A Systematic and Formal Study of the Impact of Local Differential Privacy on Fairness: Preliminary Results

Karima Makhlouf, Tamara Stefanović, Héber Hwang Arcolezi, Catuscia Palamidessi

A Systematic and Formal Study of the Impact of Local Differential Privacy on Fairness: Preliminary Results

Karima Makhlouf
karima.makhlouf@lix.polytechnique.fr
INRIA, École Polytechnique, IPP
Paris, France

Tamara Stefanović
tamara.stefanovic@mi.sanu.ac.rs
Mathematical Institute of the Serbian Academy of Sciences and Arts
Belgrade, Serbia

Héber H. Arcolezi
heber.hwang-arcolezi@inria.fr
Inria Centre at the University Grenoble Alpes
Grenoble, France

Catuscia Palamidessi
catuscia@lix.polytechnique.fr
INRIA, École Polytechnique, IPP
Paris, France

ABSTRACT

Machine learning (ML) algorithms rely primarily on the availability of training data, and, depending on the domain, these data may include sensitive information about the data providers, thus leading to significant privacy issues. Differential privacy (DP) is the predominant solution for privacy-preserving ML, and the local model of DP is the preferred choice when the server or the data collector are not trusted. Recent experimental studies have shown that local DP can impact ML prediction for different subgroups of individuals, thus affecting fair decision-making. However, the results are conflicting in the sense that some studies show a positive impact of privacy on fairness while others show a negative one. In this work, we conduct a systematic and formal study of the effect of local DP on fairness. Specifically, we perform a quantitative study of how the fairness of the decisions made by the ML model changes under local DP for different levels of privacy and data distributions. In particular, we provide bounds in terms of the joint distributions and the privacy level, delimiting the extent to which local DP can impact the fairness of the model. We characterize the cases in which privacy reduces discrimination and those with the opposite effect. We validate our theoretical findings on synthetic and real-world datasets. Our results are preliminary in the sense that, for now, we study only the case of one sensitive attribute, and only statistical disparity, conditional statistical disparity, and equal opportunity difference.

KEYWORDS

Differential Privacy, Machine learning, Fairness, Randomized Response

1 INTRODUCTION

Information gathered about individuals is frequently utilized to make decisions that affect those same individuals. For example, financial services firms use ML models for risk management and real-time market data analysis. Medical researchers also use machine learning (ML) to identify disease signs and risk factors, and doctors to help diagnose illnesses and medical conditions in patients. In these contexts, there is a tension between the requirement for accurate systems ensuring individuals receive their due and the necessity to safeguard individuals from the improper exposure of their confidential information. In other words, can these models be trusted to operate on personal and sensitive data? Are these models fair or do they potentially reproduce or exacerbate existing bias in society?

Differential privacy (DP) [18] is currently the leading privacy-preserving ML solution to protect sensitive information about individuals in data set used for statistical purposes or for training machine learning models. This is achieved by injecting controlled noise on the aggregated data or during the learning process. However, the original model of DP (aka central DP) assumes trustworthy data collectors and servers. For this reason, in recent years the local model of DP (LDP) [26] has gained more and more attention, as it achieves privacy guarantees without the above assumptions. Indeed, with LDP each data point is locally obfuscated at the data-owner side before being collected, thus protecting data from privacy leaks at both the source and the server side. LDP has been endorsed and deployed by big tech companies. For instance, Google Chrome uses LDP to collect data from users [19], and Apple uses LDP to collect emoji usage data, word usage, and other information from iPhone users (iOS keyboard) [5].

On the other hand, algorithmic fairness strives to guarantee that generated models refrain from discriminating against groups or individuals on the basis of their sensitive attributes (such as race, gender, age, etc.). Numerous fairness metrics have been formally established and suggested in the literature to evaluate or quantify discrimination [29]. These metrics can be broadly categorized into two main groups: group metrics and individual metrics. Group fairness metrics seek to guarantee uniform decisions across subpopulations, while individual fairness metrics aim to ensure that comparable individuals are treated equally [3, 28, 31, 32].

Achieving both privacy and fairness is crucial. However, it has been shown that privacy-preserving algorithms, and in particular (central and local) DP, tend to affect majority and minority groups differently, thus implying that in some cases privacy and fairness are at odds [2, 8, 12, 20, 36]. Nevertheless, in other lines of research, DP and fairness results align. For instance, Dwork et al. [17] proved that individual fairness is a generalization of DP and provided some constraints under which a DP mechanism also ensures individual fairness. Xu et al. [42] proposed algorithms to achieve both DP and fairness in logistic regression by incorporating fairness constraints in the objective function. Recently, Arcolezi et al. [7] proposed a

novel privacy budget allocation scheme that considers the varying domain size of sensitive attributes and showed that, under this scheme, LDP leads to slightly improved fairness in learning tasks. *These contrasting claims, most of which are backed only by experimental results, show that a systematic and foundational study of the relationship between privacy and fairness is highly needed.* This work is a step in that direction.

Specifically, we formally study the impact of training a model with data obfuscated by randomized response (RR) [39], a fundamental LDP protocol [23] that serves as a building block for more complex LDP mechanisms (e.g., [10, 19, 38]). The choice of RR is also motivated by its optimality for distribution estimation under several information theoretic utility functions [24] and by its design simplicity as it does not require any particular encoding. Specifically, RR provides optimal computational and communication costs for users since the output space equals the input space. Moreover, no decoding step is needed on the server side. 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.

Building on this foundation, our main contribution consists of a theoretical analysis of how the fairness of the prediction of an ML model is affected by the application of RR on the training data, depending on the level of privacy and the data distribution. In particular, we study three notions of fairness: statistical disparity [17], conditional statistical disparity [13], and equal opportunity [22], and identify the conditions under which they are improved or reduced by RR. We then empirically validate our results by performing experiments on synthetic data and four real datasets, *Compas* [4], *Adult* [15], *German credit* [16], and *LSAC* [40]. All detailed proofs supporting our findings are available in Appendix A.1.

2 RELATED WORK

Although the topics of privacy and fairness have been studied for years by philosophers and sociologists, they have only recently received significant attention from the computer science community. This section provides an overview of existing research that explores the intriguing relationship between DP and fairness, specifically in the context of ML. We also recommend the recent survey by Fioretto et al. [21], which discusses the conditions under which DP and fairness have aligned or contrasting goals in decision and learning tasks.

Central differential privacy. Pujol et al. [36] empirically measure the impact of differentially private algorithms on allocation processes. They use two privacy mechanisms: the Laplace and the Data-and-Workload-Aware algorithm. Their results show that in the settings where the introduced noise is modest (higher ε), impacts on fairness may be negligible. However, the introduced noise disproportionately impacts different groups under strict privacy constraints (smaller ε). In [8], the authors empirically showed that by introducing noise, the accuracy of a model trained using DP-SGD [1] decreases compared to the original, non-private model. More specifically, if the original model is "unfair" (in the sense that accuracy is not the same across different subgroups), then DP-SGD deepens the differences between subgroups. Mangold et al. in [30] performs a theoretical analysis of the impact of central DP on fairness in classification. They prove that the difference in fairness levels between private and non-private models diminishes at a rate of $\tilde{O}\left(\sqrt{p}/n\right)$, where n represents the number of training records and p is the number of parameters. They also provide an empirical study using the central model with Gaussian noise for DP and $l2$-regularized logistic regression models for prediction.

Local differential privacy. In [33], Mozannar et al. show how to adapt non-discriminatory learners to work with privatized attributes, giving theoretical guarantees on performance. The experimental analysis by Makhlouf et al. [27] showed that obfuscating several sensitive attributes instead of obfuscating only the sensitive attribute used to assess fairness gives better results for fairness. Also, the authors observed that combined LDP, compared to independent LDP, reduces the disparity more efficiently at low privacy guarantees (high ε). Arcolezi et al. [7] also empirically deals with the impact on fairness of applying LDP to multiple sensitive attributes. The analysis covers several fairness metrics and state-of-the-art LDP protocols. Their results contrast with those obtained with central DP, as they show that LDP slightly improves fairness in learning tasks without significant loss of the accuracy of the model.

3 PRELIMINARIES AND NOTATION

This section presents the framework we consider in this work and briefly recalls the privacy setting and the fairness metrics applied in this study.

Variables are denoted by uppercase letters, while lowercase letters denote specific values of variables (e.g., $A = a$, $Y = y$). A predictor $\hat{Y}$ of an outcome Y is a function of a set of variables (A, X) where X designates the set of non-sensitive attributes and $A \in \{0, 1\}$ represents the sensitive attribute[1]. For example, when deciding to hire an individual, the sensitive attribute could be someone's gender or race, and the non-sensitive attributes X could include the person's education level and professional experience. Note that X could include proxies to A, such as zip code, which could hint to race. We assume that $\hat{Y}$ and Y are binary random variables where $Y = 1$ (e.g., hiring a person) designates a positive outcome, and $Y = 0$ (e.g., not hiring a person) designates a negative outcome. For the remainder of this paper, we assume that we have access to a (multi)set $S = \{(a_i, x_i, y_i)\}_{i=1}^{n}$ of n i.i.d samples from the distribution on $A \times X \times Y$.

We call $A' = \mathcal{L}(A)$ the obfuscated version of the sensitive attribute A, where $\mathcal{L}$ is a certain randomized LDP mechanism. Thus, we denote a randomized version of S as $S' = (a'_i, x_i, y_i)_{i=1}^{n}$.

3.1 The framework

Figure 1 illustrates the framework deployed in our work. We assume a given decision task, such as deciding whether to release a convict on parole or admit an applicant to a college program. We assume that we dispose of a set of data $S = (A, X, Y)_{train} \bigcup (A, X, Y)_{test}$ for building an ML model to help with the task, and for evaluating it. Specifically, $(A, X, Y)_{train}$ is used for training the model, and $(A, X, Y)_{test}$ to assess the fairness of its predictions.

As shown in Figure 1, in order to measure the impact of the LDP mechanism $\mathcal{L}$, we train two models. The baseline model $\mathcal{M}$

[1] In this work, we consider a single sensitive attribute (no intersectionality [28]) and X can be a vector of variables.

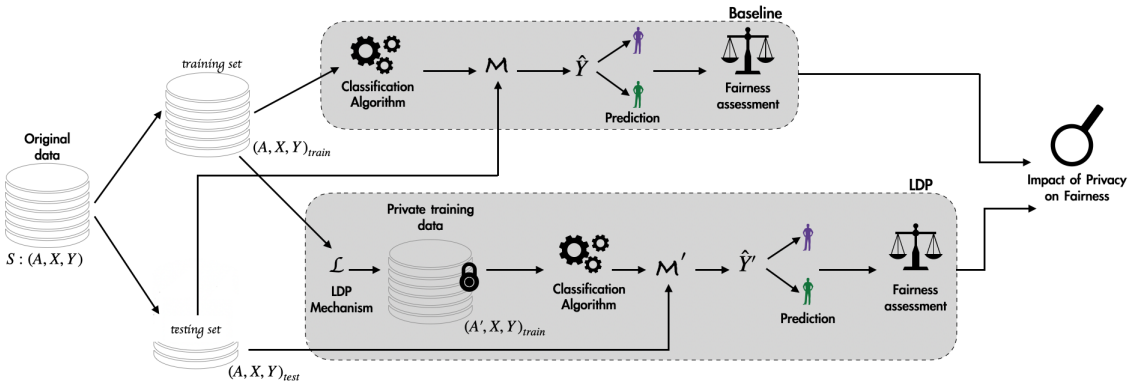


Figure 1: Our framework to assess the impact of LDP on the fairness of a ML model.

