On the Impact of Multi-dimensional Local Differential Privacy on Fairness

Karima Makhlouf, Héber Hwang Arcolezi, Sami Zhioua, Ghassen Ben Brahim, Catuscia Palamidessi

On the Impact of Multi-dimensional Local Differential Privacy on Fairness

Karima Makhlouf[1,3], Héber H. Arcolezi[2], Sami Zhioua[3], Ghassen Ben Brahim[4], Catuscia Palamidessi[1,3]

[1]Inria, Palaiseau, France.
[2]Inria Centre at the University Grenoble Alpes, Grenoble, France.
[3]École Polytechnique (IPP), Palaiseau, France.
[4]College of Computer Engineering and Science, Prince Mohammad Bin Fahd University, Dammam, Saudi Arabia.

Abstract

Automated decision systems are increasingly used to make consequential decisions on people's lives. Due to the sensitivity of the manipulated data as well as the resulting decisions, several ethical concerns need to be addressed for the appropriate use of such technologies, in particular, fairness and privacy. Unlike previous work which focused on centralized differential privacy (DP) or on local DP (LDP) for a single sensitive attribute, in this paper, we examine the impact of LDP in the presence of several sensitive attributes (*i.e., multi-dimensional data*) on fairness. Detailed empirical analysis on synthetic and benchmark datasets revealed very relevant observations. In particular, (1) multi-dimensional LDP is an efficient approach to reduce disparity, (2) the multi-dimensional approach of LDP (independent vs combined) matters only at low privacy guarantees (high $\boldsymbol{\epsilon}$), and (3) the outcome $\boldsymbol{Y}$ distribution has an important effect on which group is more sensitive to the obfuscation. Last, we summarize our findings in the form of recommendations to guide practitioners in adopting effective privacy-preserving practices while maintaining fairness and utility in ML applications.

Keywords: Differential Privacy, Machine learning, Fairness, Randomized Response

1 Introduction

Data collected about individuals is regularly used to make decisions that impact those same individuals. For example, census statistics have important implications for all aspects of daily life, including the allocation of political power, the distribution of federal funds, and research in economics and social sciences. In banking industries, machine learning (ML) models leverage data to proactively monitor customer behavior, reduce the likelihood of false positives, and prevent fraud. In these settings, there is a tension between the need for accurate systems, in which individuals receive what they deserve, and the need to protect individuals from improper disclosure of their sensitive information. Differential privacy (DP) [23] is now widely recognized as the gold standard for providing formal guarantees on the privacy level achieved by an algorithm. However, central DP can only be used on the assumption of a trustworthy server. Local DP (LDP) [32] is a variant that achieves privacy guarantees for each user locally with no assumptions on third-party servers. In other words, LDP ensures that each user's data is locally obfuscated first on the client-side and then sent to the server-side, thus protecting data from privacy leaks on both the client-side and the server-side. Many Big tech companies have deployed LDP-based algorithms to use in their industrial products (e.g., Google Chrome [24] and Apple iOS [4]).

On the other hand, algorithmic fairness aims to ensure that induced models do not discriminate against groups or individuals based on their protected[1] attributes (e.g., race, gender, age, etc.). Several fairness notions have been formally defined and proposed in the literature in order to assess/quantify discrimination [36]. These fairness notions fall into two main categories namely, group and individual notions. Group fairness notions aim to ensure that sub-populations have similar decisions while individual fairness notions aim to ensure that similar individuals are treated equally [2, 35, 38, 39].

Striking a balance between privacy and fairness while maintaining utility is crucial. However, privacy-preserving algorithms in particular, DP, may tend to disparately affect members of minority groups, implying that privacy and fairness are fundamentally at odds [9, 13, 25, 27, 28]. This tension between fairness and DP is attracting more and more attention, however, a clear understanding of the reasons for this tension is still not well explored. In another line of research, DP and fairness were viewed as aligned objectives. For instance, Dwork et al. [22] proved that individual fairness is a generalization of DP and provided some constraints under which a DP mechanism ensures individual fairness as well. Alternatively, DP and fairness have been integrated as dual objectives in a learning model. For instance, Xu et al. [45] proposed two algorithms to achieve both DP and fairness in logistic regression by combining functional mechanism and decision boundary fairness.

In this paper, we investigate the impact of training a model with obfuscated data under LDP guarantees, employing the well-known k-ary Randomized Response (k-RR) [31] mechanism. The choice of k-RR is motivated by its optimality for distribution estimation under several information theoretic utility functions [31] and also it's design simplicity since k-RR does not require any particular encoding. Specifically, since the output space is equal to the input space, k-RR provides optimal computational

[1] In this paper, we use the term *protected* to designate sensitive attributes from a fairness perspective and the term *sensitive* to designate sensitive attributes from a privacy perspective.

and communication costs for users. Moreover, on the server side, no decoding step is needed. It also means that the server is free to use any post-processing coding techniques (e.g., one-hot encoding, mean encoding, binary encoding) to improve the usefulness of the ML model.

k-RR has traditionally been mainly employed in the one-dimensional scenario in LDP and fairness literature [14, 40], where only one attribute is randomized. However, relying solely on LDP for a single sensitive attribute might be insufficient. This limitation stems from potential correlations that could allow attackers to reconstruct the privatized sensitive attribute. Hence, we specifically address scenarios involving multiple sensitive attributes, providing a more realistic representation of data collections in real-world contexts. Nevertheless, applying k-RR to multi-dimensional sensitive data presents greater challenges [20, 33]. For example, the naive approach of obfuscating each sensitive attribute independently results in the loss of any dependencies between sensitive attributes. This method has been recently employed to evaluate the impact of LDP on fairness [7]. In our study, in addition to this independent setting, we also explore a combined setting that merges all sensitive attributes into a single attribute. Indeed, combined k-RR has not been extensively studied, and its impact on fairness remains unclear, a gap in understanding that we aim to address.

More specifically, the contributions of this paper are threefold. First, we study the impact of LDP on fairness and utility by observing the behaviour of sub-populations separately. This allows for a more complete understanding of how the fairness metrics behave under different LDP guarantees. Second, we compare both independent and combined settings for obfuscating multi-dimensional sensitive attributes under LDP guarantees. Third, we study how the target distribution has an impact on the privacy-fairness-utility trade-off. The key findings of our empirical analysis are:

1. Generally, obfuscating data with LDP contributes generally to reduce disparity.
2. Obfuscating several sensitive attributes (multi-dimensional) reduces disparity more efficiently than obfuscating a single attribute (one-dimensional).
3. The multi-dimensional approaches of LDP (independent vs combined) differ in their impact on fairness only at low privacy guarantees.
4. LDP obfuscation has, typically, disproporationate impact on only one protected group, and this depends heavily on the outcome Y distribution.

Finally, to bridge the gap with practical applications, we frame the observations as concrete recommendations to practitioners considering both ethical concerns of privacy and fairness in ML applications.

2 Related Work

Fairness and (L)DP. To satisfy both privacy and fairness in ML, the literature has proposed several differentially private and fair ML models (e.g., see [26, 30, 43, 45] and references within). However, the current state-of-the-art in the intersection field of DP and fairness is multifaceted [27]. One perspective aligns DP and fairness in an individual fairness context (e.g., [22]), while the other considers them as opposing forces (e.g., [9, 25, 28]), considering group fairness. Regarding group fairness notions (our primary focus), the most popular work [9] explored the effects of training DP deep

learning models, revealing accuracy discrepancies between privileged and unprivileged groups. However, recent studies have begun to observe a negligible [18] or bounded [37] impact of DP deep learning models on group fairness. Regarding the local DP setting, some works [14, 40] proposed to obfuscate only one sensitive attribute under ϵ-LDP guarantees. With the increasing prevalence of collecting multiple sensitive attributes across various industries, relying solely on LDP for a single sensitive attribute may prove insufficient as correlations can still allow attackers to reconstruct the privatized sensitive attribute. For this reason, we consider the case of multiple sensitive attributes, reflecting real-world data collections more accurately. In this context, a recent work [7] has investigated the impact of collecting multi-dimensional under LDP on fairness. However, while [7] has only considered the *independent* setting for randomizing the users multi-dimensional data, for a more comprehensive examination, we have considered both *independent* and *combined* settings (discussed in the following). Another main difference is that we analyze the impact of LDP on fairness by varying the Y distribution (e.g., see Section 5.3).

