsilon/2} + 1} \text{ if } \mathbf{v}_i = 1, \\ q = \frac{1}{e^{\epsilon/2} + 1} \text{ if } \mathbf{v}_i = 0, \end{cases}$$

where $p + q = 1$ (*i.e.*, symmetric). Afterwards, the user sends $\mathbf{z}$ to the server.

Optimal Unary Encoding (OUE) To minimize the variance of RAPPOR, Wang *et al.* [39] proposed OUE, which perturbs the 0 and 1 bits asymmetrically, *i.e.*, $p + q \neq 1$. Thus, OUE generates $\mathbf{z}$ by perturbing $\mathbf{v}$ as follows:

$$\forall i \in [k]: \quad \Pr[\mathbf{z}_i = 1] = \begin{cases} p = \frac{1}{2} & \text{if } \mathbf{v}_i = 1, \\ q = \frac{1}{e^\epsilon + 1} & \text{if } \mathbf{v}_i = 0. \end{cases}$$

Afterwards, the user sends $\mathbf{z}$ to the server.

Subset Selection (SS) The SS [38, 41] protocol randomly selects $1 \leq \omega \leq k$ items within the input domain to report a subset of values $\Omega \subseteq A_s$. The user's true value v has higher probability of being included in the subset Ω, compared to the other values in $A_s \setminus \{v\}$. The optimal subset size that minimizes the variance is $\omega = \lfloor \frac{k}{e^\epsilon + 1} \rceil$. Given a value $v \in A_s$, $SS(v)$ starts by initializing an empty subset Ω. Afterwards, the true value v is added to Ω with probability $p = \frac{\omega e^\epsilon}{\omega e^\epsilon + k - \omega}$. Finally, it adds values to Ω as follows:

- If $v \in \Omega$, then $\omega - 1$ values are sampled from $A_s \setminus \{v\}$ uniformly at random (without replacement) and are added to Ω;
- If $v \notin \Omega$, then ω values are sampled from $A_s \setminus \{v\}$ uniformly at random (without replacement) and are added to Ω.

Afterwards, the user sends the subset Ω to the server.

Thresholding with Histogram Encoding (THE) Histogram Encoding (HE) [39] encodes the user value as a one-hot k-dimensional histogram, *i.e.*, $\mathbf{v} = [0.0, 0.0, \ldots, 1.0, 0.0, \ldots, 0.0]$ in which only the v-th component is 1.0. $HE(\mathbf{v})$ perturbs each bit of $\mathbf{v}$ independently using the Laplace mechanism [16]. Two different input values $v_1, v_2 \in A_s$ will result in two vectors with L1 distance of $\Delta = 2$. Thus, HE will output $\mathbf{z}$ such that $\mathbf{z}_i = \mathbf{v}_i + \text{Lap}\left(\frac{2}{\epsilon}\right)$. To improve the utility of HE, Wang *et al.* [39] proposed THE such that the user reports (or the server computes): $S(\mathbf{z}) = \{v \mid \mathbf{z}_v > \theta\}$, in which θ is the threshold with optimal value in $(0.5, 1)$. In this work, we use `scipy.minimize_scalar` to optimize θ for a fixed ϵ as: $\min_{\theta \in (0.5,1)} \frac{2e^{\epsilon\theta/2} - 1}{(1 + e^{\epsilon(\theta - 1/2)} - 2e^{\epsilon\theta/2})^2}$.

4 Problem Setting and Methodology

We consider the scenario in which the server collects a set of multiple sensitive attributes A_s under ϵ-LDP guarantees from n distributed users $U = \{u_1, \ldots, u_n\}$. Furthermore, in addition to the LDP-based multidimensional data, we assume that the users will also provide non-sanitized data X, which we consider as "non-sensitive" attributes. The server aims to use both sanitized $Z_s = \mathcal{M}(A_s)$ and non-sanitized data X to train an ML classifier with a binary target variable $Y = \{0, 1\}$. Notice, however, that we will be training an ML classifier on

$D_z = (X, Z_s, Y)$ but testing on $D = (X, A_s, Y)$ as the main goal is to *protect the privacy of the data used to train the ML model* (*e.g.*, to avoid membership inference attacks [22], reconstruction attacks [35], and other privacy threats [28]). In other words, instead of considering a system for on-the-fly LDP sanitization of test data, as in [32], we only sanitize the training set.

With these elements in mind, our primary goal is to study the impact of training an ML classifier on $D_z = (X, Z_s, Y)$ compared to $D = (X, A_s, Y)$ on fairness and utility, using different LDP protocols and privacy budget splitting solutions. More precisely, we consider the setting where each sensitive attribute in A_s is collected independently under LDP guarantees. In this case, to satisfy ϵ-LDP following Proposition 2, the privacy budget ϵ must be split among the total number of sensitive attributes $d_s = |A_s|$. To this end, the state-of-the-art [6,37] solution, named **uniform**, propose to split the privacy budget ϵ uniformly among all attributes, *i.e.*, allocating $\frac{\epsilon}{d_s}$ for each attribute. However, as different sensitive attributes have different domain sizes k_j, for $j \in [d_s]$, we propose a new solution named **k-based** that splits the privacy budget ϵ proportionally to the domain size of the attribute. That is, for the j-th attribute, we will allocate $\epsilon_j = \frac{\epsilon \cdot k_j}{\sum_{i=1}^{d_s} k_i}$.

In addition, each LDP protocol has a different way of encoding and perturbing user's data. We thus propose to compare all LDP protocols under the same encoding when training the ML classifier. More specifically, we will use OHE and Indicator Vector Encoding (IVE) [1] as all LDP protocols from Section 3.3 are designed for categorical data or discrete data with known domain. For example, let Ω be the reported subset of a user after using SS as LDP protocol. Following IVE, we create a binary vector $\mathbf{z} = [b_1, \ldots, b_k] \in \{0,1\}^k$ of length k, where the v-th entry is set to 1 if $v \in \Omega$, and 0, otherwise. In other words, $\mathbf{z}$ represents the subset Ω in a binary format. Fig. 1 illustrates the LDP encoding and perturbation at the user side and how to achieve a "homogeneous encoding" for all the seven LDP protocols at the server side. Last, all "non-sensitive" attributes X are encoded using OHE.

5 Experimental Evaluation

In this section, we present our experiments' setting and main results. Supplementary results can be found in Appendix A. Our main Research Questions (RQ) are:

- **RQ1.** Overall, how does preprocessing multidimensional data with ϵ-LDP affect the fairness and utility of ML binary classifiers with the same hyperparameters used before and after sanitization?
- **RQ2.** Which privacy budget-splitting solution leads to less harm to the fairness and utility of an ML binary classifier?
- **RQ3.** How do different LDP protocols affect the fairness and utility of an ML binary classifier, and which one is more suitable for the different real-world scenarios applied?

