Can Querying for Bias Leak Protected Attributes? Achieving Privacy With Smooth Sensitivity

Faisal Hamman
University of Maryland
fhamman@umd.edu

Jiahao Chen
Responsible AI LLC
jiahao@responsibleai.tech

Sanghamitra Dutta
University of Maryland
sanghamd@umd.edu

ABSTRACT

Existing regulations often prohibit model developers from accessing protected attributes (gender, race, etc.) during training. This leads to scenarios where fairness assessments might need to be done on populations without knowing their memberships in protected groups. In such scenarios, institutions often adopt a separation between the model developers (who train their models with no access to the protected attributes) and a compliance team (who may have access to the entire dataset solely for auditing purposes). However, the model developers might be allowed to test their models for disparity by querying the compliance team for group fairness metrics. In this paper, we first demonstrate that simply querying for fairness metrics, such as, statistical parity and equalized odds can leak the protected attributes of individuals to the model developers. We demonstrate that there always exist strategies by which the model developers can identify the protected attribute of a targeted individual in the test dataset from just a single query. Furthermore, we show that one can reconstruct the protected attributes of *all* the individuals from $O(N_k \log(n/N_k))$ queries when $N_k \ll n$ using techniques from compressed sensing (n is the size of the test dataset and N_k is the size of smallest group therein). Our results pose an interesting debate in algorithmic fairness: Should querying for fairness metrics be viewed as a neutral-valued solution to ensure compliance with regulations? Or, does it constitute a violation of regulations and privacy if the number of queries answered is enough for the model developers to identify the protected attributes of specific individuals? To address this supposed violation of regulations and privacy, we also propose Attribute-Conceal, a novel technique that achieves differential privacy by calibrating noise to the smooth sensitivity of our bias query function, outperforming naive techniques such as the Laplace mechanism. We also include experimental results on the Adult dataset and synthetic dataset (broad range of parameters).

CCS CONCEPTS

• **Social and professional topics** → **Governmental regulations**; *User characteristics*; • **Computing methodologies** → **Philosophical/theoretical foundations of artificial intelligence**; • **General and reference** → **Evaluation**; • **Security and privacy** → **Privacy-preserving protocols**.

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06.
https://doi.org/10.1145/3593013.3594086

KEYWORDS

algorithmic fairness, compliance, compressed sensing, differential privacy, machine learning.

ACM Reference Format:
Faisal Hamman, Jiahao Chen, and Sanghamitra Dutta. 2023. Can Querying for Bias Leak Protected Attributes? Achieving Privacy With Smooth Sensitivity. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3593013.3594086

1 INTRODUCTION

The ethical goal of algorithmic fairness [5, 60] is closely tied to the legal frameworks of both anti-discrimination and privacy. For instance, Title VII of the Civil Rights Act of 1964 [5] introduces two different notions of unfairness, namely, disparate impact [58], and disparate treatment [73], which are often at odds with each other [5]. It is widely believed that a machine learning model can avoid violating disparate treatment (and privacy concerns) if the model does not *explicitly* use the protected attributes [5]. However, it has been demonstrated that even if no protected attributes are explicitly used during training, a model might still be held liable for disparate impact, due to proxies of the protected attributes among other attributes in the dataset [17]. In fact, existing literature on algorithmic fairness demonstrates that *leveraging protected attributes during training can essentially prevent disparate impact*, e.g., by minimizing a fairness metric as a regularizer with the loss function during training [36, 43]. Thus, mitigating disparate impact often seems to be at odds with disparate treatment [49] (and privacy), depending on whether the protected attribute is being explicitly used or not.

One potential resolution (still debated [45]) is to use the protected attributes only during training to mitigate disparate treatment but not after deployment. Nonetheless, *the use of protected attributes during model training remains to be a source of active debate* and contention for various applications [49, 51]. On one hand, using protected attributes during training enables one to actively audit and account for biases, as well as understand how specific groups of people are affected. On the other hand, these protected attributes can also be used maliciously, e.g., to exacerbate discrimination [57]. An interesting example arises where the protected attributes can even be used to "mask" discrimination [22, 26, 27, 44], e.g., an expensive housing Ad is shown to only high-income White individuals and low-income Black individuals but not to low-income White individuals and high-income Black individuals (assuming an equal proportion of all these four sub-groups) [27]. The decision is clearly discriminatory against high-income Black individuals for whom

the Ad is relevant and yet they do not get to see it. This discrimination is masked since the decision might still satisfy statistical independence between the two races.

In several applications, e.g., in finance, anti-discrimination and privacy regulations adopt a stance that completely prohibits the use of protected attributes during training. In the finance domain, institutions cannot ask about an individual's race for credit decisioning, while at the same time having to prove that their decisions are non-discriminatory [15]. The Apple Card credit card was recently accused of discriminatory credit decisioning since women received lower credit limits than equally qualified men, despite not using the gender explicitly during training [67].