LDP and multi-dimensional data. Unlike the centralized DP setting, where the server collects users' original data, LDP empowers users to obfuscate their data before transmitting it to the server. While much of the existing literature on LDP has focused on the frequency estimation of one-dimensional data (e.g., [4, 24, 31]), real-world scenarios often involve servers seeking insights into multiple attributes of a population, i.e., *multi-dimensional data*. In this context, the typical process involves local perturbation of user data, followed by statistical estimation and synthetic data generation. The first phase takes place on the user side, while subsequent phases occur at the aggregator/server side. This paper diverges from this typical process in two key aspects. First, it exclusively examines the first phase, studying various approaches for perturbing user multi-dimensional data. Second, it investigates the setup of training an ML model based on the randomized data, as in [7, 14, 40]. Crucially, the objective is to analyze the impact of different data perturbation approaches on the fairness of the learned model. In the LDP literature for multi-dimensional data, prior works have adopted either an *independent* [33, 42], *sampling-based* [5, 17], or *combined* [34] (i.e., joint) perturbation of sensitive attributes. In the former, the randomization mechanism is independently applied to each sensitive attribute, leading to the loss of potentially significant dependencies among attributes and resulting in poor statistical utility. Alternatively, the latter setting, namely combined, treats the Cartesian product of the set of sensitive attributes as a single attribute [20, 33, 34], representing a natural approach to sensitive data perturbation. Our goal is then to study the impact of both the independent and combined approaches of user multi-dimensional data perturbation on the fairness of the obtained model. Notice that the sampling-based approach is not comparable since each user only sends information about one attribute.

3 Preliminaries and Notation

Variables are denoted by capital letters and small letters denote specific values of variables (e.g., $A = a$, $Y = y$). Bold capital and small letters denote a set of variables and a set of values, respectively. In particular, $\mathbf{V}$ denotes the set of all variables in

the data except the outcome. A predictor $\hat{Y}$ of an outcome Y is a function of $\mathbf{V}$ ($\hat{Y} = h(\mathbf{V})$). The set of attributes[2] $\mathbf{V}$ is composed of non-sensitive ($\mathbf{X}$) and sensitive ($\mathbf{A}$) attributes ($\mathbf{V} = (\mathbf{X}, \mathbf{A})$). A *sensitive* attribute reveals private information about an individual and hence should be obfuscated. A *protected* attribute A is an attribute that can be used for discrimination. For example, when deciding to grant a loan to an individual, the protected attribute A could be someone's race or gender. In this paper, we assume that there is only one protected attribute (no intersectionality [35]) and the protected attribute is always sensitive ($A \in \mathbf{A}$). Note that $\mathbf{X}$ could include proxies to A such as zip code which could infer race. Without loss of generality, assume that $\hat{Y}$ and Y are binary random variables where $Y = 1$ (e.g., granting a loan) designates a positive outcome and $Y = 0$ (e.g., denying a loan) designates a negative outcome. In some scenarios, the outcome Y is derived from a score s (e.g., risk score to default on a loan) and a threshold set by domain experts is used to define the cut-off point between the positive outcome and the negative outcome. For the example of granting a loan, an applicant who has a score $s >$ threshold is assigned a positive outcome ($Y = 1$) while an applicant with a score $s <=$ threshold is assigned a negative outcome ($Y = 0$). Thus, varying this threshold causes a variation in the class distribution and potentially leads to different predictions. For the remainder of this paper, we assume that we have access to a dataset D of n i.i.d samples such that $D = (\mathbf{x}_i, \mathbf{a}_i, y_i)_{i=1}^{n}$. Let $\mathcal{L}$ be a randomization[3] algorithm for sensitive attributes. We denote a randomized version of D as $D_z = (\mathbf{x}_i, \mathbf{z}_i, y_i)_{i=1}^{n}$ where $\mathbf{z}_i = \mathcal{L}(\mathbf{a}_i)$.

3.1 Problem Statement

The focus of this work is to shed light on the impact of training a classifier using locally differential private data on fairness assessment. Figure 1 depicts the framework used in this work. To assess fairness, a prediction problem is defined. An example of a prediction problem might be granting loans to individuals or admitting applicants to a college program. As a randomized mechanism $\mathcal{L}$ is applied to the sensitive part of the original data, the predictor $\hat{Y}_Z$ incurs some error. The difference between $\hat{Y}_Z$ and $\hat{Y}_A$ (predictions of the model trained on the original data) quantifies the impact of LDP on Fairness results. As shown in Figure 1, the classification model called $\mathcal{M}_A$ is first trained using the original data $D_{A_{train}}$. We refer to such a model as a baseline model. We then train the same model with the same hyper-parameters using an obfuscated version of the training set $D_{Z_{train}}$. We call this LDP model $\mathcal{M}_Z$. Note that the classification models $\mathcal{M}_A$ and $\mathcal{M}_Z$ are both tested on the original testing samples ($D_{A_{test}}$).

3.2 Local Differential Privacy

This work assumes that the centralized server in charge of aggregating data from individual users is not guaranteed trustworthy. Consequently, we consider the local DP [32] setting, where users obfuscate their data before sending it to the server to train the classification model.

[2] In the rest of the paper we use attribute and variable interchangeably.
[3] The terms randomization and obfuscation are used interchangeably.

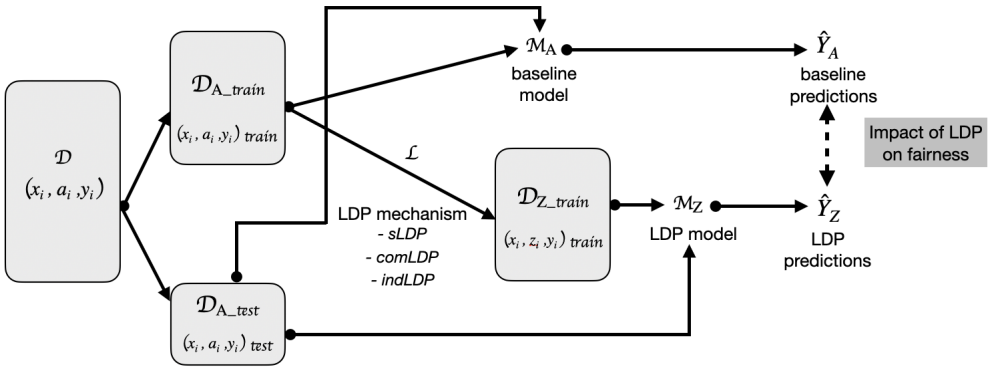


Fig. 1: Our framework for fairness assessment when learning over LDP-based data.

Definition 1 (ϵ-Local Differential Privacy)**.** An algorithm $\mathcal{L}$ with domain and range equal to the domain of $\mathbf{A}$ ($dom(\mathcal{L}) = range(\mathcal{L}) = dom(\mathbf{A})$) satisfies ϵ-local-differential-privacy (ϵ-LDP), where $\epsilon > 0$, if:

$$\max_{\mathbf{a},\mathbf{a}',\mathbf{z} \in dom(\mathbf{A})} \frac{\mathbb{P}(\mathcal{L}(\mathbf{a}) = \mathbf{z})}{\mathbb{P}(\mathcal{L}(\mathbf{a}') = \mathbf{z})} \leq e^{\epsilon}$$

Notice that Definition 1 uses sets of values ($\mathbf{a}, \mathbf{a}'$, and $\mathbf{z}$) instead of single values (a, a', and z) so that it holds when randomizing one dimensional data (one single sensitive attribute) or multi-dimensional data (several sensitive attributes). The same holds for the ϵ-LDP mechanism defined below.