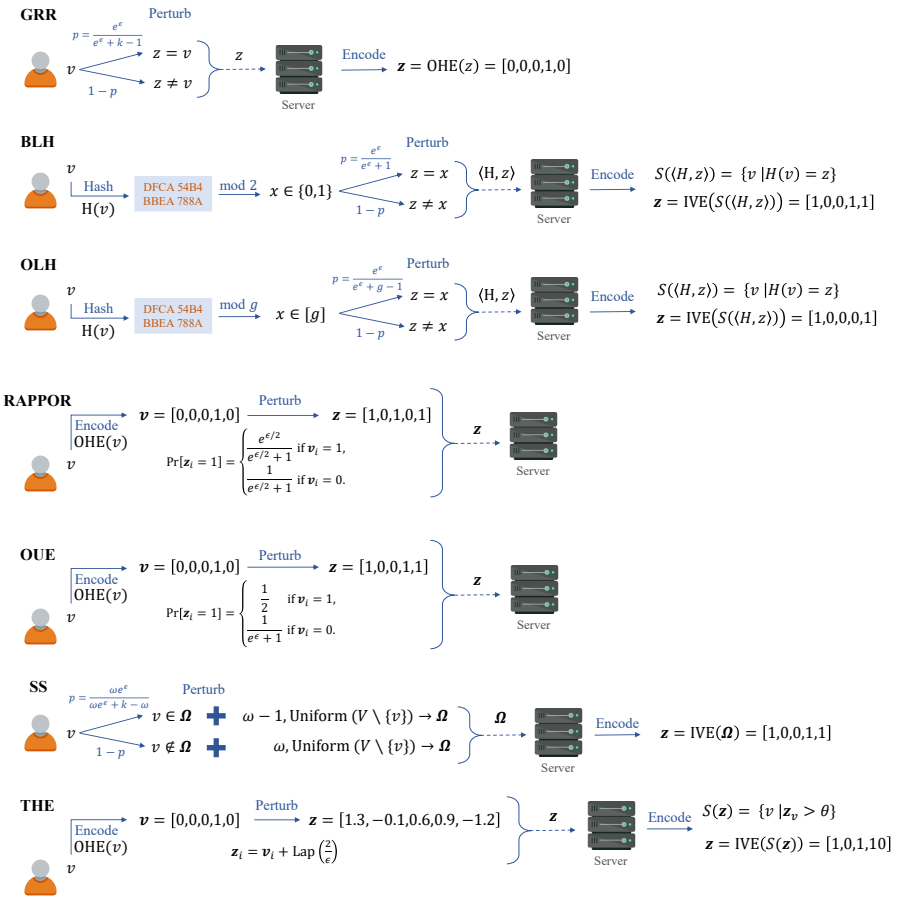

Fig. 1: Overview of client-side encoding and perturbation steps for the seven different LDP protocols applied. On the server side, there is also a post-processing step with one-hot encoding (OHE) or indicator vector encoding (IVE), if needed.

5.1 Setup of Experiments

General setting. For all experiments, we consider the following setting:

- **Environment.** All algorithms are implemented in Python 3 with Numpy [36], Numba [27], and Multi-Freq-LDPy [7] libraries, and run on a local machine with 2.50GHz Intel Core i9 and 64GB RAM. The codes we develop for all experiments are available in a **GitHub repository** [2].
- **ML classifier.** We used the state-of-the-art[3] LGBM [26] as predictor $\hat{Y}$.
- **Encoding.** We only use discrete and categorical attributes, which are encoded using OHE or IVE (see Fig. 1) and the target is binary, *i.e.*, $Y \in \{0, 1\}$.

[3] https://www.kaggle.com/kaggle-survey-2022.

– **Training and testing sets.** We randomly select 80% as training set and the remaining 20% as testing set. We apply LDP on the training set only. That is, the samples in the testing set are the original samples (*i.e.*, no LDP).
– **Stability.** Since LDP protocols, train/test splitting, and ML algorithms are randomized, we report average results over 20 runs.

Datasets. Table 2 summarizes all datasets used in our experiments. For ease of reproducibility, we use real-world and open datasets.

Table 2: Description of the datasets used in the experiments.

Dataset	n	A_p	A_s, *domain size* k	Y
Adult	45849	gender	- gender, $k = 2$	income
			- race, $k = 5$	
			- native country, $k = 41$	
			- age, $k = 74$	
ACSCoverage	98739	DIS	- DIS, $k = 2$	PUBCOV
			- AGEP, $k = 50$	
			- SEX, $k = 2$	
			- SCHL, $k = 24$	
LSAC	20427	race	- race, $k = 2$	pass bar
			- gender, $k = 2$	
			- family income, $k = 5$	
			- full time, $k = 2$	

– **Adult.** We use 26000 as threshold to binarize the target variable "income" of the *reconstructed Adult* dataset [14]. After cleaning, $n = 45849$ samples are kept. We excluded "capital-gain" and "capital-loss" and used the remaining 10 discrete and categorical attributes. We considered $A_s = \{$gender, race, native-country, age$\}$ as sensitive attributes for LDP sanitization and $A_p =$ gender as the protected attribute for fairness assessment.
– **ACSCoverage.** This dataset[4] is retrieved with the `folktables` [14] Python package and the binary target "PUBCOV" designates whether an individual is covered by public health insurance or not. We select the year 2018 and the "Texas" state, with $n = 98739$ samples. We removed "DEAR", "DEYE", "DREM", and "PINCP" and used the remaining 15 discrete and categorical attributes. We considered $A_s = \{$DIS, AGEP, SEX, SCHL$\}$ as sensitive attributes for LDP sanitization and $A_p =$ DIS as the protected attribute (*i.e.*, disability) for fairness assessment.

[4] The full documentation for the description of all attributes is in https://www.census.gov/programs-surveys/acs/microdata/documentation.html.

– **LSAC.** This dataset is from the Law School Admissions Council (LSAC) National Bar Passage Study [40] and the binary target "pass_bar" indicates whether or not a candidate has passed the bar exam. After cleaning, $n = 20427$ samples are kept. We only consider as attributes: 'gender', 'race', 'family income', 'full time', 'undergrad GPA score' (discretized to $\{1.5, 2.0, ..., 4.5\}$), and 'LSAT score' (rounded to the closest integer). The 'race' attribute was binarized to {black, other}. We set $A_s = \{$race, gender, family income, full time$\}$ as sensitive attributes for LDP sanitization and $A_p =$ race as the protected attribute for fairness assessment.

