`P2NIA`: Privacy-Preserving Non-Iterative Auditing

Jade Garcia Bourrée[1,4] ✉, Hadrien Lautraite[2], Sébastien Gambs[2], Gilles Tredan[3], Erwan Le Merrer[1,4], and Benoît Rottembourg[4]

[1] Univ Rennes, Inria, CNRS, Irisa
[2] Université du Québec à Montréal
{gambs.sebastien,lautraite.hadrien}@uqam.ca
[3] LAAS/CNRS gtredan@laas.fr
[4] Inria {jade.garcia-bourree,erwan.le-merrer, benoit.rottembourg}@inria.fr

Abstract. The emergence of AI legislation has increased the need to assess the ethical compliance of high-risk AI systems. Traditional auditing methods rely on platforms' application programming interfaces (APIs), in which responses to queries are examined through the lens of fairness requirements. However, such approaches put a significant burden on platforms, as they are forced to maintain APIs while ensuring privacy, facing the possibility of data leaks. This lack of proper collaboration between the two parties, in turn, causes a significant challenge to the auditor, who is subject to estimation bias as they are unaware of the data distribution of the platform. To address these two issues, we present `P2NIA`, a novel auditing scheme that proposes a mutually beneficial collaboration for both the auditor and the platform. Extensive experiments demonstrate `P2NIA`'s effectiveness in addressing both issues. In summary, our work introduces a privacy-preserving and non-iterative audit scheme that enhances fairness assessments using synthetic or local data, avoiding the challenges associated with traditional API-based audits.

Keywords: Algorithm auditing · synthetic data · local differential privacy · fairness estimation.

1 Introduction

Algorithm auditing refers to the evaluation of algorithmic decision-making systems. More precisely, it aims at ensuring their privacy, transparency, fairness and compliance with ethical and legal standards [19, 37]. This field is very active, in reaction to algorithms becoming increasingly ubiquitous in our daily lives in critical areas such as finance, human resources, healthcare or justice [2, 7, 12, 14, 27, 28, 34, 48, 51]. The traditional way to assess if decision-making systems and models in production satisfy ethical standards is to audit them in the so-called *black-box setting*, in which the auditor sends queries to a platform, receives answers and infers information on its behavior with respect to fairness, for instance.

Issues with black-box audits. Unfortunately, it has been shown that, in many real-world scenarios, black-box audits may not lead to accurate evaluation [5,10,11]. The main reason for this being the non-collaboration of platforms that do not want to release information about their data distribution, which is kept hidden for privacy or trade secret motives. Consider, for instance, the scenario of a bank that decides to lend to whom by using a risk scoring model. The bank uses the model to predict if clients are likely to repay the loan based on their economic profile and an auditor wants to assess the fairness of such a model. To realize this, requests composed of possible profiles are sent to the platform for assessment. In this scarce data regime, audit results can lead to erroneous conclusions, possibly due to the auditor's measurement bias, whether deliberate or subconscious or to wrong assumptions regarding the audited model. For instance, if the auditor's requests do not align with the data distribution used to train the platform model, it will lead to biased conclusions. In addition, many audit scenarios deal with sensitive data such as personal health records, income levels or demographic information like age, gender and ethnicity [29]. Such data is highly critical and cause platforms to be reluctant to open and maintain APIs to expose them, as requested by recent legislation such as the AI act [24].

Contributions. We propose a novel scheme to address these issues in a setting in which the platform and the auditor collaborate for the audit. Our solutions lead to a mutually beneficial situation in which 1) the auditor can perform an unbiased estimation of the property of interest (hereafter, we focus on fairness), and 2) the platform does not need to maintain APIs while ensuring the privacy of its data. More precisely, we first demonstrate both theoretically and with an example that in a non-collaborative audit setup, in which the auditor faces a black-box setting (*i.e.*, without being provided the data distribution of the platform), an auditor obtains a biased estimate of the model under scrutiny. This motivates our scheme P2NIA (which stands for Privacy-Preserving Non-Iterative Auditing), in which the platform collaborates with the auditor by releasing a synthetic dataset mimicking its behavior, allowing for unbiased audits while ensuring privacy. In summary, our contributions are as follows.

- We demonstrate and illustrate that an audit outcome can be biased in a black-box setting due to the *population bias*.
- To circumvent this issue, we propose P2NIA, a novel collaborative auditing scheme benefiting both the platform and the auditor by enabling accurate audits while being privacy-preserving and non-iterative.
- We have designed several ways to implement P2NIA and experimentally evaluate their performance.

Outline. First, we present the challenges associated with black-box auditing schemes in Section 2 before introducing P2NIA as an alternative approach to address them in Section 3. Then, we evaluate our proposal on two datasets and compare it to a standard audit scheme in Section 4. Afterward, we review related works in Section 5 before concluding in Section 6.

2   Auditing Black-Box Models

After formalizing the audit objective in this section, we focus on the potential bias in black-box fairness estimation. In particular, studying a common hypothesis that an auditor needs to have knowledge of the platform distribution, we theoretically show that imperfect knowledge of this distribution can lead to a linear estimation bias on the statistical distance between the two distributions. We also illustrate how such bias will likely appear in practice through a simulation.

2.1   Auditing Fairness Setup

Fairness Definitions. The concept of fairness refers to the property that algorithms should not discriminate against people based on sensitive characteristics, such as ethnical origin or gender [5]. Fairness can be defined in many ways, but we focus on *group* fairness metrics, which compare the statistics of predictions between two groups, one of which is considered potentially discriminated and is called the protected group. In particular, we focus on *demographic parity*, *equality of opportunity* and *equalized odds*, as these are arguably the most studied group fairness notions [43].

Formally, we aim at auditing model $m \in \mathcal{H} : \mathcal{X} \to \mathcal{Y}$, with $\mathcal{X}$ is its input space. Let $\mathcal{D}$ be the input distribution and $\mu_\mathcal{D}$ the target measure of the audit. Usually, this notation is shorthanded by μ. However, this notation hides a strong prerequisite for measuring fairness, namely having access to the input distribution $\mathcal{D}$. In addition, group fairness requires the definition of a protected group, usually defined through a protected attribute $A \in \{0, 1\}$. We denote Y as the target variable and $\hat{Y} = m(X)$ as the prediction of model m on input X.