Fairness assessment of these models is extremely challenging when the protected attributes are unavailable[1]. To address this, institutions often adopt a separation between the model developers and the compliance team [15]. The compliance team is responsible for ensuring methods do not violate anti-discrimination and privacy laws. As a result, the compliance team has access to the entire dataset, including the attributes protected by law (i.e., race, gender, etc.) [14, 15]. Only a subset of the data fields is visible to the model developers who train these models. The compliance team determines which attributes the model developers are allowed to see and use to train their models. Clearly, the model developers would not have access to the protected attributes. For fairness assessment, the model developers may however query the compliance team for certain group fairness metrics, e.g., statistical parity, equalized odds, etc. The model developers can then choose which model to deploy or discard based on the query responses.

In this paper, we first demonstrate that simply querying for fairness metrics (bias) can also leak the protected attributes of targeted individuals. Furthermore, we demonstrate that there exist strategies by which the model developers can always identify the protected attributes of *all* the individuals in the test dataset. We collectively refer to these strategies as *Attribute-Reveal*. Our finding poses an interesting debate in the policy aspects of fairness and privacy: *Should querying for fairness metrics be viewed as a neutral-valued solution to ensure compliance with regulations? Or, does it constitute a violation of regulations and privacy, particularly if the number of queries answered is enough for the model developers to identify the protected attributes of specific individuals?* To address this supposed violation of regulations and privacy, we also propose *Attribute-Conceal*, a novel differentially-private technique to answer queries without leaking the protected attributes.

To summarize, our main contributions are as follows:

1. Demonstrate that querying for bias can leak protected attributes: We first demonstrate that querying for fairness metrics, e.g., statistical parity, or equalized odds, can indeed leak the protected attributes of individuals. In Theorem 1, we provide the general criterion for reconstructing the protected attributes of all the individuals in the test dataset by querying for the statistical parity of several models (reduces to a linear system of equations).

2. Leverage compressed sensing to reconstruct protected attributes with fewer queries: Building on our initial result (Theorem 1), we then demonstrate how protected attributes of individuals can be leaked using a much smaller number of statistical parity queries, provided the size of one group is much smaller than the other (see Theorem 2). Our findings also extend to the absolute value of statistical parity (see Section 3.3), as well as, other fairness metrics, e.g., equalized odds or their absolute values. We collectively refer to these proposed reconstruction strategies as *Attribute-Reveal*.

3. Propose *Attribute-Conceal*, a novel technique that achieves differential privacy by calibrating noise to the smooth sensitivity of our bias query function: To avoid leaking protected attributes, we propose *Attribute-Conceal*, a technique that answers fairness queries in an $\epsilon-$differentially-private manner (see Section 4.2). Since calibrating noise to global sensitivity (e.g., using the Laplace mechanism) can often hurt the utility of the answered query (because the noise becomes too high), we employ the **smooth sensitivity** framework [56], which adds dataset-specific additive noise to achieve differential privacy (see Theorem 7 in Section 4.2).

4. Experiments: To complement our theoretical results, we also provide experimental results on the Adult dataset [25] as well as perform simulations on synthetic data for a broad range of parameters. We demonstrate how Attribute-Reveal reconstructs the protected attributes, and how Attribute-Conceal prevents the reconstruction. We also compare Attribute-Conceal to other naive differential privacy techniques such as the Laplace mechanism.

Related Works: Algorithmic fairness is an active area of research [2, 6, 15, 18, 19, 22, 26–28, 30, 33, 35, 37, 43, 44, 46–48, 55, 62, 63, 65, 68, 71] that is receiving increasing attention. Many of these techniques assume that the protected attributes are available during training, which is not always allowed in practice. In several applications, such as credit or loan decisioning [14, 15], the use of protected attributes during training is restricted by law. Our work is closely related to a body of work that addresses *fairness without access to protected attributes* [15, 18, 35, 61, 65], often using proxies to estimate the protected attributes. Our work lies in an area that relies on trusted third parties who have access to protected data necessary for improving fairness. For instance, [38, 45, 64] assume that the model has access to the protected attributes in an encrypted form via secure multi-party computation. Later, [40] noted that secure multi-party computation technique does not protect protected attributes from leaking, employing differential privacy [31] to learn fair models. [23] uses a fully homomorphic encryption scheme, allowing model developers to train models and test them for bias without revealing the protected attributes. More recent works [1, 3, 21, 69] provide schemes that allow the release of protected attributes privately for learning non-discriminatory predictors. [41] proposes a differentially private mechanism to measure differences in performance across groups while protecting the privacy of group membership in a federated setting.

Our work instead addresses a novel problem statement: Can querying for fairness metrics leak protected attributes, and if so, how can we leverage smooth sensitivity to prevent this leakage. Our problem setup also differs from existing works in attribute inference attacks [20, 34, 39, 50] where the focus is on *learning* protected information in training data from model outputs using supervised learning (we do not use group membership labels from past data). An interesting related work [70] studies query-based

[1]While [35] suggests that proxies may be used to approximate missing protected attributes for bias assessment, others [15, 42] argue that proxies may be problematic since they often underestimate bias.

auditing algorithms to estimate the statistical parity of ML models. Another related work is [59], which focuses on the issue of fair washing, where manipulation techniques are utilized to mask unfairness when presenting the model's explanations to an auditor.