Evaluated methods. The methods we use and compare are:

– **(Baseline) NonDP.** This is our baseline with LGBM trained over original data (*i.e.*, $D = (X, A_s, Y)$). We searched for the best hyperparameters using Bayesian optimization [11] through 100 iterations varying: $max\_depth \in [3, 50]$, $n\_estimators \in [50, 2000]$, and $learning\_rate \in (0.01, 0.25)$;
– **LDP protocols.** We pre-processed $Z_s = \mathcal{M}(A_s)$ of the training sets using all seven LDP protocols from Section 3.3 (*i.e.*, GRR, RAPPOR, OUE, SS, BLH, OLH, and THE) as $\mathcal{M}$. We used the best hyperparameters found for the NonDP model and trained LGBM over $D_z = (X, Z_s, Y)$. For all datasets, we set d_s to 4. That is, $d_s = |A_s| = 4$. To satisfy ϵ-LDP (*cf.* Definition 2), we split the privacy budget ϵ following the two solutions described in Section 4 (*i.e.*, the state-of-the-art uniform and our k-based solution).

Metrics. We evaluate the performance of LGBM trained over the original data (*i.e.*, NonDP baseline) and LDP-based data on privacy, utility, and fairness:

– **Privacy.** We vary the privacy parameter in the range of $\epsilon = \{0.25, 0.5, 1, 2, 4, 8, 10, 20, 50\}$. At $\epsilon = 0.25$ the ratio of probabilities is bounded by $e^{0.25} \approx 1.3$ giving nearly indistinguishable distributions, whereas at $\epsilon = 50$ almost no privacy is guaranteed.
– **Utility.** We use accuracy (acc), f1-score (f1), area under the receiver operating characteristic curve (auc), and recall as utility metrics;
– **Fairness.** We use the metrics of Section 3.1 (*i.e.*, DI, SPD, EOD, and OAD).

5.2 Main Results

LDP impact on fairness. Fig. 2 (Adult), Fig. 3 (ACSCoverage), and Fig. 4 (LSAC) illustrate the privacy-fairness trade-off for the NonDP baseline and all the seven LDP protocols, considering both uniform and our k-based privacy budget splitting solutions. From these figures, one can notice that fairness is, in general, slightly improved for all seven LDP protocols under both the uniform and the k-based solution. For instance, for the DI metric in Fig. 2, the NonDP data indicates a value of 0.44 showing discrimination against women and, by applying LDP protocols, DI tended to increase to ∼0.48 (with $\epsilon = 0.25$) resulting in a slight improvement in fairness. Similarly, SPD decreased from 0.37

to ∼0.34 after applying LDP protocols. The same behavior is obtained for EOD. The exception was in Fig. 3 for the OAD metric in which the gap between privileged and unprivileged groups was accentuated (favoring the unprivileged group). More specifically, the NonDP baseline has OAD equal to -0.17, and after satisfying LDP for both uniform and k-based solutions and using all LDP protocols, the gap between the privileged and unprivileged groups increased to -0.3. In other words, we start with favoritism towards the unprivileged group (negative value) and this favoritism increased after LDP.

Note also that when applying the uniform privacy budget splitting solution (see left-side plots), all fairness metrics were less robust to LDP than our k-based solution and, thus, returned to the NonDP baseline value in low privacy regimes. With our k-based solution (see right-side plots), all fairness metrics continued to be slightly better for all privacy regimes for the Adult dataset in Fig. 2. For the ACSCoverage dataset, not all fairness metrics returned to the NonDP baseline value and for the LSAC dataset, a similar behavior was noticed for both uniform and k-based solutions. These differences are mainly influenced by the domain size k of the sensitive attributes. For instance, while Adult has sensitive attributes with higher values of k, LSAC has many binary sensitive attributes.

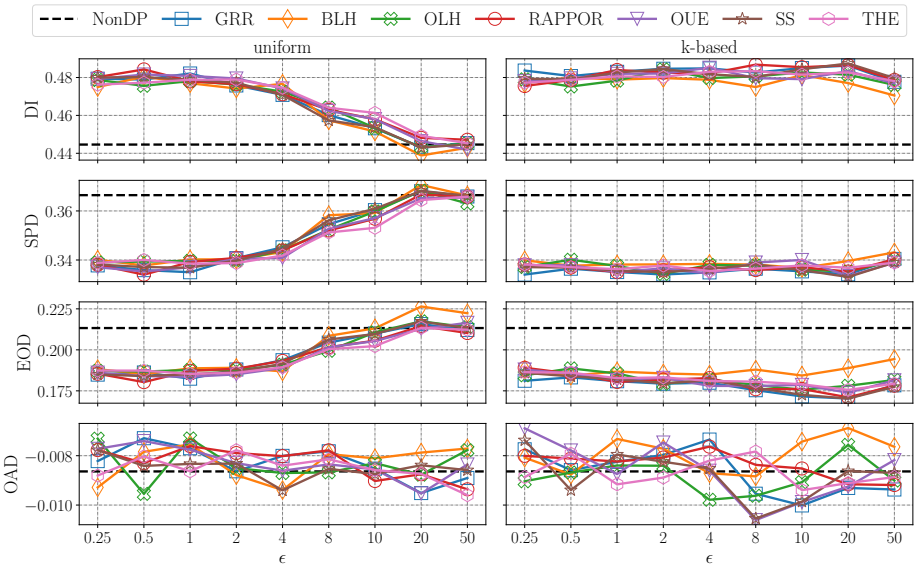

Fig. 2: Fairness metrics (y-axis) by varying the privacy guarantees (x-axis), the ϵ-LDP protocol, and the privacy budget splitting solution (*i.e.*, uniform on the left-side and our k-based on the right-side), on the Adult [14] dataset.

LDP impact on utility. Fig. 5 (Adult), Fig. 6 (ACSCoverage), and Fig. 7 (LSAC) illustrate the privacy-utility trade-off for the NonDP baseline and all