The fairness metrics we study can be formally defined as follows [9, 21, 32]:

Demographic parity: $\mu_\mathcal{D} = \left| P_\mathcal{D}[\hat{Y} = 1 | A = 1] - P_\mathcal{D}[\hat{Y} = 1 | A = 0] \right|$.

Equalized odds: $\mu_\mathcal{D} = \max_{y \in \{0,1\}} \left| P_\mathcal{D}[\hat{Y} = 1 | Y = y, A = 0] - P_\mathcal{D}[\hat{Y} = 1 | Y = y, A = 1] \right|$.

Equality of opportunity: $\mu_\mathcal{D} = \left| P_\mathcal{D}[\hat{Y} = 1 | Y = 1, A = 0] - P_\mathcal{D}[\hat{Y} = 1 | Y = 1, A = 1] \right|$.

In a nutshell, each of these metrics addresses different facets of potential bias and ensures that models perform equitably across groups. For instance, demographic parity aims at maintaining similar outcome distributions across groups, while equalized odds and equal opportunity assess fairness by comparing error rates with the latter, specifically focusing on true positive rates. Note that the last two fairness metrics require access to the true label Y of samples, which might be a strong assumption in practice.

Black-Box Auditing Setting. When assessing the fairness of a model, particularly in usual non-collaborative setups, auditors are unlikely to possess information about the distribution of training data. Hence, they will have to rely on black-box auditing strategies, which generally entails crafting a specific finite set of queries $Q = \{q_1, \ldots, q_n\}$ in the model input space $\mathcal{X}$, and obtaining the corresponding

honest[5] model inferred answers $m(Q) = \{m(q_1), \ldots, m(q_n)\}$. The pair $(Q, m(Q))$ is then processed by the auditor to produce the fairness estimates. In the specific case of equality of opportunity and equalized odds, the auditor needs domain expertise to access the true outcome Y of their requests Q.

A crucial assumption for the auditing to work is that $Q \sim \mathcal{D}$, which is equivalent to the auditor having sampling access to the input distribution[6]. Unfortunately, in a non-collaborative audit, the platform will likely keep this information confidential for privacy and trade secret reasons, as observed with recent legislation enforcing the opening of APIs [24].

2.2   Estimation Discrepancy Due to Population Bias

When the auditor does not have access to $\mathcal{D}$, they must rely on a different distribution $\mathcal{D}'$ to generate their audit queries. For instance, when auditing platforms such as TikTok or Snapchat, already known user profiles from other platforms like LinkedIn or Facebook can be used. Unfortunately, such distributions differ substantially, for example in terms of users' average age [33]. Another possible approach is to generate data from statistics published by the platforms themselves. However, public statistics are often insufficient to generate realistic data, particularly due to correlations between attributes [1]. The ignorance of $\mathcal{D}$ thus lead to an inaccurate fairness estimation $\mu_{\mathcal{D}}(m) \neq \mu_{D'}(m)$. Hereafter, we refer to this as a form of *population bias*. This term was initially introduced to qualify the bias arising when a population $x \sim \mathcal{D}$ uses a classifier trained using samples drawn from a distribution $\mathcal{D}' \neq \mathcal{D}$ [43]. In our setting, population bias is used to qualify the bias due to the evaluation of models using a biased query set, *i.e.* from a trained model on the platform using $x \sim \mathcal{D}$, then audited with a population D'. While this form of bias has been known in other contexts, to the best of our knowledge it has not been studied in the field of model audits in the black-box setting. For instance, Casper and co-authors [11] acknowledge the possible presence of bias in the black-box setup but without identifying the precise bias cause. To highlight the impact of such population bias in black-box audits, we first construct pathological examples demonstrating to which extent the estimation can be biased, before showing through simulation that population bias also has practical impacts, even in simpler scenarios.

Theorem 1. *Population bias in black-box audit. Estimating the fairness metric (*e.g., *demographic parity, equality of opportunity or equalized odds) of a model m using a distribution $\mathcal{D}'$ at a (total variation) statistical distance α from the true distribution $\mathcal{D}$ can lead to an approximation error linear in α. Formally: $\exists \mathcal{D}, \mathcal{D}', m$ s.t. $\delta(\mathcal{D}, \mathcal{D}') = \sup_{F \in \mathcal{F}} |\mathcal{D}(F) - \mathcal{D}'(F)| = \alpha$ and $m \in \mathcal{H} : \mathcal{X} \mapsto \mathcal{Y}$ s.t. $|\mu_{\mathcal{D}}(m) - \mu_{\mathcal{D}'}(m)| = \alpha$.*

---

[5] Previous works have shown that manipulative platforms may impede fairness assessment [30].

[6] In some settings, the auditor may want to deviate from $\mathcal{D}$ to optimize his sampling strategy and obtain more accurate estimates [53], but this nevertheless requires knowing $\mathcal{D}$.

Proof. First, we look at the case of demographic parity and proceed by constructing the situation considered. Let $X = a \in \{0,1\} \times b \in \{0,1\} \times c \in [0,1]$ be the input space and $Y = \{0,1\}$ the binary output space. In the following, a will play the role of the protected attribute, b will enable the tuning of the statistical distance between D and D', and c represents additional data used by the model to construct its decision. We now describe two distributions D, D' on X:

- for both D and D' we have: $c \perp\!\!\!\perp (a,b), c \sim U(0,1)$ and $a \perp\!\!\!\perp b$.
- $\mathbb{E}_D(a=1) = \mathbb{E}_{D'}(a=1) = 1/2$
- $\mathbb{E}_D(b=1) = 0$ and $\mathbb{E}_{D'}(b=1) = \alpha$

As a table:

$x \sim$	$a=1, b=1$	$a=0, b=1$	$a=1, b=0$	$a=0, b=0$
D	0	0	1/2	1/2
D'	$\alpha/2$	$\alpha/2$	$(1-\alpha)/2$	$(1-\alpha)/2$

Clearly, with $\mathcal{F}$ being a sigma-algebra of X, we have

$$\delta(D, D') = sup_{F \in \mathcal{F}} |P_D(F) - P_{D'}(F)| = P_{D'}(b=1) - P_D(b=1) = \alpha.$$

In other words, the total variation distance between D and D' is α.