2 PROBLEM SETUP

2.1 Preliminaries

Let $\mathcal{S} = (X, Y, A)$ represent a test dataset consisting of n samples, where $A = (a_1, a_2, ..., a_j, ..., a_n)$ denotes the protected/sensitive attributes (binary), $X = (x_1, x_2, ..., x_j, ..., x_n)$ denotes the model inputs with each $x_j \in \mathbb{R}^d$, and $Y = (y_1, y_2, ..., y_j, ..., y_n)$ being the corresponding true labels used for supervised learning. We let $a_j = 1$ denote the advantaged group and $a_j = 0$ denote the disadvantaged group. We consider two types of classifiers: binary and logistic classifiers. The binary classifiers are represented by the function $h(\cdot) : \mathbb{R}^d \to \{0, 1\}$. For a logistic classifier $h(\cdot) : \mathbb{R}^d \to [0, 1]$, the output corresponds to the probability of input x being accepted. When we have several classifiers, we let $h_i(x_j)$ represent the i-th classifier's output to the input x_j (input feature vector of the j-th individual in the dataset) for $i \in [m]$, where $[m] = \{1, 2, ..., m\}$ for a positive integer m. Let $h_i(X)$ represent the i-th classifier's outputs to all individuals in the dataset, i.e, $h_i(X) = \big(h_i(x_1), ..., h_i(x_j), ..., h_i(x_n)\big)$. We let N_0 and N_1 be the size of the disadvantaged and advantaged group, i.e., $N_0 = \sum_{j\in[n]} \mathbb{1}\{a_j = 0\}$ and $N_1 = \sum_{j\in[n]} \mathbb{1}\{a_j = 1\}$. Note that $N_1 + N_0 = n$, the size of the test dataset.

2.1.1 *Review of Relevant Group Fairness Metrics.*

Definition 2.1 (Statistical Parity Gap (SP_i)). *Statistical parity gap (SP_i) is defined as the difference in expected outcome between the advantaged and disadvantaged groups, i.e.,*

$$SP_i = \frac{1}{N_1} \sum_{\{j|a_j=1\}} h_i(x_j) - \frac{1}{N_0} \sum_{\{j|a_j=0\}} h_i(x_j). \tag{1}$$

Definition 2.2 (Equal Opportunity Gap (EO_i)). *Equal opportunity gap (EO_i) is defined as:*

$$EO_i = \frac{1}{N_{11}} \sum_{\{j|y_j=1,a_j=1\}} h_i(x_j) - \frac{1}{N_{10}} \sum_{\{j|y_j=1,a_j=0\}} h_i(x_j), \tag{2}$$

where $N_{11} = \sum_{j\in[n]} \mathbb{1}\{y_j{=}1, a_j{=}1\}$, $N_{10} = \sum_{j\in[n]} \mathbb{1}\{y_j{=}1, a_j{=}0\}$.

We also denote the absolute values of statistical parity and equal opportunity gap as $|SP_i|$ and $|EO_i|$ respectively.

Remark 1. *We note that although we only define statistical parity and equal opportunity, our techniques can be extended to several other group fairness measures, such as equalized odds, predictive parity, etc. as well as their absolute values. We further discuss this in Remark 5.*

2.2 Problem Statement

Institutions often adopt a separation between the model developers and the compliance team to ensure anti-discrimination and privacy laws are met. The model developers do not have access to protected attributes and therefore cannot use them for training. The compliance team, however, has access to the entire dataset, but only for auditing purposes. In our setting, the model developers train m different classifiers $h_i(\cdot)$ with $i \in [m]$ on the training dataset.

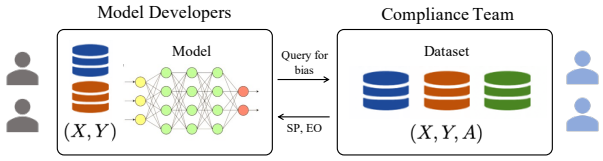

Figure 1: Illustrates an institutional separation between model developers and compliance team for ensuring fair and privacy-compliant machine learning models. Model developers train classifiers on the training dataset, but without access to protected attributes. Compliance team has access to the entire dataset for auditing purposes and is queried by model developers for fairness metrics.

For the fairness assessment of these models before deployment, the model developers are allowed to test for algorithmic bias by querying the compliance team for certain fairness metrics on the test dataset $\mathcal{S} = (X, Y, A)$. Note that the classifier $h_i(\cdot)$ is only a function of the input x_j and not the protected attributes a_j. The fairness metrics that the model developers can query for includes the statistical parity gap, and equal opportunity gap as well as their absolute values (see system model in Figure 1). The main question that we ask in this work is: Is this technique of querying for fairness metrics effective in keeping the protected attributes hidden from the model developers? Or in general, *does querying for fairness leak protected attributes?* And if so, how can one answer queries without leaking protected attributes?