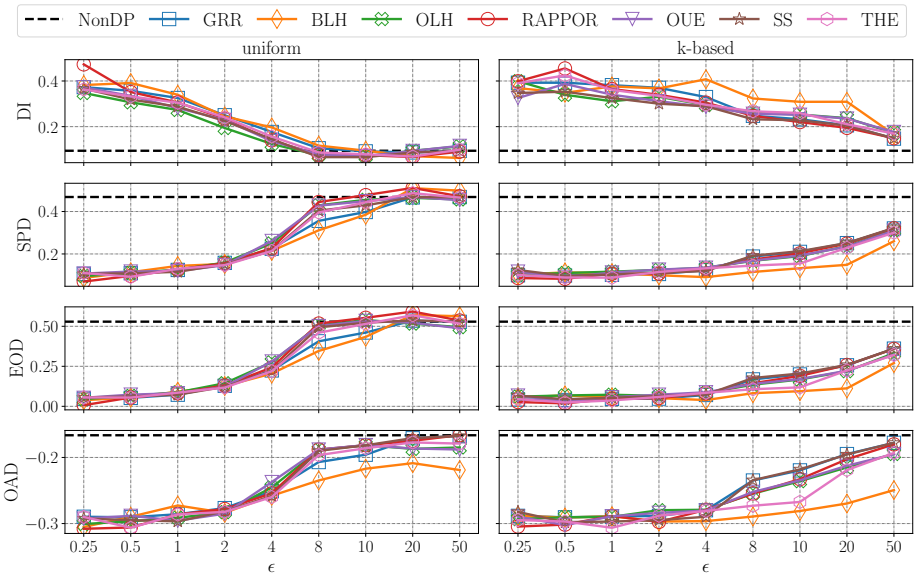

Fig. 3: Fairness metrics (y-axis) by varying the privacy guarantees (x-axis), the ϵ-LDP protocol, and the privacy budget splitting solution (*i.e.*, uniform on the left-side and our k-based on the right-side), on the ACSCoverage [14] dataset.

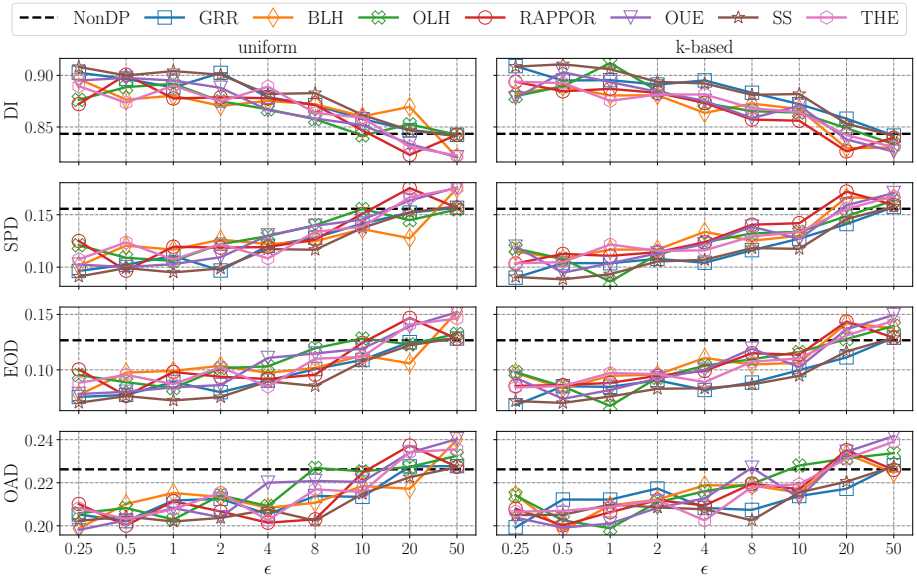

Fig. 4: Fairness metrics (y-axis) by varying the privacy guarantees (x-axis), the ϵ-LDP protocol, and the privacy budget splitting solution (*i.e.*, uniform on the left-side and our k-based on the right-side), on the LSAC [40] dataset.

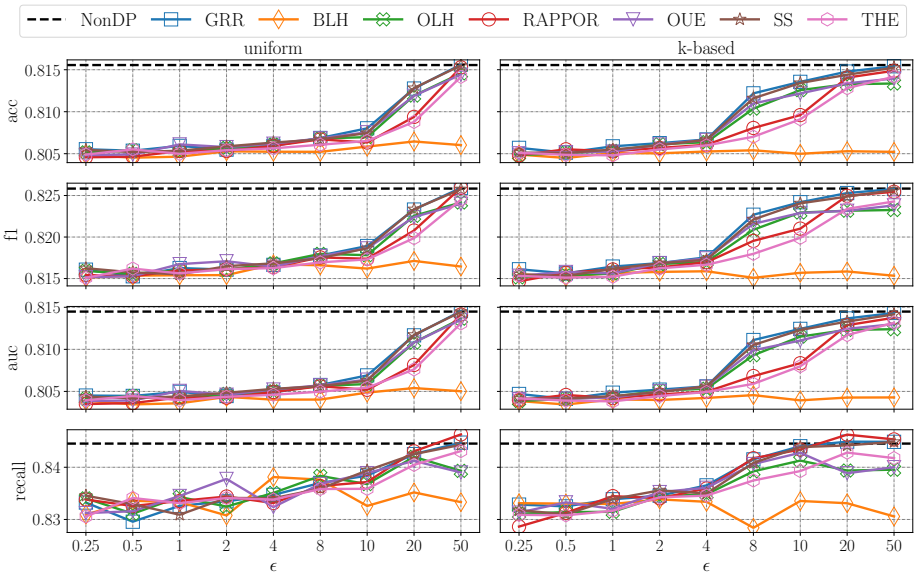

Fig. 5: Utility metrics (y-axis) by varying the privacy guarantees (x-axis), the ϵ-LDP protocol, and the privacy budget splitting solution (*i.e.*, uniform on the left-side and our k-based on the right-side), on the Adult [14] dataset.

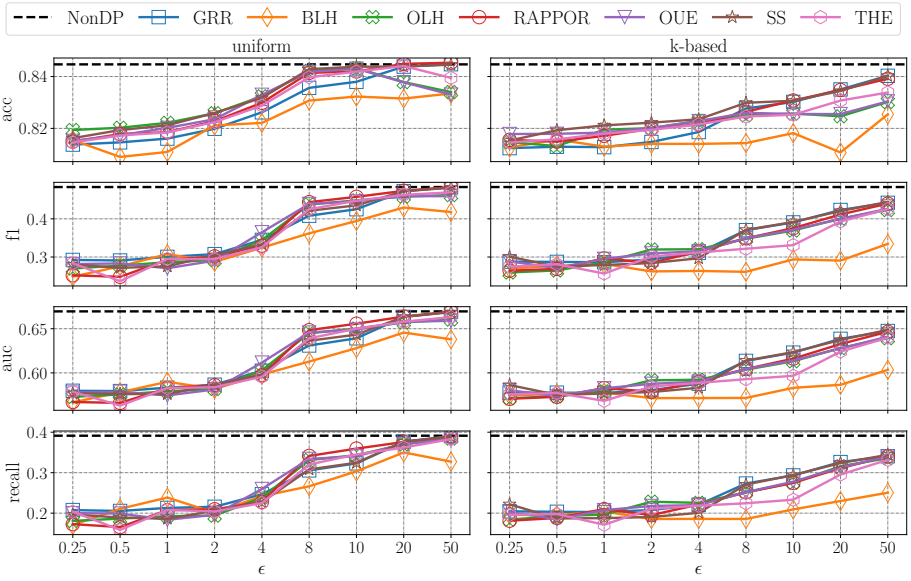

Fig. 6: Utility metrics (y-axis) by varying the privacy guarantees (x-axis), the ϵ-LDP protocol, and the privacy budget splitting solution (*i.e.*, uniform on the left-side and our k-based on the right-side), on the ACSCoverage [14] dataset.