The model $M : X \mapsto Y$ is defined as follows: $M(x) = (1-b).(c > 1/2) + b.a$. The rationale here is that b drives two distinct behaviors of M, only one of which is unfair. The demographic parity can be expressed as given D: $\mu_{x \sim \mathcal{D}}(M) = |P(Y|a) - P(Y|\bar{a})| = P(c > 1/2) - P(c \leq 1/2) = 0$. Conversely, given D', using the law of total probability on the partitions induced by b: $\mu_{x \sim \mathcal{D}'} = (P(Y|a,b) - P(Y|\bar{a},b))P(b) + (P(Y|a,\bar{b}) - P(Y|\bar{a},\bar{b}))(1 - P(b)) = (1-0)\alpha + 0 = \alpha$. Hence, $|\mu_{x \sim \mathcal{D}}(M) - \mu_{x \sim \mathcal{D}'}(M)| = \alpha$.

Second, we can generalize the previous proof to equality of opportunity and equalized odds by assessing that for all input, $Y = 1$. In such case, for all binary a, y, $P_{\mathcal{D}}[\hat{Y} = 1|Y = y, A = a] = P_{\mathcal{D}}[\hat{Y} = 1|Y = 1, A = a] = P_{\mathcal{D}}[\hat{Y} = 1|A = a]$ and the three metrics are equals. □

Intuitively, this result states that if the auditor has a bias in their prior about the distribution at play at the platform (*i.e.*, imperfect knowledge of the distribution), then the outcome is also a biased estimation. This is highly undesirable as it undermines the strength and legal impact of the audit.

While Theorem 1 corresponds to a crafted extreme scenario, we now study experimentally the impact of population bias on real data. To this end, we use the standard Folktables dataset commonly used in fairness settings [16]. Consider a platform that trained a model on data collected from a specific state (California) to predict whether an individual's income is above $50,000 [16]. We consider a black-box auditor that has no access to this training distribution, which uses the distribution of another state among the remaining fifty in the dataset. One might expect that the auditor and model distributions are sufficiently similar to

estimate the platform fairness accurately. However, as demonstrated by Figure 1, this is not the case. Thus, no matter the state the auditor selects for reference, there will be a considerable population bias in the auditor's estimation. More concretely, on average, the absolute error in estimating the demographic parity with a different state is around 0.13, which is high considering that the classical 80% rule translates into a maximal tolerated value of 0.2. Concretely, as the true demographic parity in this example is 0.01, using the distribution of (NE,SD,WY,ND,ID,UT) would lead the auditor to a wrong conclusion about the model.
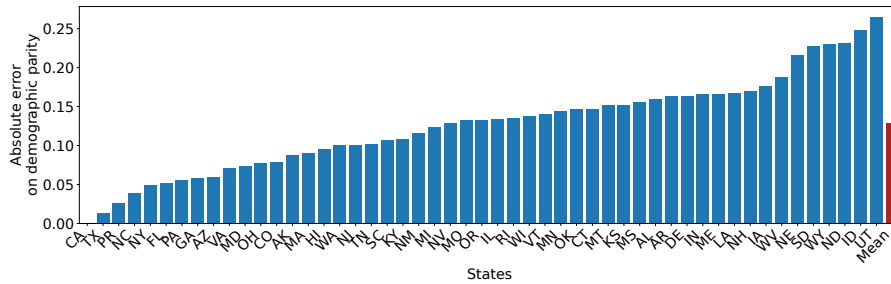

Fig. 1: Observed population bias on the evaluation of demographic parity with protected feature "sex" on a model trained in state CA to predict whether an individual's income is above $50,000$ and evaluated in other states. These results confirm the practical impact of Theorem 1.

This population bias is thus a clear impediment for accurate auditing in non-collaborative black-box audits, and is at stake even in such a simple and realistic scenario. To address this, in the following section, we propose P2NIA, a novel scheme enabling unbiased estimation while adding privacy protection with respect to the platform.

3   Auditing through a Differentially-Private Dataset

To tackle the aforementioned issues, we propose a collaborative scheme, P2NIA, which leverages privacy-preserving techniques for the benefit of the privacy protection of the platform while enabling accurate fairness auditing by the auditor.

3.1   P2NIA: A Non-Iterative Auditing scheme

We first present the P2NIA scheme before describing how it enables a fairness estimation that is 1) reliably performed and 2) without exposing the privacy of users (*i.e.*, the released dataset is differentially-private).
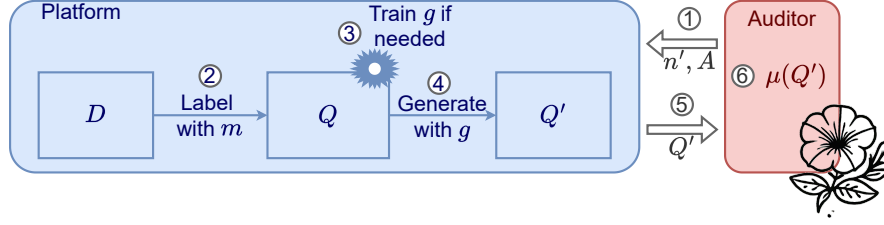
Fig. 2: Description of P2NIA scheme.

The Privacy-Preserving Non-Iterative Auditing scheme (P2NIA) works in six steps (Figure 2). First, P2NIA starts with the auditor specifying to the platform the desired number of queries n', along with A, the group of interest (**step 1**).

Upon receiving this request, the platform labels a part of its internal data (called *audit dataset* in the following) using its proprietary model m. More precisely the, model m is applied to the dataset D to generate a set of labeled queries Q, in which each query consists of a profile input X from the dataset, the target output Y and the corresponding predictions $\hat{Y}$ (**step 2**).