Remark 2. *The approach outlined in this paper can be extended to a scenario with an institution (that trains a model without access to protected attributes) and an external fairness auditing team, e.g., in [70]. The auditing team has access to the entire data for the purpose of evaluating the bias of the models and is responsible for informing the model developers on whether their deployed model passes fairness tests based on some fairness metric. In this setting, our concern is whether auditing for fairness compromise and leak the protected attributes to the institution.*

3 ATTRIBUTE-REVEAL: QUERYING FOR BIAS LEAKS PROTECTED ATTRIBUTES

3.1 Demonstrating Leakage From Querying

Here, we show that simply querying for fairness metrics such as statistical parity can reveal the protected attribute of any targeted individual to the model developers. In fact, there exist strategies (that we collectively refer to as Attribute-Reveal) that can reveal the protected attributes of all the individuals in the test dataset. We begin with a simple toy example.

Example 1 (Single Query). *The model developers train only one binary classifier $h_1(\cdot)$ and query the compliance team for statistical parity gap. Suppose, the model developers want to find the protected attribute of the first individual. They can choose a classifier that accepts only the first individual, i.e., $h_1(x_1) = 1$ and $h_1(x_j) = 0$ for $j = 2, 3, \ldots, n$. Observe that the statistical parity gap of this model will reveal the protected attribute of the first individual as follows: $SP_1 = \frac{1}{N_1}$ if $a_1 = 1$, and $SP_1 = -\frac{1}{N_0}$ if $a_1 = 0$. Thus, a positive*

statistical parity gap SP_1 would give away that the individual belongs to the advantaged group, whereas a negative gap indicates that the individual belongs to the disadvantaged group. The query also reveals the sizes of these groups N_1 and N_0.

We note that such a model might seem contrived; it might also have low accuracy. Thus, in Example 2, we demonstrate a more realistic scenario that can occur more commonly in practice: *two models of comparable accuracy* can be used to reveal the protected attribute of a targeted individual.

Example 2 (Double Query With Realistic Models). *Consider two models $h_1(\cdot)$ and $h_2(\cdot)$. The model $h_1(\cdot)$ is trained by model developers (to maximize accuracy) and accepts several individuals in the dataset. The model developers also use a second classifier, $h_2(\cdot)$, that provides the same prediction as $h_1(\cdot)$ except for one targeted individual, i.e., $h_2(x_1) = 1 - h_1(x_1)$ and $h_2(x_j) = h_1(x_j)$ for $j = 2, 3, \ldots, n$. Notice that $h_1(\cdot)$ and $h_2(\cdot)$ differ* only *in the first prediction. Their accuracies are almost similar. If one queries for statistical parity of these two models, they can identify the protected attribute of the first individual as follows: $SP_2 - SP_1 = \frac{1}{N_1}$ if $a_1 = 1$ and $SP_2 - SP_1 = -\frac{1}{N_0}$ if $a_1 = 0$.*

A similar approach can be adopted to reveal the protected attributes of all the individuals in the test dataset. Our next result provides the general criterion for reconstructing the protected attribute of all the individuals in the test dataset using the statistical parity queries (see Appendix A for proof).

Theorem 1 (Reveal From Linear System of Equations). *Let $\overline{SP} = \begin{bmatrix} SP_1 & SP_2 & \ldots & SP_m \end{bmatrix}^T$ be a vector of statistical parity gap queries for m models, $h_i(x_j)$ denote the i-th model's prediction for the j-th individual, and $\mathbf{H}$ be an $m \times n$ matrix where each row represents the binary or logistic predictions of the i-th model. If rank$(\boldsymbol{H}) = n$, then the protected attributes $A = (a_1, a_2, ..., a_j, ..., a_n)$ of the entire dataset can be identified by solving a linear system of equations:*

$$\begin{pmatrix} SP_1 \\ SP_2 \\ . \\ . \\ . \\ SP_m \end{pmatrix} = \begin{pmatrix} h_1(x_1) & h_1(x_2) & \ldots & \ldots & h_1(x_n) \\ h_2(x_1) & h_2(x_2) & \ldots & \ldots & h_2(x_n) \\ . & . & & & . \\ . & & & & . \\ . & & & & . \\ h_m(x_1) & h_m(x_2) & \ldots & \ldots & h_m(x_n) \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ . \\ . \\ . \\ v_n \end{pmatrix}. \quad (3)$$

This can also be expressed as,

$$\overline{SP} = \mathbf{H}_{m \times n}\mathbf{v}, \quad (4)$$

where $\mathbf{v}$ is the unknown vector with elements taking values:

$$v_j = \begin{cases} \frac{1}{N_1}, & \text{if } a_j = 1 \\ -\frac{1}{N_0}, & \text{if } a_j = 0. \end{cases}$$

Remark 3. *Strictly speaking, one needs $m = n - 1$ queries since the last individual can be identified using group sizes N_1 and N_0. However, one may encounter numerical errors when solving for N_1 from $1/N_1$ or, N_0 from $1/N_0$. This can happen when $N_1, N_0 >> 1$ and $N_0 \approx N_1$.*

Algorithm 1 provides a more realistic strategy by which model developers can choose practical models with comparable accuracy to reveal the protected attributes of all the individuals. Essentially, one base model $h_0(\cdot)$ could be trained, and several similar models $h_i(x_j)$ could be chosen so the prediction of $h_i(x_j)$ is flipped only when $i = j$. The accuracy of these models would remain comparable to the base model $h_0(\cdot)$ since they differ in only one prediction.