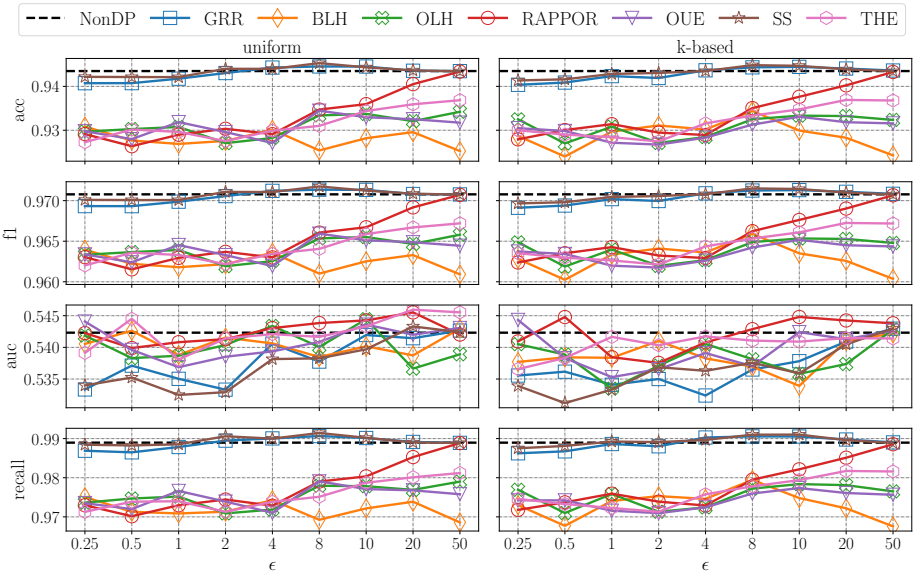

Fig. 7: Utility metrics (y-axis) by varying the privacy guarantees (x-axis), the ϵ-LDP protocol, and the privacy budget splitting solution (*i.e.*, uniform on the left-side and our k-based on the right-side), on the LSAC [40] dataset.

the seven LDP protocols, considering both uniform and our k-based privacy budget splitting solutions. From these figures, one can note that, in general, the impact of ϵ-LDP on utility metrics is minor. For instance, for the Adult dataset (Fig. 5), only $\sim 1\%$ of utility loss for all metrics is observed. Regarding privacy budget splitting, for the Adult dataset, our k-based solution is more robust to LDP as it only drops in higher privacy regimes (*i.e.*, smaller ϵ values) than the uniform solution. One main explanation for this behavior is because there is more discrepancy in the domain size k's of the sensitive attributes A_s and, consequently, more privacy budget ϵ are allocated to those attributes with high k. For this reason, the uniform solution preserved more utility for the ACSCoverage dataset in Fig. 6, and both solutions had similar results for the LSAC dataset in Fig. 7 due to sensitive attributes with small domain size k.

Summary. We summarize our main findings for the three research questions formulated at the beginning of Section 5. We highlight these findings are generic and were also confirmed in additional experiments presented in Appendix A. **(RQ1)** Using the same hypeparameters configuration, ϵ-LDP positively affects fairness in ML (see Figs. 2–4) while having a negligible impact on model's utility (see Figs. 5–7). This contrasts the findings of [8, 20] that state that under the same hyperparameters configuration, ϵ-DP negatively impacts fairness. Although the aforementioned research works concern *gradient perturbation* in central DP, the recent work of de Oliveira *et al.* [33] has shown that when searching for the best

hyperparameters for both non-private and DP models, the ϵ-DP impact on fairness is negligible. In our case, we focused on *input perturbation*, *i.e.*, randomizing multiple sensitive attributes before training any ML algorithm, and discovered a positive impact of ϵ-(L)DP on fairness. **(RQ2)** Our k-based solution consistently led to better fairness than the state-of-the-art uniform solution when there exist sensitive attributes with high domain size k (*e.g.*, for both Adult and ACSCoverage datasets). Naturally, when all sensitive attributes have a binary domain, our k-based solution is equivalent to the uniform solution. For this reason, both state-of-the-art uniform and our k-based solution led to similar privacy-utility-fairness trade-off for the LSAC dataset (see Figs. 4 and 7). Therefore, regarding utility, k-based is better when sensitive attributes have higher domain sizes k, which coincides with real-world data collections. **(RQ3)** In general, GRR and SS presented the best privacy-utility-fairness trade-off for all three datasets. This is because GRR has only one perturbed output value and because SS is equivalent to GRR when $\omega = 1$, thus, not introducing inconsistencies for a user's profile. The term *inconsistency* refers to an user being multiple categories in a given attribute, *i.e.*, being both woman and man at the same time. In fact, this is precisely what happens with UE protocols that perturb each bit independently or with LH protocols in which many values can hash to the same perturbed value. For this reason, since BLH hashes the input set $V \to \{0, 1\}$, it consistently presented the worst utility results for all three datasets, and only for ACSCoverage (see Fig. 3), it presented slightly better fairness results than all other LDP protocols.

6 Conclusion and Perspectives

This paper presented an in-depth empirical study of the impact of pre-processing multidimensional data with seven state-of-the-art ϵ-LDP protocols on fairness and utility in binary classification tasks. In our experiments, GRR [23] and SS [38, 41] presented the best privacy-utility-fairness trade-off than RAPPOR [18], OUE [39], THE [39], BLH [10], and OLH [39]. In addition, we proposed a new privacy budget splitting solution named k-based, which generally led to better fairness and performance results than the state-of-the-art solution that splits ϵ uniformly [6, 37]. Globally, while previous research [8, 20] has highlighted that DP worsens fairness in ML under the same hyperparameter configuration, our study finds that LDP slightly improves fairness and does not significantly impair utility. Indeed, there is still much to explore in the area of privacy-fairness-aware ML, and this study's empirical results can serve as a basis for future research directions. For instance, we intend to formally investigate the privacy-utility-fairness trade-off on binary classification tasks when varying the distribution of the protected attribute, the target, and their joint, and propose new methods accordingly. Last, we plan to investigate the impact of LDP pre-processing on different ML algorithms, such as deep neural networks.

Acknowledgements This work was supported by the European Research Council (ERC) project HYPATIA under the European Union's Horizon 2020 research and innovation programme. Grant agreement n. 835294.