(upper shaded box) is trained on the original data $(A, X, Y)_{train}$, and we call its prediction $\hat{Y}$. Then, we obfuscate the training set by applying $\mathcal{L}$ to the A component of each sample in $(A, X, Y)_{train}$. The resulting data set $(A', X, Y)_{train}$ is used to train (with the same classification algorithm and the same hyper-parameters) a second model $\mathcal{M}'$, whose prediction is called $\hat{Y}'$ (lower shaded box).

The difference between $\hat{Y}'$ and $\hat{Y}$ on the original testing data quantifies the impact of LDP on the fairness of the model. *It is important to emphasize that, in our framework, the individual predictions, both for $\mathcal{M}$ and $\mathcal{M}'$, are obtained by applying the models to the original testing data* $(A, X, Y)_{test}$. Namely, in testing phase, $\hat{Y} = \mathcal{M}(A, X)$ and $\hat{Y}' = \mathcal{M}(A, X)$ (instead of $\hat{Y}' = \mathcal{M}(A', X)$). This is because we argue that fairness must be evaluated on the true data. Indeed, even if a model was trained on obfuscated data, it is likely to receive the true data as input at the moment of its deployment. And in any case, the presence of proxies may reveal the true value of the sensitive variable anyway.

3.2 Local Differential Privacy

We recall the definition of LDP as given in the literature for the discrete case.

Definition 1 (ε-Local Differential Privacy [23]). An algorithm $\mathcal{L}$ satisfies ε-LDP, where ε is a positive real number representing the privacy parameter, if for any input v_1 and $v_2 \in Dom(\mathcal{L})$ and for all possible output y:

$$\mathbb{P}[\mathcal{L}(v_1) = y] \leq e^{\varepsilon}\, \mathbb{P}[\mathcal{L}(v_2) = y]$$

Several LDP mechanisms have been proposed in the literature [41]. The mechanism we consider here is randomized response (RR) [23, 39] for a binary variable $a \in \{0, 1\}$, which is defined in Equation (1).

$$\mathrm{RR}(a) = \begin{cases} a & \text{with probability } \frac{e^{\varepsilon}}{e^{\varepsilon}+1} \\ \overline{a} & \text{with probability } \frac{1}{e^{\varepsilon}+1}. \end{cases} \quad (1)$$

where $\overline{a} = 1$ if $a = 0$ and, viceversa, $\overline{a} = 0$ if $a = 1$. It is easy to see that RR satisfies ε-LDP. For simplicity, we will denote by p the probability $e^{\varepsilon}/(e^{\varepsilon}+1)$ that the reported value is the true value.

3.3 Fairness

Many fairness metrics have been proposed in the literature, and they fall into two main categories, namely, the group and the individual fairness metrics [3, 9, 28, 29, 31, 37]. This paper focuses on statistical group fairness metrics, which assess fairness based on a predefined sensitive attribute A. We will call *privileged* the group for which the prediction is favorable ($\hat{Y} = 1$) more frequently, and *unprivileged* the other group. We will consider the following metrics.

- **Statistical disparity (SD)** [17] is the most basic notion of fairness. It measures the difference in acceptance rates between groups and is defined as:
$$\mathrm{SD} = \mathbb{P}[\hat{Y} = 1 \mid A = 1] - \mathbb{P}[\hat{Y} = 1 \mid A = 0]. \quad (2)$$
- **Conditional statistical disparity (CSD)** [13] is a variant of statistical disparity obtained by conditioning on a set of explanatory attributes which may legitimate the discrimination [25]. In this paper, we assume that all variables in X are (potential) explanatory variables, and we define conditional statistical disparity for each instance x of X as follows:
$$\mathrm{CSD}_x = \mathbb{P}[\hat{Y} = 1 \mid X = x, A = 1] - \mathbb{P}[\hat{Y} = 1 \mid X = x, A = 0]. \quad (3)$$
Note that, in general, x represents a tuple of values since X may contain more than one attribute.
- **Equal opportunity difference (EOD)** [22] is one of the most popular notions of fairness nowadays. It measures the disparity between the true positive rate[2] among the two groups:
$$\mathrm{EOD} = \mathbb{P}[\hat{Y} = 1 \mid Y = 1, A = 1] - \mathbb{P}[\hat{Y} = 1 \mid Y = 1, A = 0]. \quad (4)$$

4 QUANTITATIVE ANALYSIS OF THE IMPACT OF PRIVACY ON FAIRNESS

In this section, we formally study the impact of LDP on fairness. Specifically, we perform a quantitative study of how the fairness of the prediction is affected by the application of the RR mechanism to the sensitive values in the training data, depending on the level ε of privacy and on the data distribution.

We briefly recall our setting. In addition to the sensitive attribute A and the true decision Y, which are binary, the data includes a set

[2]The true positive rate is defined as $\frac{TP}{TP+FN}$, where TP are the true positive predictions and FN are the false negative predictions.

of non-sensitive attributes X with arbitrary values. We assume that the data model is *probabilistic*, in the sense that the data may contain tuples with the same values for X and A and different values for Y. $A' = \mathrm{RR}(A)$ is an obfuscated[3] version of A obtained by applying the RR mechanism to A, and it is also binary. The prediction of the model trained on the original data is denoted by $\hat{Y}$, while that of the model trained on the obfuscated data, which we will call LDP model, is $\hat{Y}'$. Of course, $\hat{Y}$ and $\hat{Y}'$ are also binary. We assume that both models are deterministic. Namely, on a given input (x, a), $\mathcal{M}$ always outputs the same prediction. The same holds for $\mathcal{M}'$, although the prediction may be different from the one of $\mathcal{M}$.

Table 1 shows some abbreviations and definitions we use in the paper. In particular, Δ_a^x denotes the difference between the frequency of the samples with the positive true decision ($Y = 1$) and those with the negative true decision ($Y = 0$), and have $A = a$ and $X = x$. On the other hand, Γ_a^x denotes the difference between the positive and negative decision rates *given* $A = a$ and $X = x$. $\Delta_a'^x$ and $\Gamma_a'^x$ denote the corresponding quantities in the obfuscated training data (i.e., on the samples with $A' = a$ and $X = x$).

In order to reason formally about the impact of privacy on fairness, we need to make a basic assumption about the training algorithm. Namely, we assume that the baseline model, in correspondence of the input (x, a), predicts $\hat{Y} = 1$ if $\Delta_a^x \geq 0$, namely the majority of the tuples in the training set with $X = x$ and $A = a$ have $Y = 1$, and predicts $\hat{Y} = 0$, otherwise. This assumption is quite natural, as, in general, an ML model should opt for the prevailing decision seen in training[4]. We make the same assumption for the LDP model $\mathcal{M}'$ (with A replaced by A'), which is reasonable since $\mathcal{M}$ and $\mathcal{M}'$ are trained with the same algorithm. Formally:

Assumption 4.1 . *The prediction of $\mathcal{M}$ (baseline model) is:*

$$\hat{Y}_a^x = \begin{cases} 1 & \text{if} \quad \Delta_a^x \geq 0 \quad (\text{or, equivalently, } \Gamma_a^x \geq 0), \\ 0 & \text{otherwise.} \end{cases}$$

Assumption 4.2 . *The prediction of $\mathcal{M}'$ (LDP model) is:*

$$\hat{Y}_a'^x = \begin{cases} 1 & \text{if} \quad \Delta_a'^x \geq 0 \quad (\text{or, equivalently, } \Gamma_a'^x \geq 0), \\ 0 & \text{otherwise.} \end{cases}$$

The following Lemma relates the difference between the frequencies of positive and negative decisions in the obfuscated and original data. We recall that $p = e^\epsilon/(e^\epsilon+1)$ is the probability that the value reported by RR is the true value.

Lemma 4.1. $\Delta_a'^x = p\ \Delta_a^x + (1-p)\ \Delta_{\bar{a}}^x$.

See proof on page 13.

The following Lemma relates the LDP model's prediction to the original data's statistics. It follows simply by case analysis from Lemma 4.1 and Assumption 4.2.

[3]In this paper, we use the terms *obfuscated* and *randomized* exchangeably.

[4]Some learning algorithms like the Nearest Neighbours actually use a generalization of this criterion to produce the prediction.

Lemma 4.2.

$$\hat{Y}_a'^x = 1 \quad \text{if} \quad \begin{cases} \Delta_a^x,\ \Delta_{\bar{a}}^x \geq 0, & \text{or} \\ \Delta_a^x > 0 \quad \text{and} \quad \Delta_{\bar{a}}^x < 0 \quad \text{and} \quad e^\epsilon \geq -\Delta_{\bar{a}}^x/\Delta_a^x, & \text{or} \\ \Delta_a^x < 0 \quad \text{and} \quad \Delta_{\bar{a}}^x > 0 \quad \text{and} \quad e^\epsilon \leq -\Delta_{\bar{a}}^x/\Delta_a^x\ . \end{cases}$$

$$\hat{Y}_a'^x = 0 \quad \text{if} \quad \begin{cases} \Delta_a^x,\ \Delta_{\bar{a}}^x \leq 0 \quad \text{and at least one of them is strictly negative,} & \text{or} \\ \Delta_a^x > 0 \quad \text{and} \quad \Delta_{\bar{a}}^x < 0 \quad \text{and} \quad e^\epsilon < -\Delta_{\bar{a}}^x/\Delta_a^x, & \text{or} \\ \Delta_a^x < 0 \quad \text{and} \quad \Delta_{\bar{a}}^x > 0 \quad \text{and} \quad e^\epsilon > -\Delta_{\bar{a}}^x/\Delta_a^x\ . \end{cases}$$

4.1 Impact of LDP on conditional statistical disparity

In this section, we analyze the effect of RR on conditional statistical disparity with respect to a specific tuple of values x of the explaining variables. To do so, we compare CSD'_x, which represents the conditional statistical disparity of prediction of the LDP model, with CSD_x which is the one of the baseline model. Following the principle that fairness should be assessed on the true inputs, we define CSD'_x as:

$$\mathrm{CSD}'_x = \mathbb{P}[\hat{Y}' = 1 \mid X = x, A = 1] - \mathbb{P}[\hat{Y}' = 1 \mid X = x, A = 0].$$

Namely, the conditioning is on A and not on A'. Note that, since the models are deterministic, CSD_x and CSD'_x could equivalently be defined as:

$$\mathrm{CSD}_x = \hat{Y}_1^x - \hat{Y}_0^x \qquad \text{and} \qquad \mathrm{CSD}'_x = \hat{Y}_1'^x - \hat{Y}_0'^x\ .$$

The following theorem states the relation between CSD_x and CSD'_x.

Theorem 4.1 Impact of LDP on CSD_x.