Definition 2. k-Ary Randomized Response (k-RR) Let $\mathbf{A} = \{A_1, A_2, \ldots\}$ be a set of sensitive attributes with a domain $dom(\mathbf{A}) = \{\mathbf{a_1}, \ldots, \mathbf{a_k}\}$ of size k ($k = |dom(\mathbf{A})|$). Given a value $\mathbf{a} \in dom(\mathbf{A})$, k-RR($\mathbf{a}$) outputs the true value $\mathbf{a}$ with probability p, and any other value $\mathbf{a}' \in dom(\mathbf{A}) \setminus \{\mathbf{a}\}$, otherwise. More formally:

$$\forall z \in dom(\mathbf{A}): \quad \mathbb{P}(\mathbf{z} = \mathbf{a}) = \begin{cases} p = \frac{e^{\epsilon}}{e^{\epsilon}+k-1} & \text{if } \mathbf{z} = \mathbf{a}, \\ q = \frac{1}{e^{\epsilon}+k-1} & \text{if } \mathbf{z} \neq \mathbf{a}. \end{cases} \tag{1}$$

where $\mathbf{z}$ is the obfuscated version of $\mathbf{a}$ sent to the server.

It is easy to see that k-RR mechanism satisfies ϵ-LDP as $\frac{p}{q} = e^{\epsilon}$ [31]. As mentioned in Section 1, we choose k-RR as the LDP mechanism to apply because it does not use any specific user-side encoding, resulting in low computational and communication costs on the user side. Moreover, on the server side, k-RR does not require any special decoding and has proven optimal for many theoretical information losses in distribution estimation.

3.3 Fairness

The common taxonomy of fairness metrics classifies them into group and individual metrics [2, 10, 35, 36, 38, 44]. In this paper, we focus on statistical group fairness metrics. These metrics are used to assess the impact of LDP on fairness.

- **Statistical disparity**[22] is one of the most commonly applied fairness metrics. It requires the prediction to be statistically independent of the sensitive attribute ($\hat{Y} \perp A$). In other words, the predicted acceptance rates for both privileged ($A = 1$) and unprivileged ($A = 0$) groups should be equal. A classifier $\hat{Y}$ satisfies statistical parity if:

$$\mathbb{P}(\hat{Y} = 1 \mid A = 1) - \mathbb{P}(\hat{Y} = 1 \mid A = 0). \tag{2}$$

- **Equal opportunity disparity** [29] requires true positive rate equality[4] among groups:

$$\mathbb{P}(\hat{Y} = 1 \mid Y = 1, A = 1) - \mathbb{P}(\hat{Y} = 1 \mid Y = 1, A = 0). \tag{3}$$

- **Predictive equality disparity** [16] requires only the false positive rates[5] to be equal in both groups:

$$\mathbb{P}(\hat{Y} = 1 \mid Y = 0, A = 1) - \mathbb{P}(\hat{Y} = 1 \mid Y = 0, A = 0). \tag{4}$$

- **Overall accuracy disparity** [11] is satisfied when overall accuracy for both groups is the same:

$$\mathbb{P}(\hat{Y} = Y|A = 1) - \mathbb{P}(\hat{Y} = Y|A = 0) \tag{5}$$

- **Predictive rate disparity** [15] requires only the positive predictive value[6] to be equal in both groups and is achieved when:

$$\mathbb{P}(Y = 1 \mid \hat{Y} = 1, A = 1) - \mathbb{P}(Y = 1 \mid \hat{Y} = 1, A = 0) \tag{6}$$

4 Combined vs independent *k*-RR

As noted above, Definitions 1 and 2 hold for the one-dimensional as well as multi-dimensional data. That is, in addition to obfuscating a single sensitive attribute, we also consider the obfuscation of multiple sensitive attributes (Section 5 presents all the k-RR settings we consider in this study). More specifically, we assume there are d sensitive attributes $A_1, A_2, \ldots, A_d$, where the domain of each A_i is a discrete set of finite size $k_i = |dom(A_i)|$. We consider two methods to apply k-RR on multi-dimensional data [20, 33]:

Independent k-RR (k-RR-Ind). This is a naive approach that applies k-RR independently on each attribute. More precisely, k-RR-Ind splits the privacy budget ϵ among the d sensitive attributes, and reports each attribute A_i using k_i-RR parameterized with ϵ_i-LDP, where $\sum_i^d \epsilon_i = \epsilon$. The state-of-the-art approach [5, 8, 33, 42] divides ϵ evenly among the attributes, i.e., a *uniform* solution in which each attribute is reported under $\frac{\epsilon}{d}$-LDP. In this study, we apply the *k-based* solution [7]. This approach consists of splitting ϵ among sensitive attributes based on their domain size. More specifically, each sensitive attribute $A_i \in \mathbf{A}$ is obfuscated with $\epsilon_i = \frac{\epsilon \cdot k_i}{\sum_{i=1}^{d_s} k_i}$ where k_i is the domain size of the attribute A_i ($k_i = |dom(A_i)|$) and d_s is the number of attributes in $\mathbf{A}$ ($d_s = |\mathbf{A}|$).

[4] True positive rate = $\frac{TP}{TP+FN}$
[5] False positive rate = $\frac{FP}{FP+TN}$
[6] Positive predictive value = $\frac{TP}{TP+FP}$

Combined k-RR (k-RR-Comb). This mechanism considers the Cartesian product $A_1 \times A_2 \times \ldots \times A_d$ as a single attribute and sanitizes it using k-RR parameterized with ϵ-LDP, where $k = k_1 \cdot k_2 \cdot \ldots \cdot k_d$.

Independent LDP on multi-dimensional data has been studied relatively well in the literature [5, 33, 42]. Moreover its impact on fairness was the topic of a recent paper [7]. Combined LDP, on the other hand, was not studied extensively. In particular, its impact on fairness is still unclear.

5 Empirical Results and Analysis

To study the impact of k-RR on fairness, two synthetic datasets and two real-world fairness benchmark datasets, namely: *Adult* and *Compas* are used. For each of these datasets, the fairness metrics presented in Section 3.3 are applied.
Environment: All the experiments are implemented in Python 3. We use *Random Forest* model [12] for classification with its default hyper-parameters and we use the ten-fold cross-validation technique, both from Scikit-learn [41]. For k-RR mechanism, we use the implementation in Multi-Freq-LDPy [6]. The codes and datasets for all the experiments are available in the repository [1].
Stability: Since LDP protocols, k-fold cross-validation, and ML algorithms are randomized, we report average results over 20 runs.
Datasets: A summary of all datasets used in this study is provided in Table 1.

Table 1: Metadata of the datasets used in the experiments.

Dataset	n	A *(protected att.)*	**A** *(sensitive att.)*	Y	*Threshold*
Synthetic	100K	A	- A	Y	$\tau_{Q1} = .44$
			- C		$\tau_{Q2} = .52$
			- M		$\tau_{Q3} = .6$
Compas	5915	race	- race	risk score[7]	$\tau_{Q1} = 1$
			- gender		$\tau_{Q2} = 3$
			- age		$\tau_{Q3} = 5$
Adult	32561	gender	- gender	income	$\tau_{Q1} = 10K$
			- age		$\tau_{Q2} = 27K$
			- race		$\tau_{Q3} = 50K$
			- marital-status		
			- native-country		

- **Synthetic Dataset**: The causal model used to generate the synthetic dataset is depicted in Figure 2. A, C, and M are discrete variables[8], while Y is a continuous variable that is a function of all the other variables such that: $Y = h(A, C, M)$. To study the impact of k-RR on fairness while varying the class distribution,

[7] Unlike the synthetic and the *Adult* datasets, whose outcome is continuous, the outcome of the *Compas* dataset is discrete (score $\in [0, 1]$). Thus, we use scores 1, 3, and 5 as thresholds for the Y distribution to be skewed to 0, balanced and skewed to 1, respectively.
[8] C and A follow *Binomial* distributions while M follows *Multinomial* distribution.

three thresholds are set for the outcome variable Y binarisation, resulting in three synthetic datasets differing solely by the distribution of Y. The thresholds and the resulting Y distribution for all datasets are shown in Table 1. Three scenarios are considered depending on the dataset, namely, Y distribution skewed to 1, balanced Y distribution, and Y distribution skewed to 0.