References

1. Indicator vector, available online: https://en.wikipedia.org/wiki/Indicator_vector (accessed on 04 April 2023)
2. LDP impact on fairness repository, https://github.com/hharcolezi/ldp-fairness-impact
3. General data protection regulation (GDPR) (2018), available online: https://gdpr-info.eu/ (accessed on 26 March 2023)
4. Agarwal, A., Agarwal, H., Agarwal, N.: Fairness score and process standardization: framework for fairness certification in artificial intelligence systems. AI and Ethics pp. 1–13 (2022)
5. Alves, G., Bernier, F., Couceiro, M., Makhlouf, K., Palamidessi, C., Zhioua, S.: Survey on fairness notions and related tensions. arXiv preprint arXiv:2209.13012 (2022)
6. Arcolezi, H.H., Couchot, J.F., Al Bouna, B., Xiao, X.: Random sampling plus fake data: Multidimensional frequency estimates with local differential privacy. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. p. 47–57. CIKM '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3459637.3482467
7. Arcolezi, H.H., Couchot, J.F., Gambs, S., Palamidessi, C., Zolfaghari, M.: Multi-freq-ldpy: Multiple frequency estimation under local differential privacy in python. In: Atluri, V., Di Pietro, R., Jensen, C.D., Meng, W. (eds.) Computer Security – ESORICS 2022. pp. 770–775. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-17143-7_40
8. Bagdasaryan, E., Poursaeed, O., Shmatikov, V.: Differential privacy has disparate impact on model accuracy. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
9. Barocas, S., Selbst, A.D.: Big data's disparate impact. Calif. L. Rev. **104**, 671 (2016)
10. Bassily, R., Smith, A.: Local, private, efficient protocols for succinct histograms. In: Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing. p. 127–135. STOC '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2746539.2746632
11. Bergstra, J., Yamins, D., Cox, D.D.: Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In: Proceedings of the 30th International Conference on International Conference on Machine Learning. p. I–115–I–123. ICML'13, JMLR (2013)
12. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research (2018)
13. Chen, C., Liang, Y., Xu, X., Xie, S., Hong, Y., Shu, K.: On fair classification with mostly private sensitive attributes. arXiv preprint arXiv:2207.08336 (2022)
14. Ding, F., Hardt, M., Miller, J., Schmidt, L.: Retiring adult: New datasets for fair machine learning. Advances in Neural Information Processing Systems **34** (2021)

15. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. ACM (Jan 2012). https://doi.org/10.1145/2090236.2090255
16. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Theory of Cryptography, pp. 265–284. Springer Berlin Heidelberg (2006). https://doi.org/10.1007/11681878_14
17. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science **9**(3–4), 211–407 (2014)
18. Erlingsson, U., Pihur, V., Korolova, A.: RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. pp. 1054–1067. ACM, New York, NY, USA (2014). https://doi.org/10.1145/2660267.2660348
19. Fioretto, F., Tran, C., Hentenryck, P.V., Zhu, K.: Differential privacy and fairness in decisions and learning tasks: A survey. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (Jul 2022). https://doi.org/10.24963/ijcai.2022/766
20. Ganev, G., Oprisanu, B., De Cristofaro, E.: Robin hood and matthew effects: Differential privacy has disparate impact on synthetic data. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 6944–6959. PMLR (17–23 Jul 2022)
21. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
22. Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P.S., Zhang, X.: Membership inference attacks on machine learning: A survey. ACM Computing Surveys **54**(11s), 1–37 (Jan 2022). https://doi.org/10.1145/3523273
23. Kairouz, P., Bonawitz, K., Ramage, D.: Discrete distribution estimation under local privacy. In: International Conference on Machine Learning. pp. 2436–2444. PMLR (2016)
24. Kallus, N., Mao, X., Zhou, A.: Assessing algorithmic fairness with unobserved protected class using data combination. Management Science **68**(3), 1959–1981 (Mar 2022). https://doi.org/10.1287/mnsc.2020.3850
25. Kasiviswanathan, S.P., Lee, H.K., Nissim, K., Raskhodnikova, S., Smith, A.: What can we learn privately? In: 2008 49th Annual IEEE Symposium on Foundations of Computer Science. pp. 531–540 (2008). https://doi.org/10.1109/FOCS.2008.27
26. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
27. Lam, S.K., Pitrou, A., Seibert, S.: Numba: A llvm-based python jit compiler. In: Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC. LLVM '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2833157.2833162
28. Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., Lin, Z.: When machine learning meets privacy. ACM Computing Surveys **54**(2), 1–36 (Mar 2021). https://doi.org/10.1145/3436755
29. Makhlouf, K., Zhioua, S., Palamidessi, C.: Machine learning fairness notions: Bridging the gap with real-world applications. Information Processing & Management **58**(5), 102642 (Sep 2021). https://doi.org/10.1016/j.ipm.2021.102642

30. Makhlouf, K., Zhioua, S., Palamidessi, C.: On the applicability of machine learning fairness notions. ACM SIGKDD Explorations Newsletter **23**(1), 14–23 (May 2021). https://doi.org/10.1145/3468507.3468511
31. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys **54**(6), 1–35 (Jul 2021). https://doi.org/10.1145/3457607
32. Mozannar, H., Ohannessian, M., Srebro, N.: Fair learning with private demographic data. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 7066–7075. PMLR (13–18 Jul 2020)
33. de Oliveira, A.S., Kaplan, C., Mallat, K., Chakraborty, T.: An empirical analysis of fairness notions under differential privacy. PPAI 2023, 4th AAAI Workshop on Privacy-Preserving Artificial Intelligence, 13 February 2023, Washington DC, USA (2023)
34. Pujol, D., McKenna, R., Kuppam, S., Hay, M., Machanavajjhala, A., Miklau, G.: Fair decision making using privacy-protected data. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. ACM (Jan 2020). https://doi.org/10.1145/3351095.3372872
35. Salem, A.M.G., Bhattacharyya, A., Backes, M., Fritz, M., Zhang, Y.: Updates-leak: Data set inference and reconstruction attacks in online learning. In: 29th USENIX Security Symposium. pp. 1291–1308. USENIX (2020)
36. van der Walt, S., Colbert, S.C., Varoquaux, G.: The numpy array: A structure for efficient numerical computation. Computing in Science & Engineering **13**(2), 22–30 (2011). https://doi.org/10.1109/MCSE.2011.37
37. Wang, N., Xiao, X., Yang, Y., Zhao, J., Hui, S.C., Shin, H., Shin, J., Yu, G.: Collecting and analyzing multidimensional data with local differential privacy. In: 2019 IEEE 35th International Conference on Data Engineering (ICDE). IEEE (Apr 2019). https://doi.org/10.1109/icde.2019.00063
38. Wang, S., Huang, L., Wang, P., Nie, Y., Xu, H., Yang, W., Li, X.Y., Qiao, C.: Mutual information optimally local private discrete distribution estimation. arXiv preprint arXiv:1607.08025 (2016)
39. Wang, T., Blocki, J., Li, N., Jha, S.: Locally differentially private protocols for frequency estimation. In: 26th USENIX Security Symposium (USENIX Security 17). pp. 729–745. USENIX Association, Vancouver, BC (Aug 2017)
40. Wightman, L.F.: Lsac national longitudinal bar passage study. lsac research report series. (1998)
41. Ye, M., Barg, A.: Optimal schemes for discrete distribution estimation under locally differential privacy. IEEE Transactions on Information Theory **64**(8), 5662–5676 (2018). https://doi.org/10.1109/TIT.2018.2809790