To protect sensitive information while still allowing fairness evaluation, the platform has two options. More precisely, either it anonymizes the dataset by using a local differential-private mechanism as described in Section 3.2, or it trains a generative model g build to produce synthetic data based on Q to reflect the model's behavior without directly disclosing the training user data (**step 3**). Afterwards, the platform creates the audit dataset by generating or anonymizing n' synthetic queries Q' with g (**step 4**).

Then, the platform releases Q' to the auditor (**step 5**). At this stage, the auditor receives the auditing datasets and evaluates fairness using a predefined fairness metric $\mu(Q')$, applied to the protected group A (**step 6**).

Observe that in our scheme, the auditor evaluates the model using Q', without directly querying the platform model (step 6) but rather by declaring their number of requests and their group of interest (step 1). Before the generation with a mechanism g, the queries in D are labeled using the model m (step 2) in order to reflect the platform's actual behavior. This step is necessary for certain mechanisms g that require labeled data as input to achieve better result.

Thus, as the platform has generated Q' from its audit dataset D with a mechanism that ensures differential privacy, the personal data remains confidential while making it possible to audit the model (steps 2-5). As differential privacy is immune to post-processing [22] (*i.e.*, it is impossible to compute a function of the output of the private algorithm and make it less differentially-private), the assessment of μ by the auditor does not reduce the privacy of the released data. For example, in the context of the income prediction model discussed in Section 1, this would amount to accurately auditing the fairness of a platform based on synthetic demographic data without revealing the record of any user.

3.2   Implementing `P2NIA` with Differentially-private Mechanisms

The `P2NIA` scheme is generic with respect to particular privacy mechanisms implementations captured as g in Figure 2. In this paper, we focus on differential privacy and explore two distinct mechanisms to achieve it: local differential privacy and synthetic data generation. We first present the differential privacy guarantees that we aim to achieve.

Differential privacy. Differential privacy is a formal privacy model that protects individuals by bounding the impact any individual can have on the output of an algorithm [22].

Definition 1. *Differential Privacy. [22] A randomized mechanism $\mathcal{M}$ satisfies ϵ-differential privacy if for any individual $x \in \mathcal{X}$, any dataset $Q \subset \mathcal{X}$ and any subset of possible outputs S:*

$$P\left[\mathcal{M}(Q \setminus \{x\}) \in S\right] \leq exp(\epsilon)P\left[\mathcal{M}(Q) \in S\right].$$

The parameter ϵ is called the privacy budget and is considered public. The smaller ϵ, the stronger the privacy guarantees. Typical values of ϵ could be 0.01, 0.1, or in some cases, $ln(2)$ or $ln(3)$ [20]. However, in practice, ϵ can vary up to 10 [38], as, for example, Apple reports using values of 4 or 8 [3].

Data anonymization through local differential privacy. One way to share data while respecting differential privacy is to add noise to the data. For instance, in the randomized response technique [56], a respondent answers a sensitive binary question (*e.g.*, "What is your gender") by tossing a biased coin in secret. If the result is tail, the respondent answers honestly. Otherwise, the coin is tossed again, and the answer returned is "Male" if head and "No" if tail. This ensures that the auditor can reliably estimate the unknown proportions of sensitive attributes without knowing any individual's true answer with certainty. With Simple Random Sampling With Replacement (SRSWR [15]), this generalized randomized response method provides an unbiased estimator of these proportions while also ensuring compliance with the ϵ-differential privacy for any ϵ by adjusting the bias of the coin. The randomized response technique can be generalized to non-binary attributes (*e.g.*, age or marital status) or to sets of attributes [13], which is called Generalized Randomized Response (GRR). In addition, it can be combined with pre-processing techniques [54] (instead of SRSWR) for better privacy guarantee, and recent advances have further improved the performance of GRR [6, 17, 23, 55, 57].

   P2NIA implementation: Since our study aims to highlight how `P2NIA` can be applied rather than providing an exhaustive comparison of private data methods, we proceed with the basic version of GRR: each colmun of the dataset (features, target and prediction) is independently flipped with a probability p selected to reach the desired privacy guarantee ϵ: $p = e^\epsilon/(e^\epsilon + k - 1)$, in which k is the number of values the feature can have.

P2NIA Fairness Estimation Reliability: As $\hat{Y}$ is computed by the platform knowing D, the resulting fairness estimator is only biased by the noise added by GRR. The auditor knows ϵ and hence p, so they can easily debias the anonymized results to recover the correct estimator value [13,17].

Mechanisms for Differentially-private Synthetic Data Generation. Instead of adding noise to existing data, another approach consists of generating synthetic data that follows the original distribution while protecting the privacy of individual records for the original dataset. Many differentially-private data generation approaches have been proposed in the literature to achieve this objective; the interested reader can refer to [49] or [26] for recent surveys on the subject. In particular, some synthesis methods offer the possibility to specify which relationship between variables should be maintained, aligning with our goal of reliable model evaluation. Specifically, we use MST [40] and AIM [41], which are two methods operating under the *select-measure-generate* framework. In a nutshell, both methods work by first selecting a set of low-dimensional marginal queries (*i.e.*, statistics on small subsets of attributes). Afterwards, each selected marginal is measured under differential privacy by adding noise to ensure privacy. This yields a collection of noisy marginal results. Finally, a graphical model called Private-PGM [42] is constructed from those noisy marginals to infer a consistent high-dimensional data distribution to generate new synthetic data points. MST and AIM differ in the way marginals are selected. More precisely, MST performs an initial noisy measurement of marginals using a small portion of the privacy budget and then selects which one should be measured more precisely. In contrast, AIM adaptively chooses new queries in iterative rounds, using feedback from previous measurements to guide where the privacy budget should be spent. This adaptive selection can improve accuracy at the expense of increased computational cost.