Algorithm 1 Attribute-Reveal ($m = n$)

Train base model $h_0(\cdot)$ (reasonable accuracy)
Choose m=n models $h_1(\cdot), \ldots, h_n(\cdot)$ as:
for $i = 1, 2, \ldots, m$ **do**
  for $j = 1, 2, \ldots, n$ **do**

$$h_i(x_j) = \begin{cases} h_0(x_j), & \text{if } i \neq j \\ 1 - h_0(x_j), & \text{if } i = j. \end{cases}$$

Query for Statistical Parity Gap for each model and store in $\overline{SP}$
Create matrix: $\mathbf{H}_{m \times n} = [h_1(X)^T, h_2(X)^T, \ldots, h_m(X)^T]^T$
Solve linear system of equations:
  $\overline{SP} = \mathbf{Hv}$ ▷ This algorithm is based on Theorem 1
for $j = 1, 2, \ldots, n$ **do**

$$\text{Detect } \hat{a}_j = \begin{cases} 1, & \text{if } v_j > 0 \\ 0, & \text{otherwise.} \end{cases}$$

We note that if the size of the test dataset is large, it may not be desirable to have as many models as the size of the dataset. This motivates our next question: *Is it possible to obtain the protected attributes of individuals in the dataset with fewer models and queries?*

3.2 Leaking Protected Attributes with Fewer Queries using Compressed Sensing (CS)

In this section, we demonstrate that compressed sensing (CS) techniques can be used to obtain the protected attributes of individuals using a significantly smaller number of queries (m). First, we provide a brief background on compressed sensing in Section 3.2.1. Readers already familiar with this topic may skip this subsection.

3.2.1 Brief Background on Compressed Sensing. The goal [24, 53] is to recover a vector $x \in \mathbb{R}^n$ from a set of linear measurements $\eta = \Phi x$, where η is an $m \times 1$ measurement vector, $\Phi \in \mathbb{R}^{m \times n}$ is the sensing matrix. CS relies on the sparsity of x. *A vector x is k-sparse if it has only k non-zero entries (typically $k \ll n$).* The measurements m are typically much smaller than n, making this an under-determined system of equations, having many solutions for x. CS focuses on finding the sparsest solution for x. This can be expressed as an optimization problem: $\min_x ||x||_0$ s.t. $\eta = \Phi x$. Here, $||x||_0$ is the number of nonzero entries[2] of x. By minimizing an l_1 norm instead, this problem can be relaxed into a convex optimization problem which can be solved using linear programming or other CS algorithms, e.g., Orthogonal Matching Pursuit [54].

$$\min_x ||x||_1 \text{ s.t. } \eta = \Phi x. \quad (5)$$

The l_1-norm $||x||_1$ is the absolute sum of all entries of x. We refer the reader to an excellent survey [13] for more information on CS.

For accurate recovery of k-sparse vector x from measurements η, the sensing matrix Φ has to satisfy a necessary and sufficient condition called Restricted Isometry Property (RIP) [9].

[2]Note that $||x||_0$ is not a norm. $||x||_p$ denotes a standard l_p-norm for $p \geq 1$, i.e., $||x||_p = \left(\sum_{i=0}^{n} |x_i|^p\right)^{\frac{1}{p}}$ for all $x \in \mathbb{R}^n$

Definition 3.1 (k-Restricted Isometry Property). *A matrix $\Phi \in \mathbb{R}^{m \times n}$ satisfies the Restricted Isometry Property of order k if for all k-sparse vector $x \in \mathbb{R}^n$, and for some constant $\delta_k \in (0,1)$, we have*

$$(1-\delta_k)||x||_2^2 \leq ||\Phi x||_2^2 \leq (1+\delta_k)||x||_2^2. \quad (6)$$

For a matrix Φ that satisfies the RIP condition of order $2k$ with $\delta_{2k} < \sqrt{2}-1$ (see [10]), the vector x can be reconstructed from η and Φ by solving (5). Random matrices satisfy RIP of any order k with high probability provided that $m = O(k \log(n/k))$ [4, 11]. Therefore, provided x is sufficiently sparse, smaller measurements m suffice to ensure a high-quality reconstruction of x. It is also known that any CS algorithm will require at least $m = \Omega(k \log(n/k))$ for reconstruction [16].