- **Benchmark Datasets**:

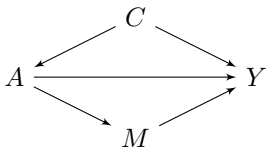
  - *Compas*: The *Compas* dataset includes data about defendants from Broward County, Florida, during 2013 and 2014 who were subject to *Compas* screening. Various information related to the defendants (e.g., race, gender, arrest date, etc.,) were gathered by ProPublica [3] and the goal is to predict the two-year violent recidivism. Only black and white defendants assigned *Compas* risk scores within 30 days of their arrest are kept for analysis leading to 5915 individuals in total. We consider race as the protected attribute. Five attributes are used in this study namely: race, sex, age, priors and risk score. We use the *Compas* risk score as the outcome. The risk score consists of rating of $1-10$ where the higher the score, the more likely the defendant is to re-offend. Following the same reasoning as the other datasets, we transform the risk score into a binary variable by choosing different thresholds to study the impact of outcome distribution on the privacy-fairness trade-off. Three thresholds are used, leading to three different outcome distributions: skewed to 1, almost balanced, and skewed to 0.

  - *Adult*: The *Adult* dataset[19] consists of $32,561$ samples and the goal is to predict the income of individuals based on several personal attributes such as gender, age, race, marital status, education, and occupation. The attributes considered in this work are age, gender, native country, education level, marital status, number of working hours per week, and income. We use the income of an individual as the outcome. Similarly to the other datasets, different thresholds are used to separate the positive outcome (high income) from the negative outcome (low income). Three thresholds are used in total, leading to three versions of the *Adult* dataset with skewed income distribution to 1 (threshold = 10K), balanced income distribution (threshold = 26K), and skewed income distribution to 0 (threshold = 50K[9]).

Applied settings: Four settings (Table 2) are used to assess the impact of LDP on fairness. We vary the privacy level in the range of $\epsilon = \{16, 8, 5, 3, 2, 1, 0.5, 0.1\}$.

- *noLDP* (Baseline): the model is trained using the original data (without privacy).
- *sLDP*: the model is trained using an obfuscated version of the data where only the protected attribute A is obfuscated using k-RR.
- *combLDP*: the model is trained using an obfuscated version of the data where a set of sensitive attributes $\mathbf{A}$, including the protected attribute A is obfuscated using k-RR-Comb (Section 4).
- *indLDP*: the model is trained using an obfuscated version of the data where the same set of sensitive attributes $\mathbf{A}$ is obfuscated using k-RR-Ind (Section 3.2). The privacy splitting solution used in the experiments is the *k-based* solution (3.2).

[9]The 50K threshold is used in the well-known Adult dataset mostly used in the literature [21].



(a) Causal Graph

$$C \sim \mathcal{B}(0.35),$$

$$A \sim \begin{cases} \mathcal{B}(0.55) \text{ if } C = 0, \\ \mathcal{B}(0.75) \text{ if } C = 1, \end{cases}$$

$$M \sim \begin{cases} \mathcal{M}(0.35, 0.4, 0.25) \text{ if } A = 0, \\ \mathcal{M}(0.5, 0.4, 0.1) \text{ if } A = 1, \end{cases}$$

$$Y = \alpha A + \beta M + \gamma C + \mathcal{U}_y, \quad \mathcal{U}_y \sim \mathcal{N}(0, 1)$$

(b) Structural equations

Fig. 2: Causal Model of the Synthetic Dataset.

Table 2: Settings applied in this study.

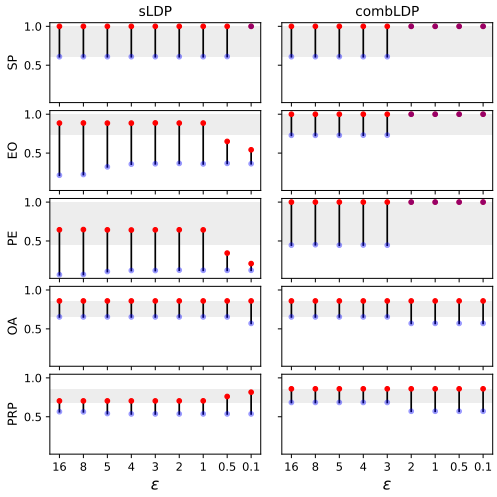
Settings applied	*k-RR applied to*
noLDP	no privacy
sLDP	A
combLDP	**A** using k-RR-Comb
indLDP	**A** using k-RR-Ind

5.1 Impact of LDP on fairness

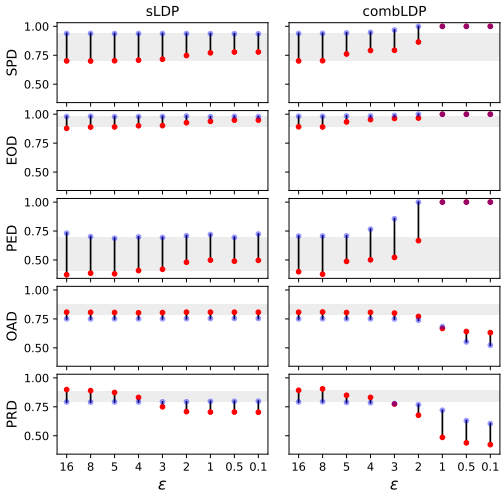
This set of experiments aims to study the effect of obfuscating data through LDP on the fairness of the model trained using that data. The experimental protocol consists of obfuscating data using either *sLDP* (not multi-dimensional) or *combLDP* (multi-dimensional) while decreasing the privacy budget ϵ toward more privacy requirements (small ϵ). Fairness is measured using the various group metrics of Section 3.3, and the experiment is repeated for all three datasets (Synthetic, *Compas*, and *Adult*). Figure 3 shows the obtained results. To better understand how LDP impacts fairness, the plots show the separate values for both groups: the privileged group ($A = 1$) in red dots and the unprivileged group ($A = 0$) in blue dots. Disparity between groups is then the difference between the two values (dots). In addition, for a better understanding of the trade-off, disparity in the baseline case (no obfuscation (*noLDP*)) is shown using a gray shaded area. The following can be observed from the empirical results.

- **[Obs1]** *More privacy leads to less disparity.* For both *sLDP* and *combLDP*, the disparity decreases when imposing stronger privacy requirements (smaller ϵ). For example, in Figure 3b, statistical disparity (first row) decreases from 0.23 to 0.15 (for *sLDP*) and 0 (for *combLDP*). The same decreasing pattern can be observed for equal opportunity disparity (second row) and predictive equality disparity (third row). For *overall accuracy disparity* and *predict rate disparity*, however, disparity either stays unaffected (Figure 3c) or increases (Figures 3a and 3b). These two fairness notions compare both groups' accuracy and precision (e.g., $Y = \hat{Y}$ for accuracy). Hence, the behavior is expected since imposing strong

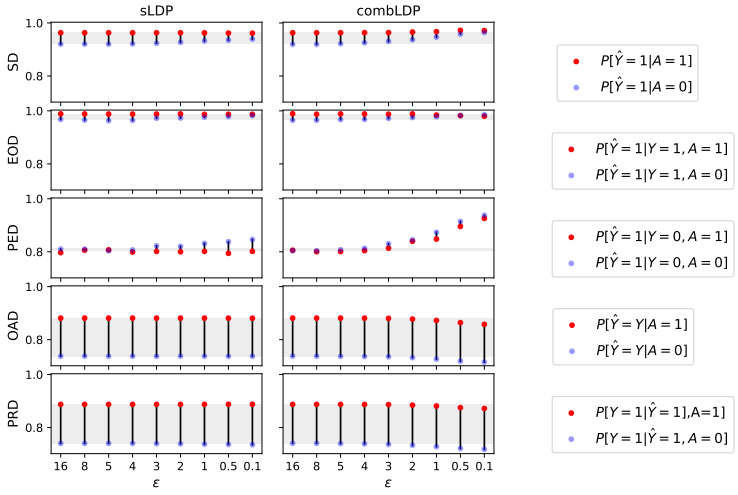
privacy guarantees typically leads to a decrease in the accuracy and precision of the classifier for one or both protected groups. But the drop is greater for one group than for the other. This is further detailed when studying the impact of the outcome distribution on the privacy-fairness-utility trade-off (Section 5.3).