A Additional Experiments

To validate our findings that LDP can improve fairness without sacrificing much utility, we conducted an additional series of experiments by considering a dynamic number of sensitive attributes d_s. Specifically, for each iteration of the 20 runs (for stability), using a specific dataset such as Adult [14], ACSCoverage [14], or LSAC [40], we randomly determined the number of sensitive attributes $2 \leq d_s \leq 6$, ensuring that the protected attribute A_p is always included in A_s, *i.e.*, $A_p \in A_s$.

Similar to Figs. 2–4 (LDP impact on fairness) and Figs. 5–7 (LDP impact on utility), Figs. 8–13 illustrate the privacy-fairness-utility trade-offs for the Adult, ACSCoverage, and LSAC dataset, respectively. These figures consider the NonDP baseline and the seven LDP protocols, as well as both the uniform and our k-based privacy budget splitting solutions. From Figs. 8–13, one can observe that the results follow similar trends as those presented in Section 5. Specifically, the LDP pre-processing positively affects fairness while only having a minor impact on the utility of the ML model.

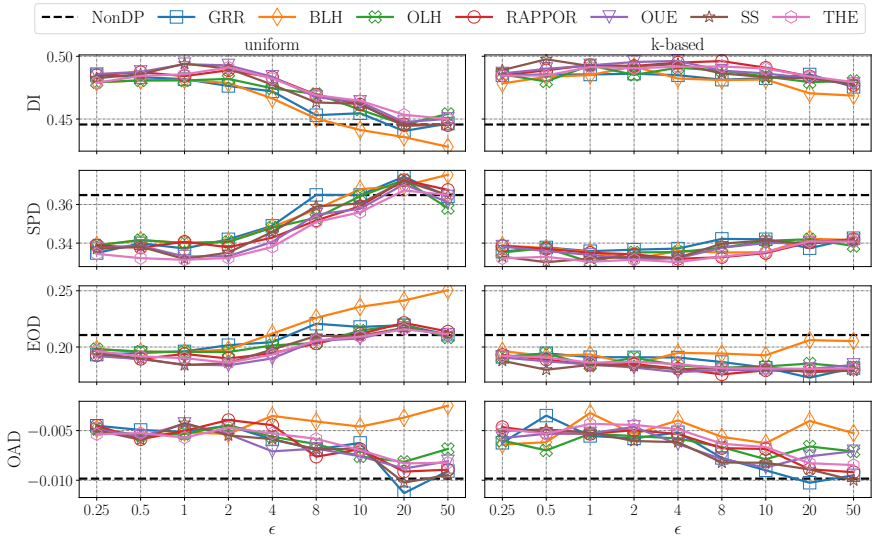

Fig. 8: Fairness metrics (y-axis) by varying the privacy guarantees (x-axis), the ϵ-LDP protocol, and the privacy budget splitting solution (*i.e.*, uniform on the left-side and our k-based on the right-side), on the Adult [14] dataset. The number of sensitive attributes $2 \leq d_s \leq 6$ is selected uniformly at random.

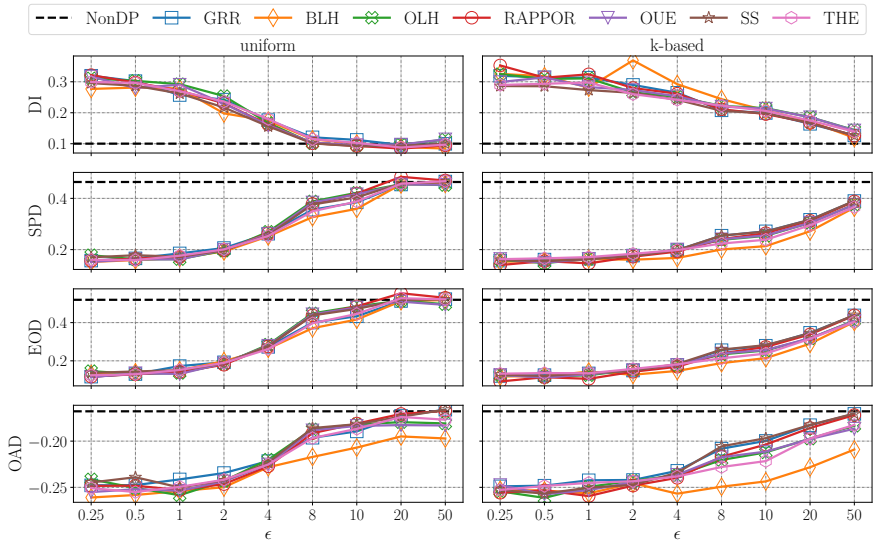

Fig. 9: Fairness metrics (y-axis) by varying the privacy guarantees (x-axis), the ϵ-LDP protocol, and the privacy budget splitting solution (*i.e.*, uniform on the left-side and our k-based on the right-side), on the ACSCoverage [14] dataset. The number of sensitive attributes $2 \leq d_s \leq 6$ is selected uniformly at random.

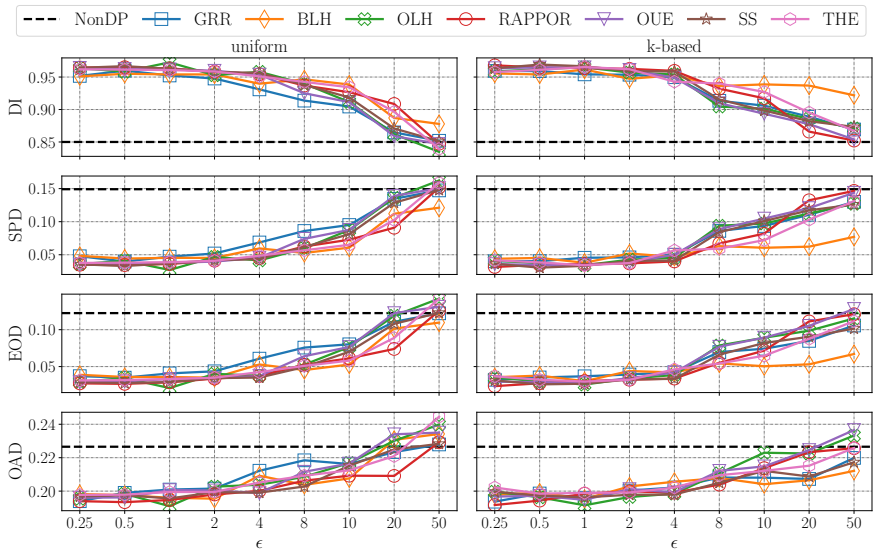

Fig. 10: Fairness metrics (y-axis) by varying the privacy guarantees (x-axis), the ϵ-LDP protocol, and the privacy budget splitting solution (*i.e.*, uniform on the left-side and our k-based on the right-side), on the LSAC [40] dataset. The number of sensitive attributes $2 \leq d_s \leq 6$ is selected uniformly at random.