In general, designing or checking whether a sensing matrix satisfies the RIP condition is computationally difficult. RIP only gives a condition on whether a matrix can be used as a sensing matrix but does not necessarily mention how to design one in practice. There are certain random matrices that are known to satisfy the RIP condition with high probability [12, 53]. The most common is the Gaussian matrix, i.e., $\Phi_{m \times n}$ consists of mn independent samples from a zero-mean Gaussian distribution with a variance of $1/m$ [12]. The random binary matrix is another well-studied sensing matrix that is known to satisfy the RIP condition [72].

3.2.2 *Reveal from Compressed Sensing.*

Theorem 2 (Reveal From Compressed Sensing). *Assume that $N_0 \ll N_1$, i.e., the size of the disadvantaged group in the dataset is much smaller than the advantaged group, and $\boldsymbol{H}_{m \times n}$ is a random matrix strongly concentrated around its mean. Then, the protected attribute vector $A = (a_1, a_2, ..., a_j, ..., a_n)$ of the entire test dataset can be obtained using $m = O(N_0 \log(n/N_0))$ statistical-parity-gap queries.*

Proof. To prove Theorem 2, we convert (4) into a compressed sensing problem (we refer the reader to Section 3.2.1 for a background on compressed sensing). Recall from the proof of Theorem 1 that the reconstruction of the protected attributes reduces to solving the linear system of equations in (4), i.e., $\overline{SP} = \mathbf{H}_{m \times n}\mathbf{v}$.

Compressed sensing allows the number of queries m to be much less than n by exploiting the sparsity of one group in the dataset. Let $\mathbf{v} = \bar{r} - \bar{s}$ where $\bar{r} = \begin{pmatrix} 1/N_1 & 1/N_1 & \dots & 1/N_1 \end{pmatrix}^T$ and,

$$\bar{s}_j = \begin{cases} 0 \quad , & \text{if } a_j = 1 \\ \frac{1}{N_1} + \frac{1}{N_0}, & \text{if } a_j = 0 \end{cases}$$

We have, $\overline{SP} = \mathbf{H}(\bar{r} - \bar{s})$, leading to $\overline{SP} - \mathbf{H}\bar{r} = \mathbf{H}\bar{s}$. Now, let $\eta = \overline{SP} - \mathbf{H}\bar{r}$. Then, we have,

$$\eta = \mathbf{H}_{m \times n}\bar{s}. \quad (7)$$

By simple manipulations, we converted (4) into a standard compressed sensing problem (7) where, η is the measurement vector, $\mathbf{H}$ is the sensing matrix, and $\bar{s}$ is a sparse vector with N_0 non-zero entries. From Theorem 1.2 in [10], the unknown vector $\bar{s}$ can be recovered if $\mathbf{H}$ satisfies RIP (see Definition 3.1) of order $2N_0$ with constant $\delta_{2N_0} < \sqrt{2} - 1$. Furthermore, Theorem 9 in Appendix B shows that a random matrix that is strongly concentrated around its expectation satisfies RIP of order $2N_0$ with high probability provided $N_0 \leq cm/\log(n/N_0)$ (implies $2N_0 \leq c'm/\log(n/(2N_0))$) for constants $c, c' > 0$. Therefore $m = O(N_0 \log(n/N_0))$ statistical parity gap queries suffice to successfully reconstruct the protected attribute vector A of the entire test dataset. □

Remark 4. *For the model developers to use the CS technique, they also might need to know N_1 and N_0. This can be found by querying using a model that only accepts one individual and first checking the sign of SP_1. If $SP_1 > 0$, then we know $SP_1 = \frac{1}{N_1}$ and $a_1 = 1$, and hence we can get N_1. We can also obtain $N_0 = n - N_1$. Alternatively, if $SP_1 < 0$, we know $SP_1 = -\frac{1}{N_0}$ (and $a_1 = 0$), and we can get N_0. We can also obtain $N_1 = n - N_0$.*

To effectively apply CS, the vector $\bar{s}$ must be sparse, meaning the size of one group (advantaged or disadvantaged) in the dataset must be significantly smaller than the other group. The sparsity requirement does not have a specific strict threshold. However, the smaller the minority group, the better CS performs with $m = O(N_0 \log(n/N_0))$ models. As the size of the minority group increases, more models are needed.

The sensing matrix $\mathbf{H}$ should satisfy the Restricted Isometry Property (RIP) for CS to work (see Definition 3.1 in Section 3.2.1). Random binary matrices are well known to satisfy this property with high probability [72]. Therefore, choosing models that predict $\{0, 1\}$ randomly would work. However, this might lead to unrealistic models that have low accuracy (since essentially it means that model developers are choosing models with random predictions). Gaussian noise is also proven to satisfy the RIP condition but it may result in a model prediction that lies outside the range of $[0, 1]$, making the models unrealistic. Even if we clip the values to lie between $[0, 1]$, the Gaussian noise may lead to a large variation in accuracy from the base model.

Thus, in Lemma 3, we show that a sensing matrix whose entries are independently sampled from a uniform distribution with variance $1/m$ will satisfy the RIP property needed for CS. In practice, this motivates us to use small bounded noise so that the output values deviate as little as possible (Algorithm 2).