(a) Synthetic data



(b) Compas



(c) Adult

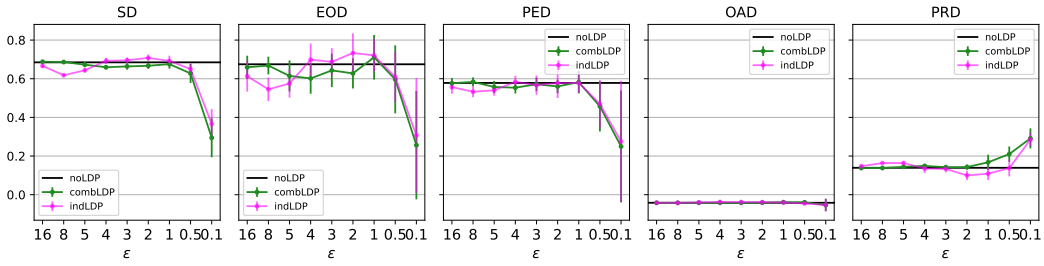
Fig. 3: Impact of LDP on disparity (y-axis) by varying the privacy level ϵ (x-axis). *sLDP* consists in obfuscating a single attribute (protected). *combLDP* consists in obfuscating all sensitive attributes. The gray shaded area represents the disparity results using the baseline model (*noLDP*).

- **[Obs2]** *Multi-dimensional LDP reduces disparity more efficiently than one-dimensional LDP.* Both *sLDP* and *combLDP* lead to a decrease in disparity (previous observation). However, with *combLDP*, the reduction can be observed with weaker privacy guarantees (higher ϵ). In other words, the more attributes are obfuscated, the less privacy level ϵ is needed to improve fairness. For instance, in Figure 3a, the disparity disappears at $\epsilon = 0.1$ for *sLDP*, but at $\epsilon = 2$ for *combLDP*. This can be explained by the fact that obfuscating the protected attribute A (equivalent to removing that attribute from the training set when the privacy guarantees are strong enough) is insufficient to improve fairness due to proxies correlated with that attribute. Thus, by additionally obfuscating all attributes correlated with the protected attribute, weaker privacy guarantees are required to reduce the disparity between groups and, therefore, improve fairness.
- **[Obs3]** *LDP has disproportionate impact on groups.* In most of the plots, one can observe that k-RR does not have an impact (or has a minor impact) on one group but a high impact on the other group. For instance, in the first three rows of Figure 3a, the change in disparity is due to a significant change related to only the unprivileged group (blue dots). In other words, considering groups separately, k-RR impacts the fairness/utility of these groups differently.

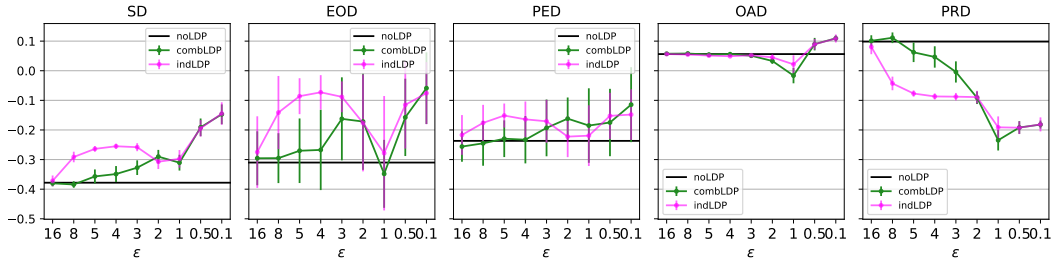
5.2 *k*-RR-Ind vs *k*-RR-Comb

The impact of LDP on the fairness level of the obtained model depends on the multi-dimensional k-RR variant (Section 4) used for obfuscation. The following experiment is performed to compare the effects of k-RR-Ind and k-RR-Comb on the disparity between the privileged and unprivileged groups. Benchmark datasets (Synthetic, *Compas*, and *Adult*) are obfuscated using k-RR-Ind and k-RR-Comb while decreasing the privacy budget ϵ toward more strict privacy guarantees (very small ϵ). The obfuscated data is then used to train a predictor and the disparity of the model is then assessed using the fairness metrics of Section 3.3. Figure 4 shows the result of the experiments.

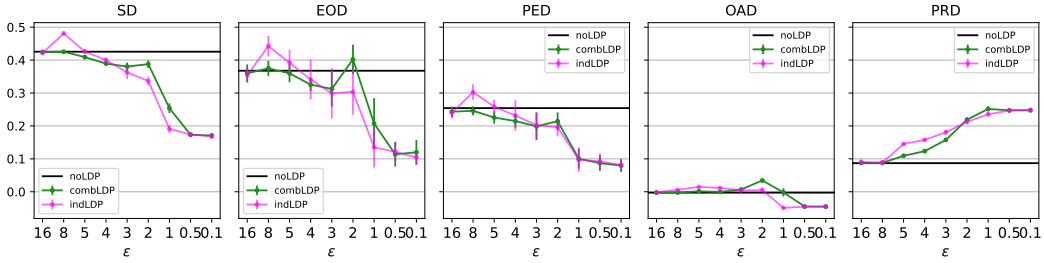
- **[Obs4]** *For large ϵ, the efficiency to reduce disparity depends on the sensitive attributes inter-dependencies. Compas* and *Adult* experiments illustrate the two different behaviors. In *Compas* experiment (Figure 4b), at $\epsilon = 4$, equal opportunity disparity (EOD) for *indLDP* is -0.09 but -0.27 for *combLDP*. Recall, from Table 1, that in *Compas* dataset, three attributes are considered sensitive (race, gender, and age) with relatively low inter-dependencies between them. This explains why k-RR-Ind is more efficient in reducing disparity than k-RR-Comb for large ϵ values. In the *Adult* experiment result (Figure 4c), k-RR-Comb is slightly more efficient than *indLDP* in reducing disparity. For instance, at $\epsilon = 8$, *EOD* is at 0.43 for *indLDP* but at 0.38 for *combLDP*. This can be explained by the relatively high inter-dependencies of the five sensitive attributes (Table 1) considered in the *Adult* dataset.
- **[Obs5]** *For small ϵ, combLDP and k-RR-Ind have a similar impact on disparity.* In all plots of Figure 4, for strict privacy guarantees (small ϵ), the disparity between protected groups converges to the same value whether the obfuscation was performed with *indLDP* or *indLDP*. In other words, by enforcing more privacy, both settings of k-RR improved fairness to the same extent.



(a) Synthetic data



(b) Compas



(c) Adult

Fig. 4: Impact of *combLDP* and *indLDP* on disparity (y-axis) by varying the privacy level ϵ (x-axis) and obfuscating a set of sensitive attributes.

5.3 The effect of changing the outcome distribution

To assess disparity using the group fairness metrics (Section 3.3), the outcome variable Y is required to be binary. However, typically, the trained model predicts a continuous numerical value representing a score as outcome. The score value needs to be thresholded to obtain a binary value. Consequently, the distribution of the outcome variable Y will depend on the threshold value. To study the effect of the outcome distribution on the disparity between protected groups while obfuscating data, the following experiment is performed. Three different distributions are considered for each dataset (Synthetic, *Compas*, and *Adult*). The first distribution is obtained by considering a

threshold value (τ_{Q1}) such that all instances in the three top quantiles have positive outcome ($Y = 1$). The threshold (τ_{Q2}) of the second distribution is selected such that the two top quantiles have positive outcome. And the third threshold (τ_{Q3}) is selected such that only the instances in the top quantile have positive outcome. Each of the obtained datasets is then obfuscated using *sLDP*, *combLDP*, and *indLDP*. Figure 5 show the experimental results for the Adult dataset (Results for Synthetic and Compas can be found in the appendix (Figures 7 and 8)). As in the experiment of Section 5.1, to better understand how fairness is impacted by the distribution of the outcome, the plots track the separate values for each protected group (dots on solid lines for privileged group and dots on dashed lines for unprivileged group). The difference between the two types of dots corresponds to the disparity. Finally, as previously mentioned, the grayed area corresponds to the disparity of the baseline model (*noLDP*).