(1) if $\mathrm{CSD}_x > 0$ *then* $0 \leq \mathrm{CSD}'_x \leq \mathrm{CSD}_x$

(2) if $\mathrm{CSD}_x < 0$ *then* $\mathrm{CSD}_x \leq \mathrm{CSD}'_x \leq 0$

(3) if $\mathrm{CSD}_x = 0$ *then* $\mathrm{CSD}'_x = \mathrm{CSD}_x = 0$

See proof on page 13.

Essentially, the above theorem says that CSD'_x is always sandwiched between CSD_x and 0. Namely, if, in the baseline model, there is discrimination against one group, then obfuscating A tends to reduce the discrimination. It never introduces discrimination against the other group. In one extreme case, it may leave things unchanged, while, in the opposite extreme case, it may remove the discrimination entirely. If, in the baseline model, we have conditional statistical parity ($\mathrm{CSD}_x = 0$), then obfuscating A maintains this property.

It is important to note that Theorem 4.1 does not depend on whether the *unprivileged* group is the minority or the majority of the population.

4.2 Impact of LDP on statistical disparity

Using the results of CSD_x, in this section, we analyze the impact of privacy on SD by comparing SD′ and SD, where SD′ is the statistical disparity of the prediction of the LDP model, defined as:

$$\mathrm{SD}' = \mathbb{P}[\hat{Y}' = 1 \mid A = 1] - \mathbb{P}[\hat{Y}' = 1 \mid A = 0]. \tag{5}$$

Again, note that we condition on A rather than A'.

We make the following assumption that we call the *uniform discrimination assumption*. Essentially, it says that if one group is discriminated against for some value $x*$ of X, then the other group cannot be discriminated against for other values x of X. This is a natural assumption in real-life scenarios. For example, consider an ML system that tries to predict whether to release an individual on parole, given the type of crime they have committed in the past. If the system (or the historical data in which it is trained) discriminates against an ethnic group in case of a minor crime, it would still discriminate against that same group in case of a major crime, or, at most, be fair. As another example, consider granting an application for a loan: If, for a certain amount of money requested, the applications from an ethnic group are accepted more frequently than those from the other group,

Table 1: Abbreviations and definitions used in the paper. $\hat{\mathbb{P}}$ denotes the empirical probability (frequency) on the training set.

Abbreviations
• $\hat{Y}_a^x \in \{0,1\}$: the prediction of the baseline model $\mathcal{M}$ on the input (x,a)
• $\text{CSD}_x = \mathbb{P}[\hat{Y}=1 \mid X=x, A=1] - \mathbb{P}[\hat{Y}=1 \mid X=x, A=0]$: Conditional statistical disparity in $\mathcal{M}$
• $\text{SD} = \mathbb{P}[\hat{Y}=1 \mid A=1] - \mathbb{P}[\hat{Y}=1 \mid A=0]$: statistical disparity in $\mathcal{M}$
• $\text{EOD} = \mathbb{P}[\hat{Y}=1 \mid Y=1, A=1] - \mathbb{P}[\hat{Y}=1 \mid Y=1, A=0]$: equal opportunity difference in $\mathcal{M}$
• $\hat{Y}_a'^x \in \{0,1\}$: the prediction of the LDP model $\mathcal{M}'$ on the input (x,a)
• $\text{CSD}'_x = \mathbb{P}[\hat{Y}'=1 \mid X=x, A=1] - \mathbb{P}[\hat{Y}'=1 \mid X=x, A=0]$: Conditional statistical disparity in $\mathcal{M}'$
• $\text{SD}' = \mathbb{P}[\hat{Y}'=1 \mid A=1] - \mathbb{P}[\hat{Y}'=1 \mid A=0]$: statistical disparity in $\mathcal{M}'$
• $\text{EOD}' = \mathbb{P}[\hat{Y}'=1 \mid Y=1, A=1] - \mathbb{P}[\hat{Y}'=1 \mid Y=1, A=0]$: equal opportunity difference in $\mathcal{M}'$
Definitions
• $\Delta_a^x = \hat{\mathbb{P}}[Y=1, X=x, A=a] - \hat{\mathbb{P}}[Y=0, X=x, A=a]$
• $\Gamma_a^x = \hat{\mathbb{P}}[Y=1 \mid X=x, A=a] - \hat{\mathbb{P}}[Y=0 \mid X=x, A=a]$
• $\Delta_a'^x = \hat{\mathbb{P}}[Y=1, X=x, A'=a] - \hat{\mathbb{P}}[Y=0, X=x, A'=a]$
• $\Gamma_a'^x = \hat{\mathbb{P}}[Y=1 \mid X=x, A'=a] - \hat{\mathbb{P}}[Y=0 \mid X=x, A'=a]$

it is unlikely that, for a different amount of money, the situation would be inverted.

Formally, the *uniform discrimination assumption* is stated as follows:

Assumption 4.3 . *Uniform discrimination assumption*

$$\textit{if } \exists x^* \ \Gamma_a^{x^*} > \Gamma_{\overline{a}}^{x^*} \textit{ then } \forall x \ \Gamma_a^x \geq \Gamma_{\overline{a}}^x$$

In the remainder of this section, we differentiate between two scenarios depending on whether X and A are independent. We will denote the case of independency by $X \perp A$, and the case of dependency by $X \not\perp A$[5].

4.2.1 *First scenario:* $X \perp A$.

We first consider the case of independency. We start by showing that we can quantitatively express SD in terms of the distribution of the data as follows:

Lemma 4.3 (Quantification of SD)**.**

$$\text{SD} = \begin{cases} \mathbb{P}[\Delta_1^X \geq 0 \ \wedge \ \Delta_0^X < 0] & \textit{if } \exists x \ \Gamma_1^x > \Gamma_0^x \\ 0 & \textit{if } \forall x \ \Gamma_1^x = \Gamma_0^x \\ -\mathbb{P}[\Delta_1^X < 0 \ \wedge \ \Delta_0^X \geq 0] & \textit{if } \exists x \ \Gamma_1^x < \Gamma_0^x \end{cases}$$

See proof on page 14.

Analogously, we have:

Lemma 4.4 (Quantification of SD′)**.**

$$\text{SD}' = \begin{cases} \mathbb{P}[\Delta_1'^X \geq 0 \ \wedge \ \Delta_0'^X < 0] & \textit{if } \exists x \ \Gamma_1'^x > \Gamma_0'^x \\ 0 & \textit{if } \forall x \ \Gamma_1'^x = \Gamma_0'^x \\ -\mathbb{P}[\Delta_1'^X < 0 \ \wedge \ \Delta_0'^X \geq 0] & \textit{if } \exists x \ \Gamma_1'^x < \Gamma_0'^x \end{cases}$$

[5]In real-life contexts, X and A are usually dependent.

See proof on page 14.

Using Lemma 4.1, by case analysis, the quantification of SD′ can be reformulated in terms of the distribution in the original data, as follows.

Lemma 4.5 (Quantification of SD′ in terms of the distribution on the original data)**.**

$$\text{SD}' = \begin{cases} \mathbb{P}\left[\begin{array}{l} \Delta_1^X > 0 \ \wedge \ \Delta_0^X < 0 \ \wedge \\ e^\varepsilon \geq -\Delta_0^X/\Delta_1^X \ \wedge \ e^\varepsilon > -\Delta_1^X/\Delta_0^X \end{array} \right] & \textit{if } \exists x \ \Gamma_1^x > \Gamma_0^x \\ 0 & \textit{if } \forall x \ \Gamma_1^x = \Gamma_0^x \\ -\mathbb{P}\left[\begin{array}{l} \Delta_1^X < 0 \ \wedge \ \Delta_0^X > 0 \ \wedge \\ e^\varepsilon > -\Delta_0^X/\Delta_1^X \ \wedge \ e^\varepsilon \geq -\Delta_1^X/\Delta_0^X \end{array} \right] & \textit{if } \exists x \ \Gamma_1^x < \Gamma_0^x \end{cases}$$

We can now state the main result of this section: If $X \perp A$, then, like in the case of conditional statistical disparity, we have that SD′ is always sandwiched between SD and 0.

Theorem 4.2 Impact of LDP on SD. Case $X \perp A$**.**

(1) $\textit{if } \text{SD} > 0 \textit{ then } 0 \leq \text{SD}' \leq \text{SD}$

(2) $\textit{if } \text{SD} < 0 \textit{ then } \text{SD} \leq \text{SD}' \leq 0$

(3) $\textit{if } \text{SD} = 0 \textit{ then } \text{SD}' = \text{SD} = 0$

The proof follows immediately from Lemmas 4.3 and 4.5 because the values of X that constitute the probability mass in the expression of SD′ are a subset of those that constitute the probability mass in the expression of SD.

Discussion. Theorem 4.2 means that, from an unfair situation (SD > 0 or SD < 0), obfuscating the sensitive attribute A in general advantages the *unprivileged* group, but it never ends up discriminating the other group. (We will see in the next section that this is not always the case when some proxies to the sensitive attribute A exist in the data.)

In one extreme case, the situation does not change (SD′ = SD). By looking at the expression quantifying SD and SD′ in Lemma 4.3 and 4.5, we can see

that this happens when the noise we inject is small, i.e., for high values of ε, and, more precisely, when ε satisfies $\forall x\ \varepsilon \geq \max\{\ln(-\Delta_0^x/\Delta_1^x), \ln(-\Delta_1^x/\Delta_0^x)\}$.

In the opposite extreme case, the discrimination is totally eliminated ($SD' = 0$). This last case raises when we inject enough noise, and more precisely, when ε satisfies $\forall x\ \varepsilon < \max\{\ln(-\Delta_0^x/\Delta_1^x), \ln(-\Delta_1^x/\Delta_0^x)\}$.

In all the other cases, i.e., when for some x we have:
$\varepsilon \geq \max\{\ln(-\Delta_0^x/\Delta_1^x), \ln(-\Delta_1^x/\Delta_0^x)\}$ and for other x we have:
$\varepsilon < \max\{\ln(-\Delta_0^x/\Delta_1^x), \ln(-\Delta_1^x/\Delta_0^x)\}$, obfuscation removes some discrimination, but not entirely. Namely $0 < SD' < SD$ if SD is positive, or $SD < SD' < 0$ if SD is negative.

Note that the extreme case in which ε is 0 is equivalent to eliminating A entirely from the data. Hence, the takeout of this section is that the disparity between groups can be eliminated by removing the sensitive attribute, but it is important to remember that this is true only because there are no proxies to the sensitive attribute in the data ($X \perp A$).

Again, we note that Theorem 4.2 does not depend on whether the *unprivileged* group is the minority or the majority of the population.

4.2.2 *Second scenario: $X \not\perp A$.*

Usually, proxy attributes to the sensitive attribute A exist in the data. In other words, A and X are dependent ($X \not\perp A$). In this section, we study the impact of privacy on SD when $X \not\perp A$. Theorem 4.3 presents the results of the impact of privacy on SD in this scenario.

Theorem 4.3 Impact of LDP on SD. Case $X \not\perp A$.

(1) if $\exists x\ \Gamma_1^x > \Gamma_0^x$ *then* $SD' \leq SD$

(2) if $\exists x\ \Gamma_1^x < \Gamma_0^x$ *then* $SD \leq SD'$

(3) if $\forall x\ \Gamma_1^x = \Gamma_0^x$ *then* $SD' = SD$

See proof on page 15.