P2NIA implementation: We explicitly maintain the relationship between protected variable A, target variable Y and predictions $\hat{Y}$ from the audited model (*i.e.*, the marginals $(A,Y), (A,\hat{Y}), (Y,\hat{Y})$ and $(A,Y,\hat{Y})$) in addition to 12 2-way marginals chosen randomly to mimic the dataset. The relationship between the required privacy level ϵ and the noise introduced in the marginals differs between MST and AIM, which are detailed in the respective original papers.

P2NIA Fairness Estimation Reliability: As we explicitly conserve all marginals involved in the definition of all three fairness measures, the resulting value is unbiased.

Hence, we consider three distinct methods to implement P2NIA: GRR, MST and AIM. While the choice of GRR is driven by its appealing simplicity, MST and AIM are recent synthetic data generation methods. With each of these methods P2NIA, **the fairness estimation is reliably performed and respects differential privacy on user's data.** While each method guarantees the achievement of an unbiased fairness estimator (for an audit set Q of infinite size), as we will see in the next section, each method has a different impact on the convergence of fairness estimators.

4   Experimental Evaluation

While involving anonymized or synthetic data in audits has theoretical advantages compared to black-box audits, this section demonstrates the effectiveness of `P2NIA` in practice. To assess the performance of the proposed approach, we conducted experiments on two datasets commonly used in audit studies while varying the function g within `P2NIA`. The results demonstrate that our proposed scheme using anonymized or synthetic data achieves significantly higher accuracy than black-box audits. We first present our experimental setup before evaluating the advantage of using anonymized or synthetic data.

4.1   Experimental Setting

Datasets and models. Experiments are performed on two standard benchmark datasets from the fair and private machine learning literature: Adult [8] and Folktables [16]. They contain demographic information from the U.S. Census with respect to 32.561 and 378.816 individuals. In both cases, the task is to predict whether an individual's income exceeds $50,000$ per year (with the task *ACSIncome* on California, year 2018 for Folktables). The protected attribute is considered to be the gender.

To solve the tasks outlined above, we employ random forests as implemented in the Scikit-learn Python Library with its **RandomForestClassifier** object. Hyper-parameters were selected using a grid search to minimize the model's loss on each dataset. Further experiments using gradient boosting algorithm (**XGBClassifier**) are provided in [31].

Experimental parameters. All datasets were split into two parts: 80% of the dataset as the training set for the model m while the remaining 20% is used as a test set. Only the test set is used as input of the function g (step 4 in Figure 2). This process is repeated ten times with the same random seeds for the train/test split and the same model, with the results averaged across these ten iterations. All experiments are conducted on a computing cluster of homogeneous nodes powered by Intel Xeon E5-2660 v2 processors.

Reference. Since the objective is to evaluate the performance of `P2NIA` in the audit of fairness, we establish a reference by defining the true fairness value of the model used by the platform. Specifically, this reference corresponds to the fairness metrics computed on the test set of the datasets once they have been labeled by the model in question and is the target value that the auditor wants to estimate.

Baseline. We compare our approach with the black-box scenario in the following manner. The auditor is assumed to have knowledge of the set of possible values for each attribute ($\mathcal{X}$), but not to the overall distribution of the data ($\mathcal{D}$). The auditor draws each attribute uniformly from its possible values to generate queries. The same approach is also applied to the outputs Y when auditing fairness metrics such as Equalized Odds or Equality of Opportunity.

The performance of `P2NIA` is compared to the black-box audit on the two binary classification tasks from Adult and Folktables. To realize this, we instantiate `P2NIA` using locally anonymized data or synthetic data generated by the methods GRR [13], MST [40] or AIM [41] described in Section 3.2. More precisely, we used publicly available implementations of those methods [39].

4.2 Impact of Sample Size on Demographic Parity

We study the practical effectiveness of `P2NIA` by evaluating its performance depending on the number of samples to assess the demographic parity. The differential privacy parameter ϵ is arbitrarily set to 10.

(a) Adult   (b) Folktables

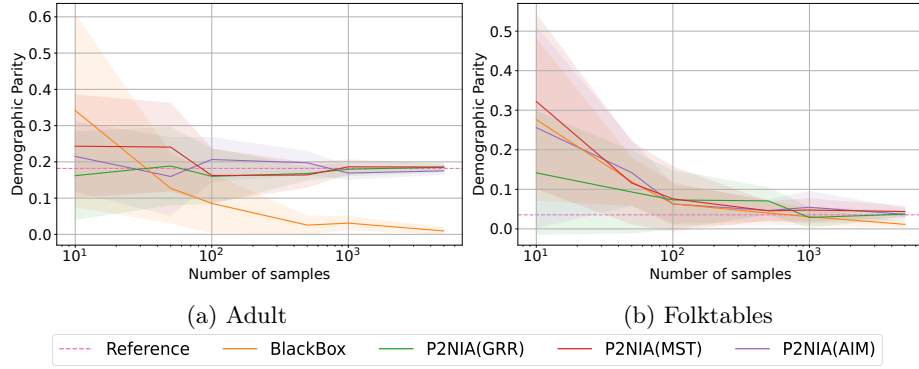- - - - Reference ──── BlackBox ──── P2NIA(GRR) ──── P2NIA(MST) ──── P2NIA(AIM)

Fig. 3: Demographic parity depending on the number of samples.

Analyzing the demographic parity depending on the number of samples (Figure 3) shows that the convergence rate of the estimator remains largely unaffected by the method used within `P2NIA`. This indicates that the selection of the method within the algorithm does not significantly affect the rate at which the estimator converges.

It is noteworthy that while increasing the number of samples enables a more accurate estimation, **`P2NIA` achieves high precision even with a modest sample size**. This contrasts with the black-box approach, in which convergence is slower and tends towards an incorrect value, resulting in unreliable fairness assessments. This outcome highlights a significant constraint of the black-box approach, namely its reliance on a dataset of significant size to achieve even an approximate solution, which is inherently biased.

4.3 Impact of Differential Privacy on Demographic Parity

We have explored the average error on demographic parity (on $5,000$ samples) for different levels of differential privacy in Figure 4.