Theorem 3. *Let $\Phi \in \mathbb{R}^{m \times n}$ be a random sensing matrix whose entries are drawn from an i.i.d Uniform$\left(-\sqrt{3/m}, \sqrt{3/m}\right)$ distribution, then the matrix Φ satisfies the Restricted Isometry Property (RIP) of order $k \leq c_1 m/\log(n/k)$ with at least probability $1 - 2e^{-c_2 m}$, for some constant $c_1, c_2 > 0$.*

See proof in Appendix B.

This motivates Algorithm 2, a novel and realistic strategy by which model developers can choose practical models with comparable accuracy to reveal the protected attributes of all the individuals. Essentially, one base model $h_0(\cdot)$ could be trained. For the other m models, small noise sampled from a uniform distribution is added to each output of the base model. If an output value goes outside $[0, 1]$, clip the value to lie between $[0, 1]$.

3.3 Extension to Absolute Statistical Parity Gap

With absolute statistical parity, it is still possible to partition individuals into two groups but no longer possible to determine which group represents the advantaged or disadvantaged populations with certainty. Being able to partition individuals in the test dataset based on their protected attributes is still a privacy infringement.

Algorithm 2 Attribute-Reveal ($m \ll n$)

Train base model $h_0(\cdot)$ (reasonable accuracy)
Choose m models $h_1(\cdot), \ldots, h_m(\cdot)$ as:
for $i = 1, 2, \ldots, m$ **do**
  for $j = 1, 2, \ldots, n$ **do**
    Sample $n_{ij} \sim \text{Unif}(-b, b)$ for a constant b
    $h_i(x_j) = h_0(x_j) + n_{ij}$ (Clip in $[0, 1]$)
Query for Statistical parity gap for each model and store in $\overline{SP}$
Create sensing matrix $\mathbf{H}_{m\times n} = [h_1(X)^T, h_2(X)^T, \ldots, h_m(X)^T]^T$
Compute $\eta = \overline{SP} - \mathbf{H}\bar{r}$
Solve: $\min_{\bar{s}} \|\bar{s}\|_1$ s.t. $\eta = \mathbf{H}\bar{s}$
for $j = 1, 2, \ldots, n$ **do**

$$\text{Detect } \hat{a}_j = \begin{cases} 0, & \text{if } \bar{s}_j > 0.5(\frac{1}{N_1} + \frac{1}{N_0}) \\ 1, & \text{otherwise.} \end{cases}$$

▷ This algorithm is based on Theorem 2

This is mostly due to the ease with which the advantaged and disadvantaged groups can be identified. If somehow the model team could only tell the protected attribute of one individual in a group, e.g., from the other attributes, the partitioning would allow them to learn the protected attribute of the entire test dataset. The partitioned sizes can also be used to determine which group is which. In many cases, the disadvantaged group is often known to be significantly smaller than the advantaged group.

Theorem 4. *Given $m = n$ absolute-statistical-parity-gap queries, there exists a strategy that partitions individuals in the test dataset into two different groups based on their protected attributes.*

Proof. We discuss such a strategy in the proof. Let us use α and β to represent the two partitions of the dataset, i.e., $A \in \{\alpha, \beta\}^n$. Let N_α and N_β denote the size of α and β partitions respectively. Note that $N_\alpha + N_\beta = n$.

First, obtain N_α and N_β. This can be done by querying a model that accepts only one individual. The query will return $|SP_1| = 1/N_\alpha$ or $1/N_\beta$, revealing the size of the partitions. Now, consider the two cases.

Case 1: $N_\alpha \neq N_\beta$. If the size of the two groups is not equal, query a model $h_1(X)$ that accepts only the first individual in the dataset x_1. Assume that the individual belongs to the α partition, i.e., $a_1 = \alpha$. $|SP_1|$ would therefore be $1/N_\alpha$. Then, query a second model, $h_2(X)$, that accepts only the second individual x_2. If $|SP_2| = 1/N_\alpha$, then $a_2 = \alpha$. If $|SP_2| \neq 1/N_\alpha$ then $|SP_2|$ must equal $1/N_\beta$, implying that $a_2 = \beta$. Continue this procedure for every individual until everyone is classified into $a_j = \alpha$ or β.

Case 2: $N_\alpha = N_\beta$. If the size of the two groups is the same, it would not be possible to differentiate between $1/N_\alpha$ and $1/N_\beta$. Hence, a slightly different approach is taken. First, query a model $h_1(X)$ that only accepts x_1 and assume $a_1 = \alpha$, resulting in $|SP_1| = 1/N_\alpha$. Next, query a second model $h_2(X)$ that accepts only x_1 and x_2. The protected attribute of x_2 can be obtained using the query $|SP_2|$, i.e.,

$$|SP_2| = \begin{cases} \frac{2}{N_\alpha}, & \text{if } a_2 = \alpha \\ \frac{1}{N_\alpha} - \frac{1}{N_\beta} = 0, & \text{if } a_2 = \beta \end{cases}$$