4.2.3 *Discussion.*

Theorem 4.3 confirms that also in the case $X \not\perp A$, in general, the *unprivileged* group benefits from privacy, and again, it does not depend on the *privileged* group being the majority or not. This finding is validated by our experiments on both synthetic and real-world datasets (cf. Figures 6 and 8 in Section 5).

Theorem 4.3 differs from Theorem 4.2 mainly on two points. First, SD and SD′ can have opposite signs. In other words, from a scenario where there is discrimination against one group, for instance, the group $A = 0$ ($SD > 0$), we can have, after obfuscation, a discrimination against the other group $A = 1$ ($SD' < 0$). We can even have scenarios in which, after obfuscation, the magnitude of unfairness against the other group is higher than the original one. This result is quite surprising. We simulated such a scenario using synthetic data (S5) and presented the results in Figure 4 (Section 5).

Second, we note that in case 1 we can have $SD < 0$ despite the fact that $\exists x\ \Gamma_1^x > \Gamma_0^x$ (which, by the Assumption 4.3, implies that $\forall x\ \Gamma_1^x \geq \Gamma_0^x$), and similarly for case 3. From the proof of the above theorem, we can see that it is particularly likely to happen when $\mathbb{P}[X = x|A = 1] \ll [X = x|A = 0])$. This is a form of the *Simpson's paradox* called *Association Reversal* [34]: we have a scenario in which for all sub-populations (i.e., for all x) there is discrimination against one group, while when considering the whole population, the discrimination is against the other group. Note that privacy obfuscation does not break the paradox, because also SD′ is negative.

Another form of the *Simpson's paradox* called the *Yule's Association Paradox* [14] can happen when for all sub-populations, the model shows fair results (i.e., $\forall x\ \ CSD_x = 0$), while for the whole population, it shows unfair results ($SD \neq 0$). In Section 5, we generated a synthetic dataset (S4) to illustrate such a paradox. Note that in this case, the privacy obfuscation has no effect on fairness: all the metrics remain the same. Indeed, if $\forall x\ CSD_x = 0$, then $\forall x\ \ CSD'_x = 0$, and all the metrics under consideration in this paper are based on CSD'_x.

4.3 Impact of LDP on equal opportunity difference

This section considers the impact of privacy on EOD (Eq. 4). This notion of fairness, by contrast to SD (Eq. 2), considers, in addition to the prediction $\hat{Y}$, the true decision Y (cf. Equation 4).

The justification for the EOD as a notion of fairness is that Y is supposed to be reliable and not to incorporate any bias (Hardt et al. [22]). Hence, if $\hat{Y}$ is consistent with Y, the prediction should be fair as well. Furthermore, thanks to this compatibility, and in contrast to other notions of fairness, EOD is, in general, going well along with accuracy (although there are exceptions: [35] has shown that, for certain distributions, Equal Opportunity implies trivial accuracy). We capture this principle in Assumption 4.4 here below, which states that the true decision Y is independent of the sensitive attribute A given X.

Assumption 4.4 . *Reliable Y. The decision Y is independent of the sensitive attribute for any value of the explaining variable. Namely:*

$$\mathbb{P}[Y = 1 \mid X = x, A = 1] = \mathbb{P}[Y = 1 \mid X = x, A = 0].$$

The limitation of EOD is that the "true" Y may not always be available. In its stead, the data may contain decisions that have been made in the past (which may not always have been fair), or decisions based on some proxy to the true Y. In any case, Assumption 4.4, may not always be satisfied in the data. When it is satisfied, however, we can obtain a strong result about the effect of privacy on EOD, similar to the one for SD. This is expressed by the theorem below.

Theorem 4.4 Impact of LDP on EOD.

(1) if $EOD > 0$ *then* $0 \leq EOD' \leq EOD$

(2) if $EOD < 0$ *then* $EOD \leq EOD' \leq 0$

(3) if $EOD = 0$ *then* $EOD' = EOD = 0$

See proof on page 16.

We recall that the above theorem holds under Assumption 4.4. On the other hand, it is valid regardless of whether X and A are independent.

5 EXPERIMENTAL RESULTS AND DISCUSSION

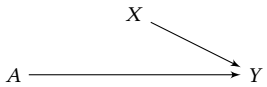
5.0.1 Main results and discussion. To validate our theoretical results, we have conducted a set of experiments on both synthetic and real-world fairness benchmark datasets. To each of these datasets, the fairness metrics presented in section 3.3 are applied to the baseline model $\mathcal{M}$ (model trained on the original samples) and to the LDP model $\mathcal{M}'$ (model trained on the obfuscated samples). Then, to assess the impact of privacy on fairness, in relation to Theorems 4.1, 4.2, 4.3, and 4.4, the predictions of these two models are compared. Recall that the testing samples for both $\mathcal{M}$ and $\mathcal{M}'$ are always kept original without obfuscation. We vary the privacy parameter ε in the $\{16, 8, 2, 1, 0.85, 0.5, 0.4, 0.3, 0.2, 0.1\}$ for the synthetic datasets and in the $\varepsilon = \{16, 8, 5, 4, 3, 2, 1, 0.5\}$ for the real-world datasets. At $\varepsilon = 0.1$ (strong privacy), the ratio of probabilities is bounded by $\varepsilon^{0.1} \approx 1.05$, giving nearly indistinguishable distributions between the two groups, whereas at $\varepsilon = 16$ (weak privacy), the distributions are nearly the same as in the original data.

5.1 Data and Experiments

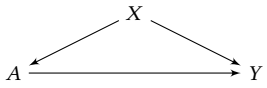
Environment: All the experiments are implemented in Python 3. We use *Random Forest* model [11] for classification with its default hyper-parameters and randomly select 80% as the training set and the remaining 20% as the testing set. For the RR mechanism, we use the implementation in Multi-Freq-LDPy [6].

Stability: Since LDP protocols, train/test splitting, and ML algorithms are randomized, we report average results over 100 runs.

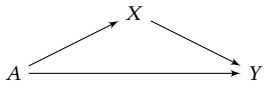
Datasets: We validate our theoretical results with six synthetic datasets (Section 5.2) and four real-world datasets (Section 5.3).



(a) Causal Graph for S1.



(b) Causal Graph for S2 and S6.



(c) Causal Graph for S3, S4 and S5.

Figure 2: Causal graphs of the Synthetic Datasets.

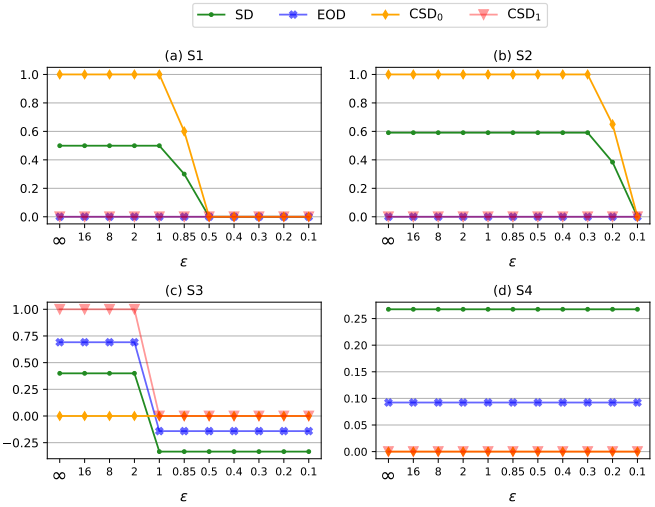


Figure 3: Results for the synthetic dataset S1-S4, illustrating the impact of LDP on fairness (y-axis) for privacy level ε (x-axis). Note that in S3 we have $X \not\perp A$ and the fairness measure SD is inverted after obfuscation. Also, EOD is inverted after obfuscation. This is because Assumption 4.4 is not verified in this dataset. S4 illustrates Yule's Association Paradox, a variant of the *Simpson's paradox*. The fairness values on the original data (no privacy) are the values for $\varepsilon = \infty$.

5.2 Synthetic Datasets

The causal graphs used to generate the synthetic datasets are depicted in Figure 2, and the joint empirical probabilities (frequencies) for the various combinations of values are shown in Table 2. S1 differs from all other datasets in that X and A are independent, whereas in all other datasets, namely S2-S6, X and A are dependent. A and Y are binary variables while X is a discrete variable. In S1, S2, and S4, X is also a binary variable. S3 and S5 are generated to simulate the scenario where privacy shifts discrimination between groups. S5 shows an extreme scenario where $|\text{SD}'| > |\text{SD}|$, while $|\text{SD}'| < |\text{SD}|$ in S3. And finally, S4 includes a case of *Yule's Association Paradox* [14] (Section 4.2.2).

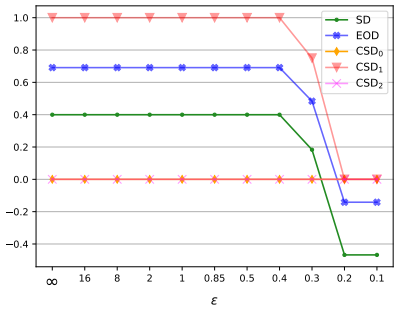


Figure 4: Results for the synthetic dataset S5. Note that EOD is also inverted here after obfuscation. Again, this is because Assumption 4.4 is not verified in this dataset.

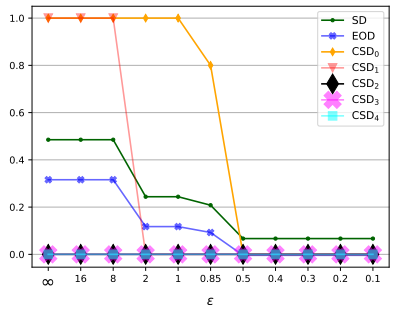


Figure 5: Results for the synthetic dataset S6.

In the plots that follow, the vertical dashed line in each plot shows the fairness values when the model is trained on original samples (no privacy). It turns out that they are always the same as for the level of privacy $\varepsilon = 16$.

Figure 3 shows the obtained results for S1-S4 presented above while S5 and S6 results are depicted in Figure 4 and Figure 5, respectively. For example, in S1, where some fairness measures show fair results in the baseline model $\mathcal{M}$, namely EOD and CSD_1, enforcing privacy helped maintain these fair results: $\text{SD}' = 0$ and $\text{CSD}'_1 = 0$. However, some fairness measures show unfair results against group $A = 0$ in the baseline model, namely SD, and CSD_0; thus, enforcing privacy removed discrimination when enough noise is added. In particular, at $\varepsilon = \ln(-\Delta^x_0/\Delta^x_1) = 0.85$, SD' and CSD'_0 values started to decrease and continued to decrease reaching full parity between groups.

As we proved theoretically in Theorem 4.3, and explained in Section 4.2.2, in S3 and S5 and from a scenario where SD and EOD show discrimination against the group $A = 0$, by applying privacy, the discrimination became against the other group $A = 1$. Note that this does not contradict Theorem 4.4, because S3 and S5 do not verify Assumption 4.4[6]. For S3, although this inversion of fairness conclusions (discrimination switching from one group to another when applying privacy), the disparity after obfuscation decreased: $|\text{SD}'| < |\text{SD}|$ ($|\text{SD}'| = 0.33$ and $|\text{SD}| = 0.39$). However, S5 (Figure 4) shows an extreme case where the disparity between groups after obfuscation increased: $|\text{SD}'| > |\text{SD}|$ ($|\text{SD}'| = 0.46$ and $|\text{SD}| = 0.39$). In other words, not only has the discrimination switched from one group to the other after obfuscation, but also the level of unfairness has increased.