- **[Obs6]** *When enforcing privacy, which group witnesses more accuracy drop depends on the outcome distribution.* Depending on the threshold for positive outcome (and hence the outcome distribution), the drop in accuracy[10] due to more tight privacy guarantees (smaller ϵ) is higher for one group than the other. In particular, the accuracy drops more for the unprivileged group $A = 0$ when the Y distribution is either skewed to 1 (τ_{Q1}) or balanced (τ_{Q2}), which correspond to the first and second columns in Figure 5. Whereas it drops more for the privileged group $A = 1$ when the Y distribution is skewed to 0 (τ_{Q1})[11].
- **[Obs7]** *When enforcing privacy, which group contributes more to reduce the disparity depends on the outcome distribution.* Similarly to the above observation, the outcome distribution has significant impact on how each group (privileged vs unprivileged) contributes to the disparity reduction while enforcing more privacy. In particular, the prediction rates per group (e.g. $P(\hat{Y} = 1|A = 1)$ for SD) *increased more for the unprivileged group $A = 0$ when the outcome distribution is skewed to* 1 *(τ_{Q1} and τ_{Q2}) but* decreased more for the privileged group $A = 1$ when the outcome distribution is skewed to 0 (τ_{Q3})[12].
- **[Obs8]** *For a fair baseline model, enforcing privacy amplifies disparity.* The outcome distribution experiment exhibited an interesting behavior illustrated clearly in the *Adult* dataset results (Figure 5). In particular, for the PED metric with outcome distribution at threshold τ_{Q1}, the disparity in the baseline predictor is relatively small. However, training the predictor using obfuscated data resulted in disparity amplification. A similar behavior is observed for OAD with τ_{Q2}.

Based on the above observations, one can conclude the following statements:

Statement 1: *If $A = a$ is the privileged group (has a majority of $Y = 1$) then if Y is skewed to* 1*, adding noise affects more the accuracy of the unprivileged group $A \neq a$ else (Y is skewed to* 0*) adding noise affects more the accuracy of $A = a$.*

Statement 2: *If $A = a$ is the privileged group (has a majority of $Y = 1$), then if Y is skewed to* 1*, adding noise increases more the predicted rates for the unprivileged*

[10] As this observation is about the accuracy, only the last two fairness metrics are concerned, that is, OAD and PRD corresponding to the two lower rows of Figure 5.

[11] Note that this observation is also confirmed in the *Compas* dataset (Figure 8) but inverted since the privileged group in this dataset is the group $A = 0$. To confirm the inversed behavior, we generated a second synthetic dataset where the group $A = 0$ is privileged. The plots can be found in Appendix A.1.

[12] Again, the behavior is reversed for the *Compas* dataset (Figure 8) for the same reason as the previous observation.

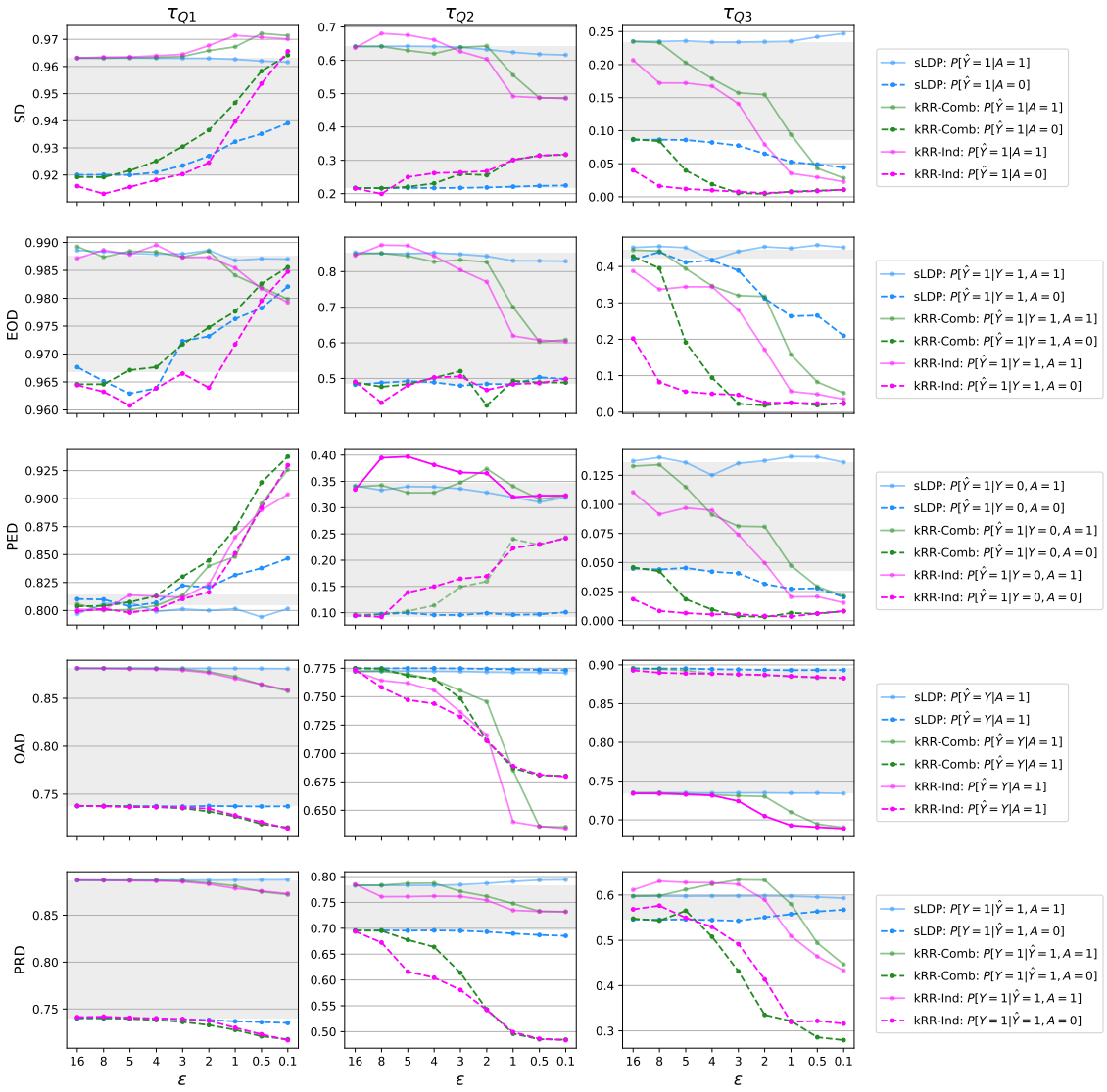


Fig. 5: Impact of Y distribution on the privacy-fairness trade-off. Columns 1, 2, and 3 illustrate the results for the *Adult* dataset when the Y distribution is skewed to 1, balanced, and skewed to 0, respectively.

group $A \neq a$ else (Y is skewed to 0), adding noise decreases more the predicted rates for group $A = a$.

5.4 Recommendations

Based on the observations obtained from the experimental analysis, one can propose the following recommendations for a practitioner who is considering a mechanism satisfying privacy and fairness guarantees. That is, a mechanism allowing individual users to share their data while at the same time protecting their sensitive information and guaranteeing that the obtained model is fair with respect to sub-populations and/or individuals.

A. LDP data obfuscation is an efficient mechanism to reduce disparity.

Almost all observations from the experimental analysis confirm the conclusion that LDP obfuscation reduces disparity (**Obs1, Obs2, Obs4, Obs7**). The disparity reduction is often due to one group being more sensitive to the LDP obfuscation rather the other (**Obs2**). The only exception is when the predicted model using baseline (not obfuscated) data is already fair. In that case, LDP may create disparity (**Obs8**).

B. Obfuscating several sensitive attributes allows to reduce disparity more efficiently than a single attribute.

If a practioner is interested in producing a fair model but with a minimal privacy enforcement, it is recommended that she uses multi-dimensional LDP obfuscating as many sensitive attributes as possible (**Obs2**).