(a) Adult   (b) Folktables

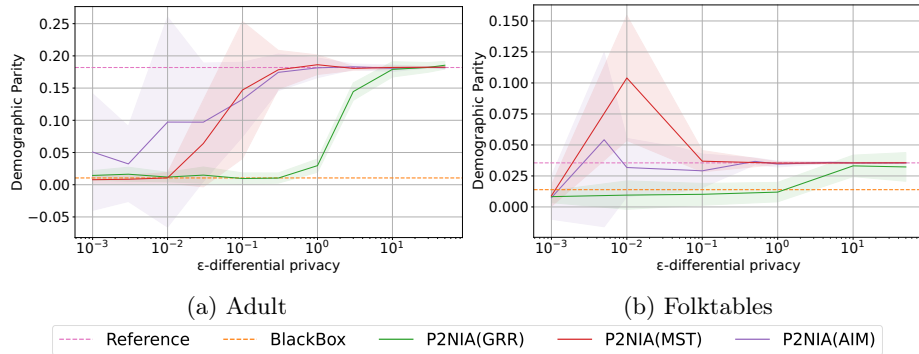--- Reference   --- BlackBox   — P2NIA(GRR)   — P2NIA(MST)   — P2NIA(AIM)

Fig. 4: Demographic parity depending on the ϵ-differential privacy in P2NIA.

Regardless of the value of ϵ, the black-box audit consistently produces the worst results for the Adult dataset, highlighting its inefficiency. Specifically, when the value of ϵ is large, the estimated demographic parity with P2NIA exhibits high accuracy, but with a low privacy level. Conversely, for small values of ϵ, the accuracy of the demographic parity estimator decreases, but the privacy protection increases.

On the Folktables dataset, P2NIA exhibits similar asymptotic behavior to that observed on the Adult dataset, reinforcing the generalizability of its performance. However, a key difference emerges in the stability of intermediate values of ϵ, for P2NIA (MST) and P2NIA (AIM). In these cases, the lack of stability can be attributed to the complexity introduced by the noisy marginals, affecting the consistency of the fairness estimates. In contrast, P2NIA (GRR) demonstrates a more controlled and predictable behavior despite converging more slowly.

These findings reveal the existing trade-off between differential privacy and estimator accuracy on demographic parity. The optimal balance between these two metrics is likely to depend on the characteristics of the dataset and the generation method employed with realistic ϵ Section 3.2. Notably, P2NIA **enables effective platform audit of demographic parity**, a capability that the black-box approach lacks due to its inherent bias.

4.4   Head-to-Head Comparison of P2NIA on Other Fairness Metrics

In this section, we show how P2NIA can be used to audit other fairness metrics, such as equalized odds and equality of opportunity. We set the differential privacy parameter ϵ to 10 but additional experiments for $\epsilon = 1$ can be found in [31].

For a constant value of $\epsilon = 10$, Figure 5 demonstrates that the earlier results extend beyond demographic parity to other fairness metrics, such as equalized odds and equality of opportunity. Specifically, our scheme with locally anonymized data (*i.e.*, P2NIA(GRR)) performs comparably to, or even better than, black-box auditing. Consequently, by releasing locally anonymized data, the platform makes it possible to reliably estimate fairness. Furthermore, our
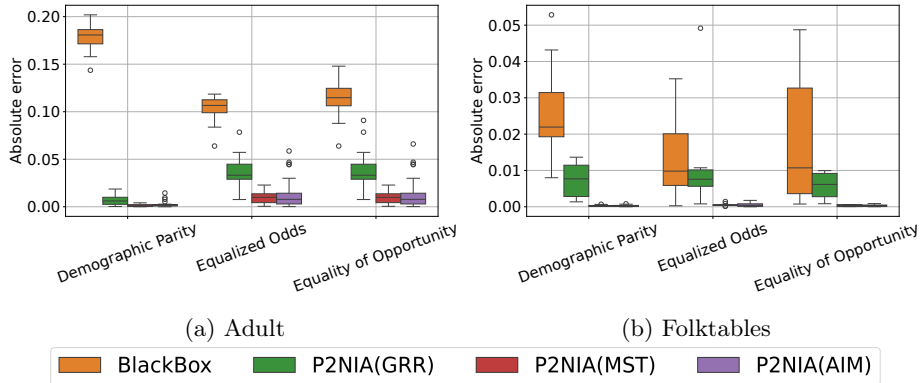
(a) Adult   (b) Folktables

| BlackBox | P2NIA(GRR) | P2NIA(MST) | P2NIA(AIM) |

Fig. 5: Absolute difference compared to the reference for the three standard fairness metrics with $\epsilon = 10$.

scheme instantiated with synthetic data (*i.e.*, P2NIA(MST) and P2NIA(AIM)) also results in a significant error reduction. These findings highlight the usefulness of our approach as a fairness auditing tool, demonstrating **the ability of P2NIA to evaluate not only Demographic Parity but also Equalized Odds and Equality of Opportunity effectively**.

5   Related work

Fairness auditing, particularly in black-box settings, has been extensively explored [4, 14, 19, 46, 48]. These approaches rely on analyzing the outputs of the model based on a set of queries designed to assess fairness. For black-box auditing, the fundamental lever the auditor relies on is their choice of queries. This could involve i.i.d. sampling [25], stratified sampling [50] or more complex setups [44]. However, all these methods rely on the central assumption that the auditor knows the model distribution in advance, which is unrealistic due to the knowledge asymmetry between non collaborative platforms and auditors.

Thus, there has some been recent work [10,11], arguing for more transparent ways to evaluate the assessment of algorithms with respect to ethical standards as audits are imprecise due to the lack of comprehension of these algorithms. Alternative approaches for fairness auditing include, for instance, the whitebox setting in which the model is provided to the auditor [5, 11, 11] or fairness certification [18,35,47]. The most closely related work to ours is the recent study by Yuan and Wang [58] but nonetheless their auditing setting differs significantly from ours in that the auditor is granted direct access to real labeled data from the platform. The responsibility for privacy preservation is then delegated to the auditor, who received data from real users. Although the original data must be deleted after use, privacy has already been compromised due to the data disclosure. In contrast, our approach is built to be privacy-preserving by design, without exposing raw labeled data to the auditor.