[6]We provide in Appendix A.2 a dataset called S7 that satisfies Assumption 4.4

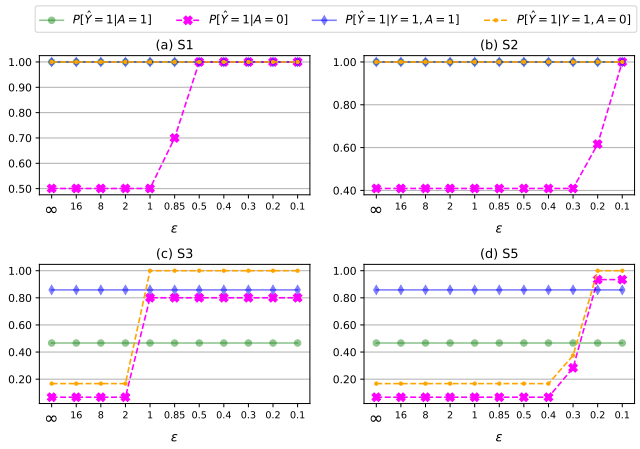


Figure 6: Impact of LDP on disparity (y-axis) by varying the privacy level ε (x-axis) showing the behavior of fairness measures on groups separately when applying privacy. For readability, only SD and EOD are illustrated. Results for the synthetic datasets S1, S2, S3, and S5.

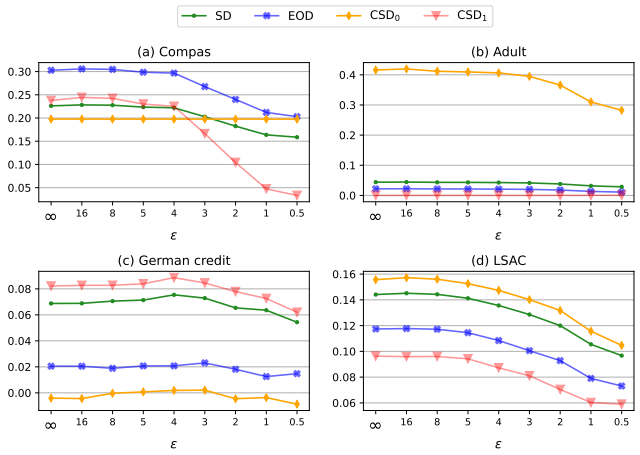


Figure 7: Results for the real-world datasets. The *German credit* dataset does not satisfy Assumption 4.3, which explains its unstable behavior.

S4 shows a case of the *Yule's Association Paradox* [14], a variant of the *Simpson's paradox*. That is, the model $\mathcal{M}$ shows fair results for all subpopulations: $\text{CSD}_0 = \text{CSD}_1 = 0$. However, $\mathcal{M}$ shows unfair results for the whole population: SD= 0.26. As shown in Figure 3(d), the paradox stayed even under a strong privacy regime ($\varepsilon = 0.1$) and hence obfuscating the sensitive attribute solely didn't remove the paradox from the data.

To better understand how privacy impacts fairness, the plots in Figure 6 show how the impact of privacy on $\mathbb{P}[\hat{Y} = 1 \mid A = a]$ and $\mathbb{P}[\hat{Y} = 1 \mid Y = 1, A = a]$ for both groups $A = 1$ and $A = 0$ while varying ε.

As mentioned in Section 4.2.2, the *unprivileged* group $A = 0$ benefits more from privacy. In other words, when obfuscating the sensitive attribute and aligning with our Theorems 4.1- 4.4, the results of the acceptance rates and the true positive rates of the unprivileged group tend to increase. For instance, for all the synthetic datasets, it is clear that it is group $A = 0$ who advantages from privacy as shown in Figure 6. In other words, there is an increasing trend of $\mathbb{P}[\hat{Y} = 1 \mid A = 0]$ and $\mathbb{P}[\hat{Y} = 1 \mid Y = 1, A = 0]$.

5.3 Real-world Datasets

We consider the following four real-world datasets:

- *Compas*: This 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, age, arrest date, etc.) were gathered by ProPublica [4], and the goal is to predict a risk score of recidivism[7]. 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 ($A = 1$ for non-black individuals and $A = 0$ for black individuals) as the sensitive attribute and the risk of recidivism as the outcome. $Y = 1$ designates a low risk score of recidivism while $Y = 0$ denotes a high risk score. The number of priors of an individual is used as an explaining variable to compute CSD_x where $X = 1$ denotes a high number of priors, and $X = 0$ denotes a low number of priors.
- *Adult* [15]: This dataset [15] 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. Gender is the sensitive attribute ($A = 1$ for men and $A = 0$ for women), and income is the outcome where $Y = 1$ designates a high income while $Y = 0$ denotes a low income. Education level is the attribute used as an explaining variable to compute CSD_x where $X = 1$ denotes a high education level and $X = 0$ denotes a low education level.
- *German credit* [16]: This dataset includes data of 1000 individuals applying for loans. This dataset is designed for binary classification to predict whether an individual will default on the loan ($Y = 0$) or not ($Y = 1$) based on personal attributes such as gender, job, credit amount, credit history, etc. We consider gender the sensitive attribute where female applicants ($A = 0$) are compared to male applicants ($A = 1$). Credit history is the explaining attribute used to compute CSD_x where $X = 1$ denotes an applicant who has duly repaid in the past while $X = 0$ denotes a critical account for which the applicant has had late payments and/or defaults in the past.
- *LSAC*: This dataset originates from the Law School Admissions Council (LSAC) National Bar Passage Study [40]. The outcome, denoted as "pass bar", indicates whether a candidate has successfully passed the bar exam ($Y = 1$) or not ($Y = 0$). The prediction is based on personal information such as race, gender, family income, LSAT, undergraduate GPA score, etc. The sensitive attribute is race ($A = 0$ for blacks and $A = 1$ for other ethnic groups). The explaining variable is the undergraduate GPA score of an applicant where $X = 1$ indicates a high GPA and $X = 0$ denotes a low GPA.

The real-world datasets' distributions are shown in Table 3.

Figure 7 shows the results of applying privacy on the four real-world datasets. As with the synthetic datasets and in alignment with our proofs, obfuscating the sensitive attribute tends to improve the fairness metrics considered in this study in all the datasets *except the German credit* one (we will discuss this latter case below). We believe that this is due to the fact that the real-world datasets do not always follow the "ideal" situation represented by our assumptions. In particular, the datasets we are considering contain other variables besides the X that we use as an explaining variable, which can influence the prediction.

[7]The risk of recidivism in the *Compas* dataset is a value between 1 and 10 where the higher the score, the higher the risk of recidivism for the defendant.

For instance, in the *Compas* dataset, starting from discrimination against black individuals ($A = 0$), privacy reduced the disparity from SD = 0.21 to 0.15. Similarly, privacy decreased discrimination against black individuals from SD = 0.13 to SD = 0.09 in the *LSAC* dataset, and a similar decrease pattern is observed for all the other fairness measures. We notice a very slight increase of CSD_1 at $\varepsilon = 1$, mainly due to ML inaccuracy. The *Adult* dataset also shows a slight disparity decrease caused by privacy. However, starting from a very high disparity between groups given a low level of education ($\text{CSD}_0 = 0.39$), the disparity is reduced to 0.22 at $\varepsilon = 0.1$.

Concerning the *German credit* dataset, the results show an unstable trend. This is because this data set does not satisfy the *uniform discrimination* assumption (Assumption 4.3). Indeed, we for $X = 0$, we have, for group $A = 1$:

$$\begin{aligned}\Gamma_1^0 &= \mathbb{P}[Y = 1 \mid X = 0, A = 1] - \mathbb{P}[Y = 0 \mid X = 0, A = 1] \\ &= \frac{0.23}{0.29} - \frac{0.06}{0.29} \\ &\approx 0.58\end{aligned}$$

while for the same $X = 0$, for group $A = 0$ we have:

$$\begin{aligned}\Gamma_0^0 &= \mathbb{P}[Y = 1 \mid X = 0, A = 0] - \mathbb{P}[Y = 0 \mid X = 0, A = 0] \\ &= \frac{0.08}{0.09} - \frac{0.01}{0.09} \\ &\approx 0.77\end{aligned}$$

Hence $\Gamma_1^0 < \Gamma_0^0$.
On the other hand, for $X = 1$ and group $A = 1$ we have:

$$\begin{aligned}\Gamma_1^1 &= \mathbb{P}[Y = 1 \mid X = 1, A = 1] - \mathbb{P}[Y = 0 \mid X = 1, A = 1] \\ &= \frac{0.27}{0.40} - \frac{0.13}{0.40} \\ &\approx 0.35\end{aligned}$$

while for the same $X = 1$, for group $A = 0$ we have:

$$\begin{aligned}\Gamma_0^1 &= \mathbb{P}[Y = 1 \mid X = 1, A = 0] - \mathbb{P}[Y = 0 \mid X = 1, A = 0] \\ &= \frac{0.13}{0.22} - \frac{0.09}{0.22} \\ &\approx 0.18\end{aligned}$$

Hence, $\Gamma_1^1 > \Gamma_0^1$, which means that the *German credit* dataset does not satisfy Assumption 4.3 It may also mean that the attribute "Credit history" is badly chosen as an explaining variable, and/or that it is not the main attribute influencing the decision.

To better understand how privacy impacts fairness, the plots in Figure 8 show how the impact of privacy on $\mathbb{P}[\hat{Y} = 1 \mid A = a]$ and $\mathbb{P}[\hat{Y} = 1 \mid Y = 1, A = a]$ for both groups $A = 1$ and $A = 0$ while varying ε.

For instance, for the *Adult* dataset, we can observe that women's acceptance rate ($\mathbb{P}[\hat{Y} = 1 \mid A = 0]$) and true positive rate increase ($\mathbb{P}[\hat{Y} = 1 \mid Y = 1, A = 0]$) from 0.91 to 0.93 and from 0.96 to 0.99, respectively. However, no change is observed for men ($A = 1$) even at strong privacy ($\varepsilon = 0.5$). A similar behavior is observed for the *LSAC* dataset. For the *Compas* dataset, while no change is observed for the black defendants' ($A = 0$) rates, a decrease is observed for the non-black defendants ($A = 1$). Similar behavior is also observed for the *German credit* dataset, where a slight increase in the acceptance rate and the true positive rate for women is observed while almost no change is observed for men.

5.4 LDP impact on model accuracy

Figures 9 and 10 illustrate the impact of LDP on the accuracy of the model for the synthetic datasets and the real-world datasets, respectively. From these figures, one can note that, in general, the impact of obfuscating the sensitive attribute on model accuracy of the real-world datasets is minor. The drop in the utility is more apparent for the synthetic datasets but remains reasonable with a maximum drop of 0.2 in S2.