C. Independent and combined variants of multi-dimensional LDP are different only with weak privacy guarantees.

The choice of the multi-dimensional approach of LDP (combined vs independent) matters only at low privacy guarantees (large ϵ) (**Obs4**). In that case, the practitioner's choice should depend on the level of interdependency between sensitive attribute. For high interdependency, a combined approach is more efficient to reduce disparity. For low or no interdependency, an independent approach is more efficient. At strict privacy guarantees (low ϵ), however, both approaches have similar effect on disparity (**Obs5**).

D. Obfuscating data impacts disproportionally only one group depending on the outcome distribution.

A practitioner who obfuscates individual data with LDP should expect that only one group will be significantly affected. And she can *guess* which group will be more affected by studying the outcome distribution. More precisely, if the outcome distribution is skewed towards the positive outcome (typically $Y = 1$), it is the unpriviliged group who will be more affected. Otherwise (outcome distribution is skewed to the negative outcome (typically $Y = 0$), it is the privileged group who will be more affected (**Obs7** and **Obs8**).

6 Conclusion

This paper investigates how the accuracy and fairness of the decisions made by the model change under local differential privacy (LDP), in particular, k-ary Randomized Response (k-RR) mechanism, given different levels of privacy and different class

distributions. To broaden the scope of our study, we employed various group fairness metrics and evaluated two settings for obfuscating multi-dimensional sensitive attributes under LDP, namely, independent and combined, on one synthetic and two benchmark datasets to substantiate our claims. The experimental analysis revealed very relevant observations that we framed as concrete recommendations for machine learning practitioners aiming at guaranteeing both ethical concerns of privacy and fairness. To the best of our knowledge, this is the first work which studies the effect of combined multi-dimensional LDP on fairness. In particular, we observed that combined LDP reduces more efficiently the disparity at low privacy guarantees (high ϵ). As future work, we aim to formalize the privacy-utility-fairness trade-off when learning over LDP-based data, as well as to propose LDP- and fairness-aware ML models.

References

[1] Impact ldp on fairness repository. https://github.com/KarimaMakhlouf/Impact_of_LDP_on_Fairness.

[2] Guilherme Alves, Fabien Bernier, Miguel Couceiro, Karima Makhlouf, Catuscia Palamidessi, and Sami Zhioua. Survey on fairness notions and related tensions. *arXiv preprint arXiv:2209.13012*, 2022.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. propublica. *See https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing*, 2016.

[4] Differential Privacy Team Apple. Learning with privacy at scale, Dec 2017.

[5] Héber H. Arcolezi, Jean-François Couchot, Bechara Al Bouna, and Xiaokui Xiao. Random sampling plus fake data: Multidimensional frequency estimates with local differential privacy. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 47–57, New York, NY, USA, 2021. Association for Computing Machinery.

[6] Héber H. Arcolezi, Jean-François Couchot, Sébastien Gambs, Catuscia Palamidessi, and Majid Zolfaghari. Multi-freq-ldpy: Multiple frequency estimation under local differential privacy in python. In Vijayalakshmi Atluri, Roberto Di Pietro, Christian D. Jensen, and Weizhi Meng, editors, *Computer Security – ESORICS 2022*, pages 770–775, Cham, 2022. Springer Nature Switzerland.

[7] Héber H. Arcolezi, Karima Makhlouf, and Catuscia Palamidessi. (local) differential privacy has NO disparate impact on fairness. In *Data and Applications Security and Privacy XXXVII*, pages 3–21. Springer Nature Switzerland, 2023.

[8] Héber H. Arcolezi, Jean-François Couchot, Bechara Al Bouna, and Xiaokui Xiao. Improving the utility of locally differentially private protocols for longitudinal and multidimensional frequency estimates. *Digital Communications and Networks*, 2022.

[9] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32, 2019.

[10] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

[11] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

[12] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[13] Hongyan Chang and Reza Shokri. On the privacy risks of algorithmic fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 292–303. IEEE, 2021.

[14] Canyu Chen, Yueqing Liang, Xiongxiao Xu, Shangyu Xie, Yuan Hong, and Kai Shu. When fairness meets privacy: Fair classification with semi-private sensitive attributes. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*, 2022.

[15] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[16] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2017.

[17] José Serafim Costa Filho and Javam C Machado. Felip: A local differentially private approach to frequency estimation on multidimensional datasets. In *Proceedings of the 26th International Conference on Extending Database Technology, EDBT 2023, Ioannina, Greece, March 28 - March 31, 2023*, pages 671–683. OpenProceedings.org, 2023.

[18] Anderson Santana de Oliveira, Caelin Kaplan, Khawla Mallat, and Tanmay Chakraborty. An empirical analysis of fairness notions under differential privacy. *arXiv preprint arXiv:2302.02910*, 2023.

[19] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[20] Josep Domingo-Ferrer and Jordi Soria-Comas. Multi-dimensional randomized response. *IEEE Transactions on Knowledge and Data Engineering*, 34(10):4933–4946, 2022.

[21] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[22] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[23] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284. Springer Berlin Heidelberg, 2006.

[24] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.

[25] Tom Farrand, Fatemehsadat Mireshghallah, Sahib Singh, and Andrew Trask. Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. In *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice*, pages 15–19, 2020.

[26] Bogdan Ficiu, Neil D. Lawrence, and Andrei Paleyes. Automated discovery of trade-off between utility, privacy and fairness in machine learning models. *arXiv preprint arXiv:2311.15691*, 2023.

[27] Ferdinando Fioretto, Cuong Tran, Pascal Van Hentenryck, and Keyu Zhu. Differential privacy and fairness in decisions and learning tasks: A survey. *arXiv preprint arXiv:2202.08187*, 2022.

[28] Georgi Ganev, Bristena Oprisanu, and Emiliano De Cristofaro. Robin hood and matthew effects: Differential privacy has disparate impact on synthetic data. In *International Conference on Machine Learning*, pages 6944–6959. PMLR, 2022.

[29] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

[30] Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi Malvajerdi, and Jonathan Ullman. Differentially private fair learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3000–3008. PMLR, 09–15 Jun 2019.

[31] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. In *International Conference on Machine Learning*, pages 2436–2444. PMLR, 2016.

[32] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on*

Computing, 40(3):793–826, 2011.

[33] Hiroaki Kikuchi. Castell: Scalable joint probability estimation of multi-dimensional data randomized with local differential privacy. *arXiv preprint arXiv:2212.01627*, 2022.

[34] Gaoyuan Liu, Peng Tang, Chengyu Hu, Chongshi Jin, and Shanqing Guo. Multi-dimensional data publishing with local differential privacy. In *Proceedings of the 26th International Conference on Extending Database Technology, EDBT 2023, Ioannina, Greece, March 28 - March 31, 2023*, pages 183–194. OpenProceedings.org, 2023.

[35] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management*, 58(5):102642, 2021.

[36] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. On the applicability of machine learning fairness notions. 23(1):14–23, may 2021.

[37] Paul Mangold, Michaël Perrot, Aurélien Bellet, and Marc Tommasi. Differential privacy has bounded impact on fairness in classification. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 23681–23705. PMLR, 23–29 Jul 2023.

[38] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), jul 2021.

[39] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163, 2021.

[40] Hussein Mozannar, Mesrob Ohannessian, and Nathan Srebro. Fair learning with private demographic data. In *International Conference on Machine Learning*, pages 7066–7075. PMLR, 2020.

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[42] Xuebin Ren, Chia-Mu Yu, Weiren Yu, Shusen Yang, Xinyu Yang, Julie A McCann, and S Yu Philip. Lopub: high-dimensional crowdsourced data publication with local differential privacy. *IEEE Transactions on Information Forensics and Security*, 13(9):2151–2166, 2018.

[43] Cuong Tran, Ferdinando Fioretto, and Pascal Van Hentenryck. Differentially private and fair deep learning: A lagrangian dual approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):9932–9939, May 2021.

[44] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.

[45] Depeng Xu, Shuhan Yuan, and Xintao Wu. Achieving differential privacy and fairness in logistic regression. In *Companion proceedings of The 2019 world wide web conference*, pages 594–599, 2019.