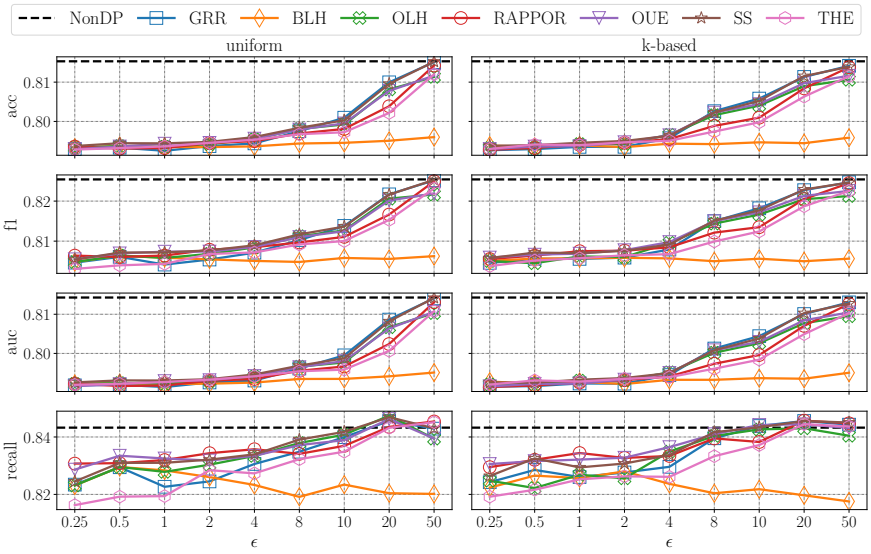

Fig. 11: Utility metrics (y-axis) by varying the privacy guarantees (x-axis), the ϵ-LDP protocol, and the privacy budget splitting solution (*i.e.*, uniform on the left-side and our k-based on the right-side), on the Adult [14] dataset. The number of sensitive attributes $2 \leq d_s \leq 6$ is selected uniformly at random.

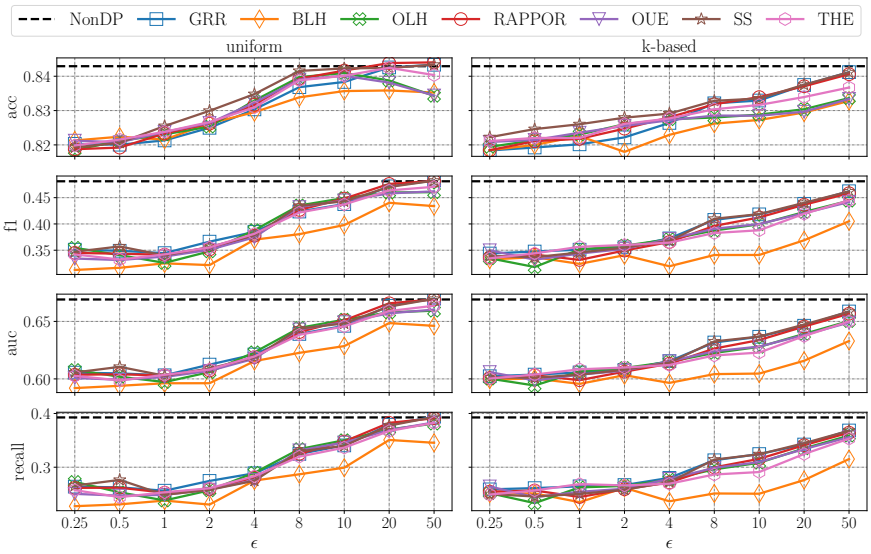

Fig. 12: Utility metrics (y-axis) by varying the privacy guarantees (x-axis), the ϵ-LDP protocol, and the privacy budget splitting solution (*i.e.*, uniform on the left-side and our k-based on the right-side), on the ACSCoverage [14] dataset. The number of sensitive attributes $2 \leq d_s \leq 6$ is selected uniformly at random.

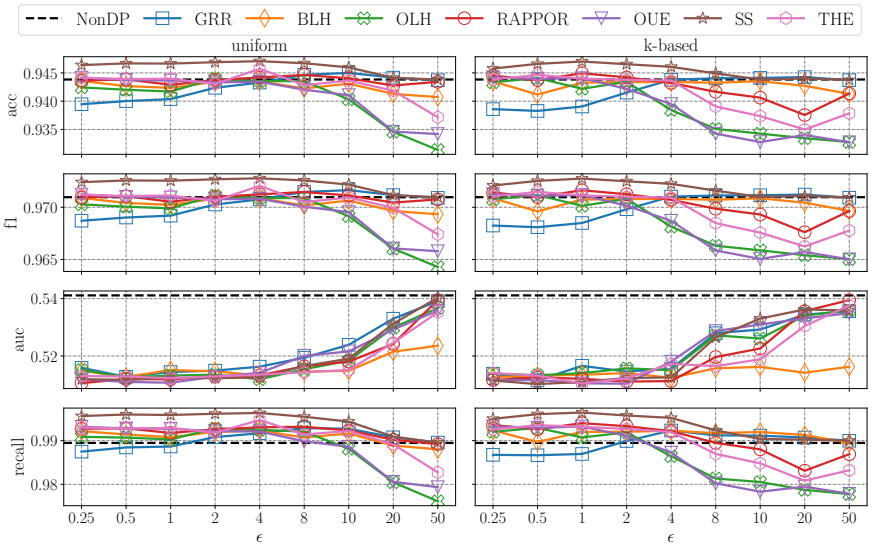

Fig. 13: Utility metrics (y-axis) by varying the privacy guarantees (x-axis), the ϵ-LDP protocol, and the privacy budget splitting solution (*i.e.*, uniform on the left-side and our k-based on the right-side), on the LSAC [40] dataset. The number of sensitive attributes $2 \leq d_s \leq 6$ is selected uniformly at random.