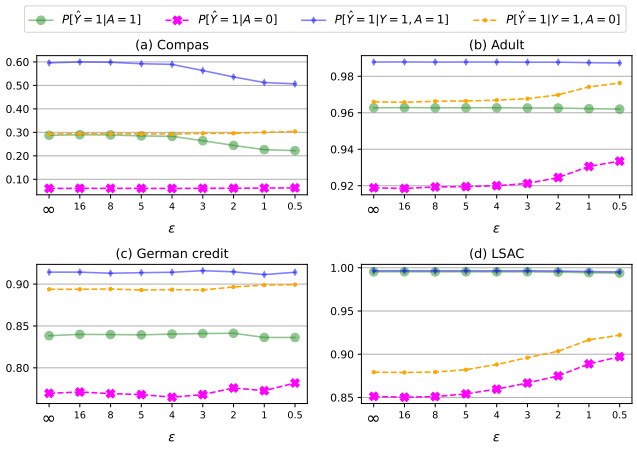


Figure 8: Impact of LDP on disparity (y-axis) by varying the privacy level ε (x-axis) showing the behavior of fairness measures on groups separately when applying privacy. For readability, only SD and EOD are illustrated. Results for the real-world datasets.

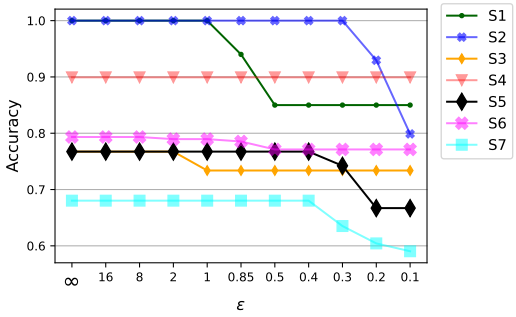


Figure 9: Impact of LDP on the model accuracy for the synthetic datasets.

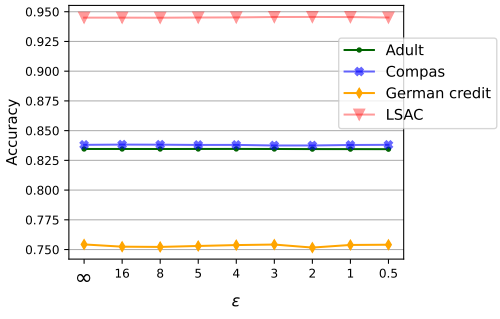


Figure 10: Impact of LDP on the model accuracy for the real-world datasets.

6 CONCLUSION

This study formally examines how LDP affects fairness. More specifically, we provide bounds in terms of the joint distributions and the privacy level,

delimiting the extent to which LDP can impact the fairness of the model. Our findings show that the *unprivileged* group benefits more than the *privileged* group when injecting enough noise into the sensitive attribute. Furthermore, for conditional statistical disparity and for equal opportunity difference, injecting noise, in general, improves fairness. This also holds for statistical disparity when the data contain no proxies to the sensitive attribute. However, when the data contains proxies, in certain cases, by injecting enough noise, while the discrimination was originally against one group, it may be shifted to the other group after obfuscation, and the level of unfairness may be worse than before. Note that none of our results depend on whether the *unprivileged* group is the minority or the majority. Additionally, our work focuses on the RR mechanism, a fundamental LDP protocol [23] that serves as a building block for more complex LDP mechanisms (e.g., [10, 19, 38]).

In future work, we aim to extend our work to more fairness measures, particularly overall accuracy equality and others. We also believe that hiding only the sensitive attribute is crucial but not sufficient because proxies for this attribute may exist in the data and thus reveal sensitive information. Therefore, we plan to study formally the impact of LDP on multidimensional data.

Acknowledgement. This work was partially supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement 835294) and by the "ANR 22-PECY-0002" IPOP (Interdisciplinary Project on Privacy) project of the Cybersecurity PEPR. The work of Héber H. Arcolezi has been partially supported by MIAI @ Grenoble Alpes ("ANR-19-P3IA-0003").

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] Sushant Agarwal. Trade-offs between fairness and interpretability in machine learning. In *IJCAI 2021 Workshop on AI for Social Good*, 2021.

[3] 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.

[4] 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.

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

[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. *arXiv preprint arXiv:2304.12845*, 2023.

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

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

[10] Raef Bassily and Adam Smith. Local, private, efficient protocols for succinct histograms. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, page 127–135, New York, NY, USA, 2015. Association for Computing Machinery.

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

[12] 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.

[13] 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.

[14] H. A. David and A. W. F. Edwards. *Yule's Paradox ("Simpson's Paradox")*, pages 137–143. Springer New York, New York, NY, 2001.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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, New York, NY, USA, 2014. ACM.

[20] 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.

[21] 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.

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

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

[24] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. *The Journal of Machine Learning Research*, 17(1):492–542, 2016.

[25] Faisal Kamiran, Indrė Zliobaite, and Toon Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and information systems (Print)*, 35(3):613–644, 2013.

[26] 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.

[27] Karima Makhlouf, Heber H. Arcolezi, Sami Zhioua, Ghassen Ben Brahim, and Catuscia Palamidessi. On the impact of multi-dimensional local differential privacy on fairness. *arXiv preprint arXiv:2312.04404*, 2023.

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

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

[30] Paul Mangold, Michaël Perrot, Aurélien Bellet, and Marc Tommasi. Differential privacy has bounded impact on fairness in classification. *International Conference on Machine Learning, 2023*, 2023.

[31] 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.

[32] 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.

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

[34] Judea Pearl. Comment: understanding simpson's paradox. In *Probabilistic and causal inference: The works of judea Pearl*, pages 399–412. The American Statistician, 2022.

[35] Carlos Pinzón, Catuscia Palamidessi, Pablo Piantanida, and Frank Valencia. On the impossibility of non-trivial accuracy in presence of fairness constraints. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 7993–8000. AAAI Press, 2022.

[36] David Pujol, Ryan McKenna, Satya Kuppam, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 189–199, 2020.

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

[38] Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. Locally differentially private protocols for frequency estimation. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 729–745, Vancouver, BC, August 2017. USENIX Association.

[39] Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, March 1965.

[40] Linda F Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998.

[41] Xingxing Xiong, Shubo Liu, Dan Li, Zhaohui Cai, and Xiaoguang Niu. A comprehensive survey on local differential privacy. *Security and Communication Networks*, 2020:1–29, October 2020.

[42] 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.

Table 2: Distributions of the synthetic datasets.

(a) S1.

$Y=1$	$X=0$	$X=1$
$A=1$	0.35	0.35
$A=0$	0	0.15
$Y=0$	$X=0$	$X=1$
$A=1$	0	0
$A=0$	0.15	0

(b) S2.

$Y=1$	$X=0$	$X=1$
$A=1$	0.28	0.38
$A=0$	0	0.12
$Y=0$	$X=0$	$X=1$
$A=1$	0	0
$A=0$	0.22	0

(c) S3.

$Y=1$	$X=0$	$X=1$	$X=2$
$A=1$	0.03	0.17	0.03
$A=0$	0	0.17	0.03
$Y=0$	$X=0$	$X=1$	$X=2$
$A=1$	0.24	0.03	0
$A=0$	0.1	0.2	0

(d) S4.

$Y=1$	$X=0$	$X=1$
$A=1$	0	0.4
$A=0$	0.03	0.34
$Y=0$	$X=0$	$X=1$
$A=1$	0.03	0.07
$A=0$	0.13	0

(e) S5.

$Y=1$	$X=0$	$X=1$	$X=2$
$A=1$	0.03	0.17	0.03
$A=0$	0	0.17	0.03
$Y=0$	$X=0$	$X=1$	$X=2$
$A=1$	0.24	0.03	0
$A=0$	0.03	0.27	0

(f) S6.

$Y=1$	$X=0$	$X=1$	$X=2$	$X=3$	$X=4$
$A=1$	0.05	0.08	0.09	0.13	0.14
$A=0$	0.02	0.03	0.06	0.03	0.04
$Y=0$	$X=0$	$X=1$	$X=2$	$X=3$	$X=4$
$A=1$	0.04	0.02	0.01	0.06	0
$A=0$	0.06	0.04	0.02	0.08	0

Table 3: Distributions of the real-world datasets.

(a) Compas.

$Y=1$		$X=0$	$X=1$
	$A=1$	0.12	0.03
	$A=0$	0.06	0.03
$Y=0$			
	$A=1$	0.15	0.1
	$A=0$	0.25	0.26

(b) Adult.

$Y=1$		$X=0$	$X=1$
	$A=1$	0.06	0.53
	$A=0$	0.02	0.21
$Y=0$			
	$A=1$	0.03	0.06
	$A=0$	0.02	0.07

(c) German credit.

$Y=1$		$X=0$	$X=1$
	$A=1$	0.23	0.27
	$A=0$	0.08	0.13
$Y=0$			
	$A=1$	0.06	0.13
	$A=0$	0.01	0.09

(d) LSAC.

$Y=1$		$X=0$	$X=1$
	$A=1$	0.43	0.47
	$A=0$	0.03	0.01
$Y=0$			
	$A=1$	0.02	0.02
	$A=0$	0.01	0.01

A APPENDIX

A.1 Proofs

Lemma 4.1. $\Delta'^{x}_{a} = p\ \Delta^{x}_{a} + (1-p)\ \Delta^{x}_{\overline{a}}$.

Proof of Lemma 4.1.

$$\begin{aligned}
\Delta'^{x}_{a} &= \hat{\mathbb{P}}[Y=1, X=x, A'=a] - \hat{\mathbb{P}}[Y=0, X=x, A'=a] \\
&= p\ \hat{\mathbb{P}}[Y=1, X=x, A=a] + (1-p)\hat{\mathbb{P}}[Y=1, X=x, A=\overline{a}] - \left(p\ \hat{\mathbb{P}}[Y=0, X=x, A=a] + (1-p)\hat{\mathbb{P}}[Y=0, X=x, A=\overline{a}]\right) \\
&= p\left(\hat{\mathbb{P}}[Y=1, X=x, A=a] - \hat{\mathbb{P}}[Y=0, X=x, A=a]\right) + (1-p)\left(\hat{\mathbb{P}}[Y=1, X=x, A=\overline{a}] - \hat{\mathbb{P}}[Y=0, X=x, A=\overline{a}]\right) \\
&= p\ \Delta^{x}_{a} + (1-p)\ \Delta^{x}_{\overline{a}}
\end{aligned}$$

□

Theorem 4.1 Impact of LDP on CSD_x.

(1) if $\mathrm{CSD}_x > 0$ *then* $0 \le \mathrm{CSD}'_x \le \mathrm{CSD}_x$

(2) if $\mathrm{CSD}_x < 0$ *then* $\mathrm{CSD}_x \le \mathrm{CSD}'_x \le 0$

(3) if $\mathrm{CSD}_x = 0$ *then* $\mathrm{CSD}'_x = \mathrm{CSD}_x = 0$

Proof of Theorem 4.1.