As fairness auditing often requires querying the model on user data, privacy concerns have also emerged as a fundamental issue. Dwork and co-authors [21] have already identified and prove the link between individual fairness and local differential privacy. It has also been studied in the context of fair learning [38] (*e.g.*, before model being in production) but not for audit purpose. However, [36, 45] introduced differential privacy techniques to mitigate the risk of information leakage during fairness audits. Other methods, such as secure multi-party computation [52] rely on complex cryptographic techniques to ensure the privacy of queries. A key limitation of these previous works is that they focus on protecting the auditor's queries rather than the data of the entity being audited. Our approach, `P2NIA`, aims to bridge this gap by allowing effective fairness auditing without compromising user privacy.

6   Conclusion

To summarize, in this paper, we have first underlined an important problem for black-box audits in real scenarios: population bias due to the lack of a perfect prior for the auditor on the data distribution of the platform. This bias arises mainly due to the lack of collaboration between the platform and the auditor. We then proposed a novel audit scheme, named `P2NIA`, in which both parties collaborate out of mutual benefit, namely accurate estimation for the auditor and data privacy preservation plus ease of operation for the platform as the scheme is non-iterative. We empirically show on standard datasets that our scheme operates as intended for three group fairness metrics, with accurate assessment and controllable privacy guarantees.

Future work includes the study of other forms of collaborations driven by different benefits at each party. For instance, it might be beneficial for the platform to operate in a "grey-box" setting, in which it will reveal some information about the internal workings of its model, resulting in a more lightweight querying procedure by an auditor (*e.g.*, with respect to the number of queries required).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alaa, A., Van Breugel, B., Saveliev, E.S., Van Der Schaar, M.: How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In: International Conference on Machine Learning. PMLR (2022)

2. American city and council: Avoiding bumps in the road when designing ai-powered traffic management systems, `https://www.americancityandcounty.com/2024/01/26/avoiding-bumps-in-the-road-when-designing-ai-powered-traffic-management-systems/`
3. Apple Inc.: Differential privacy overview (2017), `https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf`
4. Bandy, J.: Problematic machine behavior: A systematic literature review of algorithm audits. Proceedings of the acm on human-computer interaction (2021)
5. Barocas, S., Hardt, M., Narayanan, A.: Fairness and machine learning: Limitations and opportunities. MIT press (2023)
6. Bassily, R., Smith, A.: Local, private, efficient protocols for succinct histograms. In: Proceedings of the forty-seventh annual ACM symposium on Theory of computing. pp. 127–135 (2015)
7. BBC: Ai hiring tools may be filtering out the best job applicants, `https://www.bbc.com/worklife/article/20240214-ai-recruiting-hiring-software-bias-discrimination`
8. Becker, B., Kohavi, R.: Adult. UCI Machine Learning Repository (1996), DOI: https://doi.org/10.24432/C5XW20
9. Besse, P., del Barrio, E., Gordaliza, P., Loubes, J.M., Risser, L.: A survey of bias in machine learning through the prism of statistical parity. The American Statistician **76**(2), 188–198 (2022)
10. Birhane, A., Steed, R., Ojewale, V., Vecchione, B., Raji, I.D.: Ai auditing: The broken bus on the road to ai accountability. In: 2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). pp. 612–643. IEEE (2024)
11. Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T.L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., et al.: Black-box access is insufficient for rigorous ai audits. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency. pp. 2254–2272 (2024)
12. CBS News: Argentina plans to use ai to "predict future crimes and help prevent them", `https://www.cbsnews.com/news/argentina-plans-to-use-ai-to-predict-future-crimes-and-help-prevent-them/`
13. Chaudhuri, A., Mukerjee, R.: Randomized response: Theory and techniques. Routledge (2020)
14. Chen, L., Mislove, A., Wilson, C.: An empirical analysis of algorithmic pricing on amazon marketplace. In: Proceedings of the 25th international conference on World Wide Web. pp. 1339–1349 (2016)
15. Cochran, W.G.: Sampling Techniques. John Wiley & Sons (1977)
16. Ding, F., Hardt, M., Miller, J., Schmidt, L.: Retiring adult: New datasets for fair machine learning. Advances in Neural Information Processing Systems **34** (2021)
17. Domingo-Ferrer, J., Soria-Comas, J.: Multi-dimensional randomized response. IEEE Transactions on Knowledge and Data Engineering **34**(10), 4933–4946 (2020)
18. Duddu, V., Das, A., Khayata, N., Yalame, H., Schneider, T., Asokan, N.: Attesting distributional properties of training data for machine learning. In: European Symposium on Research in Computer Security. pp. 3–23. Springer (2024)
19. Dunna, A., Keith, K.A., Zuckerman, E., Vallina-Rodriguez, N., O'Connor, B., Nithyanand, R.: Paying attention to the algorithm behind the curtain: Bringing transparency to youtube's demonetization algorithms. Proceedings of the ACM on human-computer interaction **6**(CSCW2), 1–31 (2022)
20. Dwork, C.: A firm foundation for private data analysis. Communications of the ACM **54**(1), 86–95 (2011)

21. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through aware-ness. In: Proceedings of the 3rd innovations in theoretical computer science con-ference. pp. 214–226 (2012)
22. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science (2014)
23. Erlingsson, Ú., Pihur, V., Korolova, A.: Rappor: Randomized aggregatable privacy-preserving ordinal response. In: Proceedings of the 2014 ACM SIGSAC conference on computer and communications security. pp. 1054–1067 (2014)
24. European Union: Artificial intelligence act. `http://data.europa.eu/eli/reg/2024/1689/oj` (2024)
25. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. pp. 259–268 (2015)
26. Figueira, A., Vaz, B.: Survey on synthetic data generation, evaluation methods and gans. Mathematics **10**(15), 2733 (2022)
27. Financial Times: Volatility: how 'algos' changed the rhythm of the market, `https://www.ft.com/content/fdc1c064-1142-11e9-a581-4ff78404524e`
28. Forbes: Sports ai can be a game-changing partner for coaches, `https://www.forbes.com/sites/geristengel/2024/05/15/sports-ai-can-be-a-game-changing-partner-for-coaches/`
29. French Government: Penal code - article 434-10 (2020), `https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000042026716/`, ac-cessed: August 20, 2024
30. Fukuchi, K., Hara, S., Maehara, T.: Faking fairness via stealthily biased sampling. In: Proceedings of the AAAI Conference on Artificial Intelligence (2020)
31. Garcia Bourrée, J., Lautraite, H., Gambs, S., Tredan, G., Le Merrer, E., Rot-tembourg, B.: P2nia: Privacy-preserving non-iterative auditing. arXiv preprint arXiv:2504.00874 (2025)
32. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
33. Hargittai, E.: Is bigger always better? potential biases of big data derived from social network sites. The ANNALS of the American Academy of Political and Social Science **659**(1), 63–76 (2015)
34. Harvard Business review: Algorithms can save networking from being business card roulette, `https://hbr.org/2014/03/algorithms-can-save-networking-from-being-business-card-roulette`
35. Kang, M., Li, L., Weber, M., Liu, Y., Zhang, C., Li, B.: Certifying some distribu-tional fairness with subpopulation decomposition. Advances in Neural Information Processing Systems **35**, 31045–31058 (2022)
36. Kilbertus, N., Gascón, A., Kusner, M., Veale, M., Gummadi, K., Weller, A.: Blind justice: Fairness with encrypted sensitive attributes. In: International Conference on Machine Learning. pp. 2630–2639. PMLR (2018)
37. Le Merrer, E., Pons, R., Trédan, G.: Algorithmic audits of algorithms, and the law. AI and Ethics pp. 1–11 (2023)
38. Makhlouf, K., Stefanović, T., Arcolezi, H.H., Palamidessi, C.: A systematic and formal study of the impact of local differential privacy on fairness: Preliminary results. In: IEEE 37th Computer Security Foundations Symposium. pp. 1–16 (2024)
39. McKenna, R., Miklau, G., Sheldon, D.: Private-pgm (2021), `https://github.com/journalprivacyconfidentiality/private-pgm-jpc-778/tree/v2021-10-04-jpc`

40. McKenna, R., Miklau, G., Sheldon, D.: Winning the nist contest: A scalable and general approach to differentially private synthetic data. arXiv preprint arXiv:2108.04978 (2021)
41. McKenna, R., Mullins, B., Sheldon, D., Miklau, G.: Aim: An adaptive and iterative mechanism for differentially private synthetic data. arXiv preprint arXiv:2201.12677 (2022)
42. McKenna, R., Sheldon, D., Miklau, G.: Graphical-model based estimation and inference for differential privacy. In: International Conference on Machine Learning. pp. 4435–4444. PMLR (2019)
43. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM computing surveys **54**(6), 1–35 (2021)
44. Panigutti, C., Perotti, A., Panisson, A., Bajardi, P., Pedreschi, D.: Fairlens: Auditing black-box clinical decision support systems. Information Processing & Management **58**(5), 102657 (2021)
45. Pentyala, S., Melanson, D., De Cock, M., Farnadi, G.: Privfair: a library for privacy-preserving fairness auditing. arXiv preprint arXiv:2202.04058 (2022)
46. Sandvig, C., Hamilton, K., Karahalios, K., Langbort, C.: Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and discrimination: converting critical concerns into productive inquiry (2014)
47. Shamsabadi, A.S., Wyllie, S.C., Franzese, N., Dullerud, N., Gambs, S., Papernot, N., Wang, X., Weller, A.: Confidential-profitt: confidential proof of fair training of trees. In: ICLR (2022)
48. Silva, M., Santos de Oliveira, L., Andreou, A., Vaz de Melo, P.O., Goga, O., Benevenuto, F.: Facebook ads monitor: An independent auditing system for political ads on facebook. In: Proceedings of the web conference 2020. pp. 224–234 (2020)
49. Tao, Y., McKenna, R., Hay, M., Machanavajjhala, A., Miklau, G.: Benchmarking differentially private synthetic data generation algorithms. arXiv preprint arXiv:2112.09238 (2021)
50. Taskesen, B., Blanchet, J., Kuhn, D., Nguyen, V.A.: A statistical test for probabilistic fairness. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. pp. 648–665 (2021)
51. The Guardian: Ai better than biopsy at assessing some cancers, study finds, `https://www.bbc.com/worklife/article/20240214-ai-recruiting-hiring-softwar-bias-discrimination`
52. Toreini, E., Mehrnezhad, M., Van Moorsel, A.: Verifiable fairness: privacy–preserving computation of fairness for machine learning systems. In: European symposium on research in computer security. pp. 569–584. Springer (2023)
53. de Vos, M., Dhasade, A., Garcia Bourrée, J., Kermarrec, A.M., Le Merrer, E., Rottembourg, B., Tredan, G.: Fairness auditing with multi-agent collaboration. In: ECAI 2024, pp. 1116–1123. IOS Press (2024)
54. Wang, S., Peng, Y., Li, J., Wen, Z., Li, Z., Yu, S., Wang, D., Yang, W.: Privacy amplification via shuffling: Unified, simplified, and tightened. arXiv preprint arXiv:2304.05007 (2023)
55. Wang, T., Blocki, J., Li, N., Jha, S.: Locally differentially private protocols for frequency estimation. In: 26th USENIX Security Symposium. pp. 729–745 (2017)
56. Warner, S.L.: Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American statistical association **60**(309), 63–69 (1965)
57. Ye, M., Barg, A.: Optimal schemes for discrete distribution estimation under locally differential privacy. IEEE Transactions on Information Theory (2018)
58. Yuan, C.C.R., Wang, B.Y.: Quantitative auditing of ai fairness with differentially private synthetic data. arXiv preprint arXiv:2504.21634 (2025)