A Appendix

A.1 Results of the Synthetic dataset 2

The synthetic dataset 2 follows the exact same causal model depicted in Figure 2. The data distribution is the only difference between the Synthetic datasets 1 and 2. More specifically, synthetic data 2 differed from synthetic data 1 solely by Y distribution.

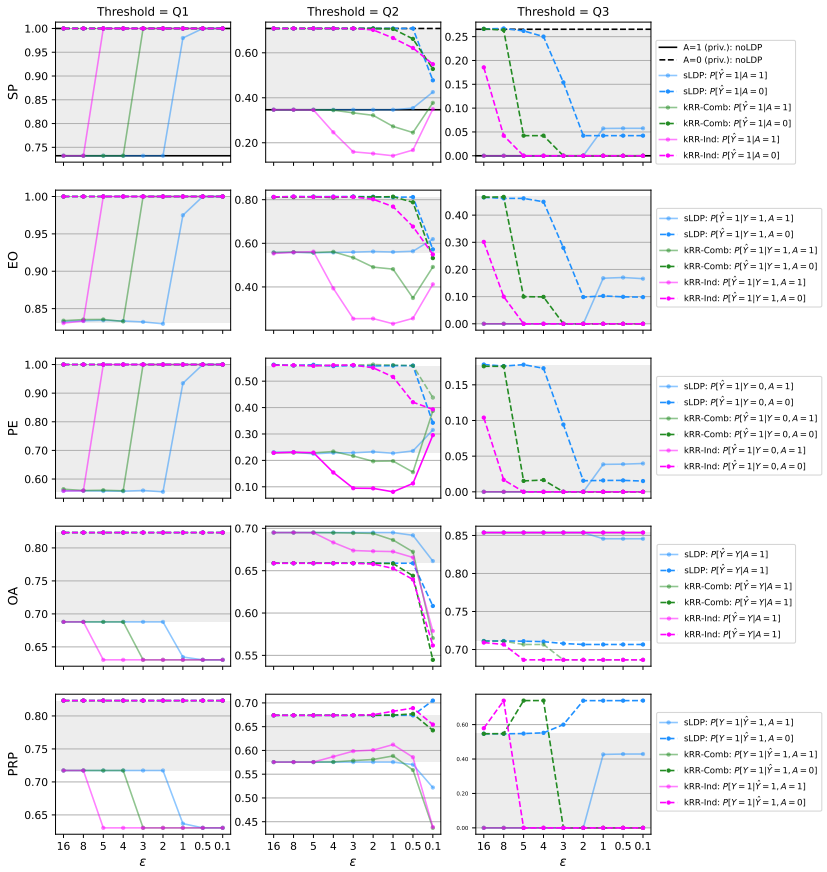


Fig. 6: Impact of k-RR on fairness for the Adult datasets generated with three different thresholds leading to different Y distributions. Synthetic data 2

A.2 Synthetic and Compas experimental results for Section 5.3

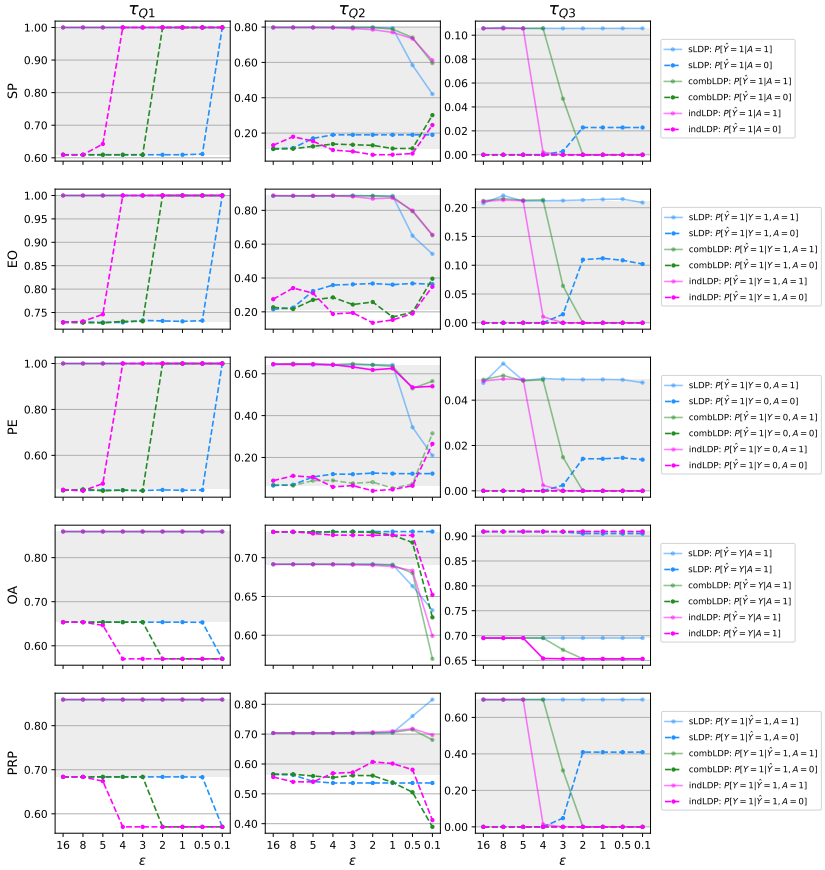


Fig. 7: Impact of Y distribution on the privacy-fairness trade-off. Columns 1, 2, and 3 illustrate the results for the synthetic dataset when the Y distribution is skewed to 1, balanced, and skewed to 0, respectively.

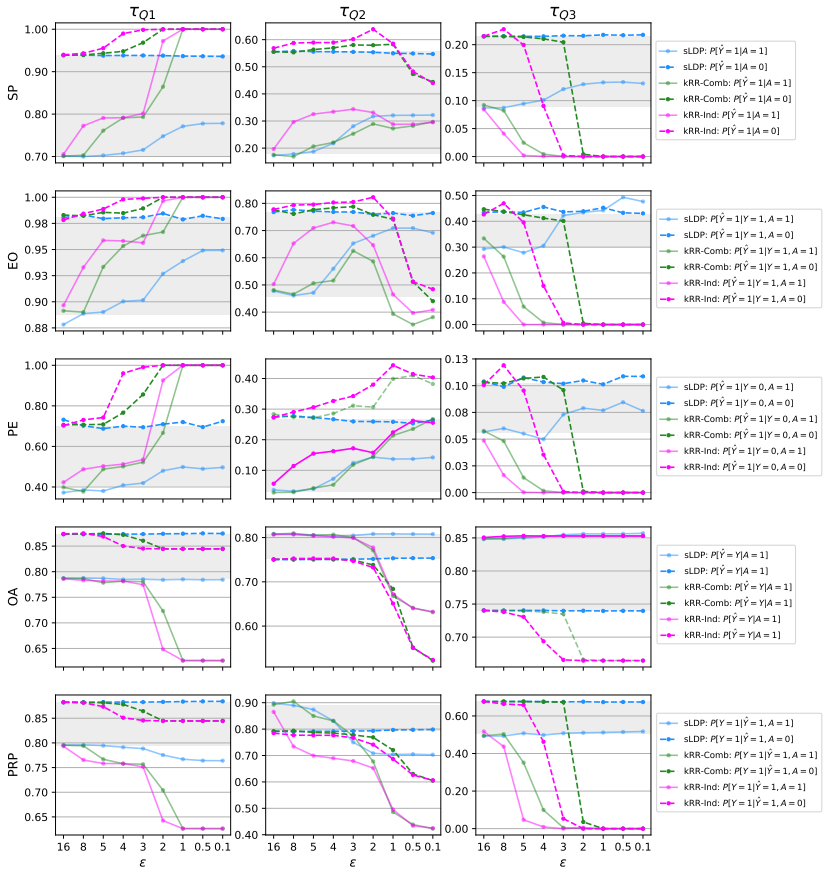


Fig. 8: Impact of Y distribution on the privacy-fairness trade-off. Columns 1, 2, and 3 illustrate the results for the *Compas* dataset when the Y distribution is skewed to 1, balanced, and skewed to 0, respectively.