(1) if $\mathrm{CSD}_x > 0$ then, according to Assumption 4.1, $\hat{Y}_1 = 1$ and $\hat{Y}_0 = 0$.
Hence, $\Delta^x_1 \ge 0$ and $\Delta^x_0 < 0$.
Using Lemma 4.2, we have:

$$\hat{Y}'^{x}_{1} = \begin{cases} 1 \text{ if } \Delta^x_1 > 0 \text{ and } e^{\varepsilon} \ge -\Delta^x_0/\Delta^x_1 \\ 0 \text{ if } (\Delta^x_1 > 0 \text{ and } e^{\varepsilon} < -\Delta^x_0/\Delta^x_1) \text{ or } \Delta^x_1 = 0 \end{cases}$$

and

$$\hat{Y}'^{x}_{0} = \begin{cases} 1 \text{ if } \Delta^x_1 > 0 \text{ and } e^{\varepsilon} \le -\Delta^x_1/\Delta^x_0 \\ 0 \text{ if } (\Delta^x_1 > 0 \text{ and } e^{\varepsilon} > -\Delta^x_1/\Delta^x_0) \text{ or } \Delta^x_1 = 0 \end{cases}$$

Consequently, three scenarios are possible:

- $\hat{Y}'^{x}_{1} = 0 \wedge \hat{Y}'^{x}_{0} = 0$ if $\Delta^x_1 = 0$ or $\Delta^x_1 > 0$ and $e^{\varepsilon} < -\Delta^x_0/\Delta^x_1$ and $e^{\varepsilon} > -\Delta^x_1/\Delta^x_0$
- $\hat{Y}'^{x}_{1} = 1 \wedge \hat{Y}'^{x}_{0} = 0$ if $\Delta^x_1 > 0$ and $e^{\varepsilon} \ge -\Delta^x_0/\Delta^x_1$ and $e^{\varepsilon} > -\Delta^x_1/\Delta^x_0$
- $\hat{Y}'^{x}_{1} = 1 \wedge \hat{Y}'^{x}_{0} = 1$ if $\Delta^x_1 > 0$ and $e^{\varepsilon} \ge -\Delta^x_0/\Delta^x_1$ and $e^{\varepsilon} \le -\Delta^x_1/\Delta^x_0$

Note that the case $\hat{Y}'^{x}_{1} = 0 \ \wedge\ \hat{Y}'^{x}_{0} = 1$ is not possible. Indeed, $\hat{Y}'^{x}_{1} = 0 \ \wedge\ \hat{Y}'^{x}_{0} = 1$ implies $e^{\varepsilon} < -\Delta^x_0/\Delta^x_1$ and $e^{\varepsilon} \le -\Delta^x_1/\Delta^x_0$. Note that the two fractions are one the inverse of the other. Hence, one of them is smaller than 1, or both are 1. Therefore, we would have $e^{\varepsilon} < 1$, which is not possible because $\varepsilon \ge 0$.
Hence we have $\mathrm{CSD}'_x = 0$ or $\mathrm{CSD}'_x = 1$, i.e., $0 \le \mathrm{CSD}'_x \le \mathrm{CSD}_x$.

(2) Case 2 ($\mathrm{CSD}_x < 0$) is analogous to case 1. That is, proving this case amounts to replacing 0 by 1 and 1 by 0 in case 1 proof.

(3) if $\mathrm{CSD}_x = 0$, two cases are possibles:
- $\hat{Y}^x_1 = 0 \wedge \hat{Y}^x_0 = 0$. This means that $\Delta^x_1 < 0 \wedge \Delta^x_0 < 0$. By Lemma 4.2, we derive $\hat{Y}'^{x}_{1} = 0 \wedge \hat{Y}'^{x}_{0} = 0$. Hence, $\mathrm{CSD}'_x = 0$.
- $\hat{Y}^x_1 = 1 \wedge \hat{Y}^x_0 = 1$. This means that $\Delta^x_1 \ge 0 \wedge \Delta^x_0 \ge 0$. By Lemma 4.2, we derive $\hat{Y}'^{x}_{1} = 1 \wedge \hat{Y}'^{x}_{0} = 1$. Hence, $\mathrm{CSD}'_x = 0$.

□

Lemma 4.3 (Quantification of SD)**.**

$$\mathrm{SD} = \begin{cases} \mathbb{P}[\Delta^X_1 \ge 0 \ \wedge\ \Delta^X_0 < 0] & \text{if } \exists x\ \Gamma^x_1 > \Gamma^x_0 \\ 0 & \text{if } \forall x\ \Gamma^x_1 = \Gamma^x_0 \\ -\mathbb{P}[\Delta^X_1 < 0 \ \wedge\ \Delta^X_0 \ge 0] & \text{if } \exists x\ \Gamma^x_1 < \Gamma^x_0 \end{cases}$$

Proof of Lemma 4.3.

$$
\begin{aligned}
\mathrm{SD} &\stackrel{\text{def}}{=} \mathbb{P}[\hat{Y}=1|A=1] - \mathbb{P}[\hat{Y}=1|A=0] \\
&= \sum_x \mathbb{P}[\hat{Y}=1, X=x|A=1] - \sum_x \mathbb{P}[\hat{Y}=1, X=x|A=0] \\
&= \sum_x \mathbb{P}[\hat{Y}=1|X=x, A=1]\cdot\mathbb{P}[X=x|A=1] - \sum_x \mathbb{P}[\hat{Y}=1|X=x, A=0]\cdot\mathbb{P}[X=x|A=0] \\
&\stackrel{(a)}{=} \sum_x \hat{Y}_1^x\ \mathbb{P}[X=x|A=1] - \sum_x \hat{Y}_0^x\ \mathbb{P}[X=x|A=0] \\
&\stackrel{(b)}{=} \sum_x \hat{Y}_1^x\ \mathbb{P}[X=x] - \sum_x \hat{Y}_0^x\ \mathbb{P}[X=x] \\
&\stackrel{(c)}{=} \sum_{\substack{x:\\ \Delta_1^x\geq 0}} \mathbb{P}[X=x] - \sum_{\substack{x:\\ \Delta_0^x\geq 0}} \mathbb{P}[X=x]
\end{aligned} \tag{6}
$$

In step (a), we replace $\mathbb{P}[\hat{Y}=1|X=x, A=1]$ and $\mathbb{P}[\hat{Y}=1|X=x, A=0]$ by their corresponding abbreviated forms $\hat{Y}_1^x$ and $\hat{Y}_0^x$. Step (b) follows from $X \perp A$. Step (c) follows because $\hat{Y}_1^x = 1$ when $\Delta_1^x \geq 0$, and $\hat{Y}_1^x = 0$, otherwise. Similarly, $\hat{Y}_0^x = 1$ when $\Delta_0^x \geq 0$, and $\hat{Y}_0^x = 0$, otherwise.

Then, we consider three cases:

- Case $\exists x\ \Gamma_1^x > \Gamma_0^x$. By Assumption 4.3 (uniform discrimination) we have that $\forall x\ \Gamma_1^x \geq \Gamma_0^x$. Also, recall that $\Gamma_a^x \geq 0$ if and only if $\Delta_a^x \geq 0$. Therefore, in the expression (6), for each x such that $\Delta_0^x \geq 0$, we also have $\Delta_1^x \geq 0$, which concludes the proof for this case.
- Case $\forall x\ \Gamma_1^x = \Gamma_0^x$. We have that $\Delta_0^x \geq 0$ if and only if $\Delta_1^x \geq 0$, hence the two terms in the expression (6) are equal.
- Case $\exists x\ \Gamma_1^x < \Gamma_0^x$. This is the symmetric of the first case. Following the same reasoning (with 0 and 1 exchanged), we have that, in the expression (6), for each x such that $\Delta_1^x \geq 0$, we also have $\Delta_0^x \geq 0$.

□

Lemma 4.4 (Quantification of SD′)**.**

$$
\mathrm{SD}' = \begin{cases} \mathbb{P}[\Delta_1'^X \geq 0\ \wedge\ \Delta_0'^X < 0] & \text{if } \exists x\ \Gamma_1'^x > \Gamma_0'^x \\ 0 & \text{if } \forall x\ \Gamma_1'^x = \Gamma_0'^x \\ -\mathbb{P}[\Delta_1'^X < 0\ \wedge\ \Delta_0'^X \geq 0] & \text{if } \exists x\ \Gamma_1'^x < \Gamma_0'^x \end{cases}
$$

Proof of Lemma 4.4.

$$
\begin{aligned}
\mathrm{SD}' &\stackrel{\text{def}}{=} \mathbb{P}[\hat{Y}'=1|A=1] - \mathbb{P}[\hat{Y}'=1|A=0] \\
&= \sum_x \mathbb{P}[\hat{Y}'=1, X=x|A=1] - \sum_x \mathbb{P}[\hat{Y}'=1, X=x|A=0] \\
&= \sum_x \mathbb{P}[\hat{Y}'=1|X=x, A=1]\cdot\mathbb{P}[X=x, A=1] - \sum_x \mathbb{P}[\hat{Y}'=1|X=x, A=0]\cdot\mathbb{P}[X=x, A=0] \\
&= \sum_x \hat{Y}'^x_1\ \mathbb{P}[X=x|A=1] - \sum_x \hat{Y}'^x_0\ \mathbb{P}[X=x|A=0] \\
&= \sum_{\substack{x:\\ \Delta_1'^x\geq 0}} \mathbb{P}[X=x] - \sum_{\substack{x:\\ \Delta_0'^x\geq 0}} \mathbb{P}[X=x]
\end{aligned}
$$

The proof proceeds like the one in Lemma 4.3. We need, however, the following result, which states that LDP obfuscation preserves *uniform discrimination assumption* (Assumption 4.3).

Lemma A.1. *If* $\exists x^*\ \Gamma_a'^{x^*} > \Gamma_{\overline{a}}'^{x^*}$ *then* $\forall x\ \Gamma_a'^x \geq \Gamma_{\overline{a}}'^x$

Proof

We prove the property by showing that $\Gamma_a'^x > \Gamma_{\overline{a}}'^x$ if and only if $\Gamma_a^x > \Gamma_{\overline{a}}^x$, and that $\Gamma_a'^x = \Gamma_{\overline{a}}'^x$ if and only if $\Gamma_a^x = \Gamma_{\overline{a}}^x$. Then, clearly, the statement of the theorem derives from the assumption of *uniform discrimination* for the original data (before obfuscation).

It is easy to see that

$$
\Gamma_a'^x = \frac{p\Delta_a^x + (1-p)\Delta_{\overline{a}}^x}{p\mathbb{P}[X=x, A=a] + (1-p)\mathbb{P}[X=x, A=\overline{a}]}
$$

Let us prove that $\Gamma'^x_a > \Gamma'^x_{\overline{a}}$ if and only if $\Gamma^x_a > \Gamma^x_{\overline{a}}$:

$$\begin{array}{rcl}
\Gamma'^x_a & > & \Gamma'^x_{\overline{a}} \\
& \Leftrightarrow & \\
\frac{p\Delta^x_a+(1-p)\Delta^x_{\overline{a}}}{p\mathbb{P}[X=x,A=a]+(1-p)\mathbb{P}[X=x,A=\overline{a}]} & > & \frac{p\Delta^x_{\overline{a}}+(1-p)\Delta^x_a}{p\mathbb{P}[X=x,A=\overline{a}]+(1-p)\mathbb{P}[X=x,A=a]} \\
& \Leftrightarrow & \\
p^2\Delta^x_a\mathbb{P}[X=x,A=\overline{a}] + (1-p)^2\Delta^x_{\overline{a}}\mathbb{P}[X=x,A=a] & > & p^2\Delta^x_{\overline{a}}\mathbb{P}[X=x,A=a] + (1-p)^2\Delta^x_a\mathbb{P}[X=x,A=\overline{a}] \\
& \Leftrightarrow & \\
\left.\begin{array}{r} p^2\Gamma^x_a\mathbb{P}[A=a,X=x]\mathbb{P}[A=\overline{a},X=x] \\ + \\ (1-p)^2\Gamma^x_{\overline{a}}\mathbb{P}[A=\overline{a},X=x]\mathbb{P}[A=a,X=x] \end{array}\right\} & > & \left\{\begin{array}{l} p^2\Gamma^x_{\overline{a}}\mathbb{P}[A=\overline{a},X=x]\mathbb{P}[A=a,X=x] \\ + \\ (1-p)^2\Gamma^x_a\mathbb{P}[A=\overline{a},X=x]\mathbb{P}[A=a,X=x] \end{array}\right. \\
& \Leftrightarrow & \\
\mathbb{P}[A=\overline{a},X=x]\mathbb{P}[A=a,X=x]\left(p^2\Gamma^x_a + (1-p)^2\Gamma^x_{\overline{a}}\right) & > & \mathbb{P}[A=\overline{a},X=x]\mathbb{P}[A=a,X=x]\left(p^2\Gamma^x_{\overline{a}} + (1-p)^2\Gamma^x_a\right) \\
& \Leftrightarrow & \\
p^2\left(\Gamma^x_a - \Gamma^x_{\overline{a}}\right) & > & (1-p)^2\left(\Gamma^x_a - \Gamma^x_{\overline{a}}\right) \\
& \Leftrightarrow & \\
\Gamma^x_a & > & \Gamma^x_{\overline{a}}
\end{array}$$

The property $\Gamma'^x_a = \Gamma'^x_{\overline{a}}$ if and only if $\Gamma^x_a = \Gamma^x_{\overline{a}}$ can be proved similarly, just replace the "$>$" symbol by "$=$". □

Theorem 4.3 Impact of LDP on SD. Case $X \not\perp A$.

(1) if $\exists x\ \Gamma^x_1 > \Gamma^x_0$ *then* $\mathrm{SD}' \leq \mathrm{SD}$

(2) if $\exists x\ \Gamma^x_1 < \Gamma^x_0$ *then* $\mathrm{SD} \leq \mathrm{SD}'$

(3) if $\forall x\ \Gamma^x_1 = \Gamma^x_0$ *then* $\mathrm{SD}' = \mathrm{SD}$

Proof of Theorem 4.3. We prove the result for the case $\exists x\ \Gamma^x_1 > \Gamma^x_0$, the other two cases can be proven similarly.
Recall that:

$$\mathrm{SD} = \sum_{\substack{\boldsymbol{x}: \\ \Delta^x_1 \geq 0}} \mathbb{P}[X=x|A=1] - \sum_{\substack{\boldsymbol{x}: \\ \Delta^x_0 \geq 0}} \mathbb{P}[X=x|A=0] \quad \hat{Y}^x_1, \hat{Y}^x_0 = 1$$

Since we are considering the case $\exists x\ \Gamma^x_1 > \Gamma^x_0$, from Assumption 4.3 (*uniform discrimination*) we derive $\forall x\ \Gamma^x_1 \geq \Gamma^x_0$, hence:

$$\mathrm{SD} = \sum_{\substack{\boldsymbol{x}: \\ \Delta^x_1, \Delta^x_0 \geq 0}} \mathbb{P}[X=x|A=1] - \mathbb{P}[X=x|A=0] \ + \sum_{\substack{\boldsymbol{x}: \\ \Delta^x_1 \geq 0, \Delta^x_0 < 0}} \mathbb{P}[X=x|A=1]$$

After obfuscation, from Lemma A.1 we have that $\forall x\ \Gamma'^x_1 > \Gamma'^x_0$. Hence:

$$\begin{aligned}
\mathrm{SD}' &= \sum_{\substack{\boldsymbol{x}: \\ \Delta'^x_1, \Delta'^x_0 \geq 0}} \mathbb{P}[X=x|A=1] - \mathbb{P}[X=x|A=0] \ + \sum_{\substack{\boldsymbol{x}: \\ \Delta'^x_1 \geq 0, \Delta'^x_0 < 0}} \mathbb{P}[X=x|A=1] \\
&= \sum_{\substack{\boldsymbol{x}: \\ \Delta^x_1, \Delta^x_0 \geq 0}} \mathbb{P}[X=x|A=1] - \mathbb{P}[X=x|A=0] \ + \sum_{\substack{\boldsymbol{x}: \\ \Delta^x_1 > 0, \Delta^x_0 < 0, \\ -\frac{\Delta^x_0}{\Delta^x_1} \leq e^\varepsilon \leq -\frac{\Delta^x_1}{\Delta^x_0}}} \mathbb{P}[X=x|A=1] - \mathbb{P}[X=x|A=0] \\
&\quad + \sum_{\substack{\boldsymbol{x}: \\ \Delta^x_1 \geq 0, \Delta^x_0 < 0 \\ e^\varepsilon \geq -\frac{\Delta^x_0}{\Delta^x_1}, \\ e^\varepsilon > -\frac{\Delta^x_1}{\Delta^x_0}}} \mathbb{P}[X=x|A=1]
\end{aligned}$$

By case analysis, and similar to the proof of Theorem 4.2, we can conclude Theorem 4.3. The main difference with Theorem 4.2, is that SD' contains the additional term

$$\sum_{\substack{\boldsymbol{x}: \\ \Delta^x_1 > 0, \Delta^x_0 < 0, \\ -\frac{\Delta^x_0}{\Delta^x_1} \leq e^\varepsilon \leq -\frac{\Delta^x_1}{\Delta^x_0}}} \mathbb{P}[X=x|A=1] - \mathbb{P}[X=x|A=0]$$

which can be negative and large enough to cause SD' to go below 0. Hence SD' and SD can be of opposite signs. □

Theorem 4.4 Impact of LDP on EOD.

(1) if $\mathrm{EOD} > 0$ *then* $0 \leq \mathrm{EOD}' \leq \mathrm{EOD}$

(2) if $\text{EOD} < 0$ *then* $\text{EOD} \leq \text{EOD}' \leq 0$

(3) if $\text{EOD} = 0$ *then* $\text{EOD}' = \text{EOD} = 0$

Proof of Theorem 4.4.

$$
\begin{aligned}
\text{EOD} &\overset{\text{def}}{=} \mathbb{P}[\hat{Y}=1|Y=1,A=1] - \mathbb{P}[\hat{Y}=1|Y=1,A=0] \\
&= \sum_{x} \mathbb{P}[\hat{Y}=1, X=x|Y=1,A=1] - \sum_{x} \mathbb{P}[\hat{Y}=1, X=x|Y=1,A=0] \\
&= \sum_{x} \mathbb{P}[\hat{Y}=1|X=x,Y=1,A=1] \cdot \mathbb{P}[X=x|Y=1,A=1] - \sum_{x} \mathbb{P}[\hat{Y}=1|X=x,Y=1,A=0] \cdot \mathbb{P}[X=x|Y=1,A=0] \\
&= \sum_{x} \frac{\mathbb{P}[\hat{Y}=1,Y=1|X=x,A=1]}{\mathbb{P}[Y=1|X=x,A=1]} \cdot \mathbb{P}[X=x|Y=1,A=1] - \sum_{x} \frac{\mathbb{P}[\hat{Y}=1,Y=1|X=x,A=0]}{\mathbb{P}[Y=1|X=x,A=0]} \cdot \mathbb{P}[X=x|Y=1,A=0] \\
&\overset{(a)}{=} \sum_{\substack{x: \\ \Delta_1^x \geq 0}} \mathbb{P}[X=x|Y=1,A=1] - \sum_{\substack{x: \\ \Delta_0^x \geq 0}} \mathbb{P}[X=x|Y=1,A=0] \\
&\overset{(b)}{=} \sum_{\substack{x: \\ \Delta_1^x \geq 0}} \mathbb{P}[X=x|Y=1] - \sum_{\substack{x: \\ \Delta_0^x \geq 0}} \mathbb{P}[X=x|Y=1]
\end{aligned}
$$

(a) follows from the fact that both $\frac{\mathbb{P}[\hat{Y}=1,Y=1|X=x,A=1]}{\mathbb{P}[Y=1|X=x,A=1]}$ and $\frac{\mathbb{P}[\hat{Y}=1,Y=1|X=x,A=0]}{\mathbb{P}[Y=1|X=x,A=0]}$ are equal to 1 for $x : \Delta_1^x \geq 0$ and $x : \Delta_0^x \geq 0$, respectively. And (b) follows because of the reliability assumption 4.4.

Now, after obfuscation and following the same reasoning as in the proofs of Theorems 4.2 and 4.3, we have:

$$
\begin{aligned}
\text{EOD}' = &\sum_{\substack{x: \\ \Delta_1^x, \Delta_0^x \geq 0}} \mathbb{P}[X=x|Y=1] - \mathbb{P}[X=x|Y=1] \quad + \sum_{\substack{x: \\ \Delta_1^x > 0, \Delta_0^x < 0, \\ -\frac{\Delta_0^x}{\Delta_1^x} \leq e^{\epsilon} \leq -\frac{\Delta_1^x}{\Delta_0^x}}} \mathbb{P}[X=x|Y=1] - \mathbb{P}[X=x|Y=1] \\
&+ \sum_{\substack{x: \\ \Delta_1^x, \Delta_0^x \geq 0, \\ e^{\epsilon} \geq -\frac{\Delta_0^x}{\Delta_1^x}, \\ e^{\epsilon} > -\frac{\Delta_1^x}{\Delta_0^x}}} \mathbb{P}[X=x|Y=1] - \mathbb{P}[X=x|Y=1]
\end{aligned}
$$

The rest is deduced by case analysis.

□

A.2 Results for S7

Below are the data distribution (Table 4) and the results of the dataset S7. The data was generated following the causal graph depicted in Figure 2(c). The results of applying privacy on fairness are illustrated in Figure 11. Note that in this dataset, the Assumption 4.4 is satisfied.

Table 4: Distributions of the synthetic dataset S7.

$Y=1$		$X=0$	$X=1$	$X=2$	$X=3$	$X=4$
	$A=1$	0.05	0.07	0.04	0.06	0.05
	$A=0$	0.05	0.07	0.04	0.06	0.05
$Y=0$		$X=0$	$X=1$	$X=2$	$X=3$	$X=4$
	$A=1$	0	0.06	0.05	0.02	0
	$A=0$	0.09	0.04	0.06	0.02	0.12

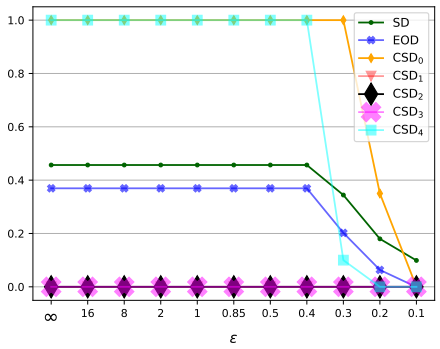


Figure 11: Results for the synthetic dataset S7.