In general, to obtain the group of the j-th individual a_j, select a model that accepts only x_1 and x_j. To partition the whole dataset using this technique, the model developers would need at most $m = n$ models and queries. □

Remark 5. *Our results extend to other fairness metrics, such as equalized odds, equal opportunity, and predictive rate parity* [3]*. However, when querying for measures like equal opportunity, the model developers can only identify the protected attributes of individuals with true label $Y = 1$. Since equal opportunity conditions on $Y = 1$, one does not get any information about individuals with $Y = 0$.*

4 DIFFERENTIALLY-PRIVATE APPROACHES TO BIAS ASSESSMENTS

In this section, we discuss approaches to prevent the problem of leaking protected attributes. The main goal is to answer fairness queries as accurately as possible but without leaking the protected attributes of any individual in the test dataset. This motivates us to leverage differential privacy [29, 31].

The notion of ϵ-differential privacy was introduced in [29, 31]. The definition of differential privacy used in this work focuses on keeping the protected attributes private. Because the model developers already have access to a portion of the test dataset (X, Y), we define neighboring datasets as datasets that differ only on one individual's protected attribute A. For $A, A' \in \{0, 1\}^n$, $\mathcal{S} = (X, Y, A)$ and $\mathcal{S}' = (X, Y, A')$ are neighboring if $\|A - A'\|_1 = 1$. Let $\mathcal{D}$ denote a universe of all possible datasets.

Definition 4.1 ((ϵ, δ)-Differential privacy). *Consider any two test datasets $\mathcal{S} = (X, Y, A)$ and $\mathcal{S}' = (X, Y, A')$, where A and A' differ on the protected attribute A of one individual. We say that a randomized mechanism $\mathcal{M}$ is (ϵ, δ)-differentially private if, for all neighbouring $\mathcal{S}, \mathcal{S}'$, and all $\tau \subseteq Range(\mathcal{M})$, we have:*

$$\Pr[\mathcal{M}(\mathcal{S}) \in \tau] \leq e^\epsilon \Pr[\mathcal{M}(\mathcal{S}') \in \tau] + \delta \quad \forall \mathcal{S}, \mathcal{S}' \in \mathcal{D},$$

where the randomness is over the choices made by $\mathcal{M}$ and $\epsilon > 0$ is the privacy budget parameter.

A smaller ϵ introduces greater noise, resulting in enhanced privacy but reduced output accuracy. On the other hand, a larger ϵ incorporates less noise, leading to weaker privacy guarantees but increased output accuracy. Here, δ is the probability of information being accidentally leaked. If $\delta = 0$, $\mathcal{M}$ is ϵ-differentially private. A popular mechanism that achieves ϵ-differential privacy is the Laplace mechanism [32]. The Gaussian mechanism achieves (ϵ, δ)-DP for numeric queries (details in [32, Theorem A.1]).

4.1 Laplace Mechanism for Answering Bias Queries Using Global Sensitivity

We first introduce the definition of global sensitivity for a set of bias queries, e.g., SP queries for a set of m models.

Definition 4.2 (l_1-Global sensitivity [32]). *The l_1-sensitivity of a query function f for all neighboring $\mathcal{S}, \mathcal{S}' \in \mathcal{D}$ is:*

$$\Delta_f = \max_{\mathcal{S}, \mathcal{S}'} \|f(\mathcal{S}) - f(\mathcal{S}')\|_1.$$

[3] A classifier satisfies equalized odds if the individuals in the advantaged and disadvantaged groups have equal expected outcomes given their true labels. A classifier satisfies predictive rate parity if both groups have an equal probability of a subject with positive predictive value truly belonging to the positive class [66].

A naive differentially private technique the compliance team could employ is the Laplace mechanism.

Laplace Mechanism: Given a query function $f : \mathcal{D} \to \mathbb{R}^m$, the Laplace mechanism releases queries as follows:

$$\mathcal{M}(\mathcal{S}, f(\cdot), \epsilon) = f(\mathcal{S}) + (n_1, n_2, ..., n_j, ..., n_m)$$

where n_j are i.i.d. random variables drawn from $\text{Lap}(\Delta_f/\epsilon)$.

Theorem 5. *Given statistical parity gap queries (SP) for m models, the Laplace mechanism that adds noise from* $\text{Lap}(\Delta_{\text{SP}}/\epsilon)$ *to each query is ϵ-differential private, where* $\Delta_{SP} = \frac{m}{2} + \frac{m}{n-1}$.

Theorem 6. *Given absolute statistical parity gap queries ($|SP|$) for m models, the Laplace mechanism that adds noise from* $\text{Lap}(\Delta_{\text{SP}}/\epsilon)$ *to each query is ϵ-differential private, where* $\Delta_{|SP|} = \frac{m}{2}$.

Similarly, for equal opportunity gap EO and absolute equal opportunity gap $|EO|$ the Laplace mechanism adds noise with sensitivity $\Delta_{EO} = \frac{m}{2} + \frac{m}{n-1}$ and $\Delta_{|EO|} = \frac{m}{2}$. See Appendix C for proofs.

Remark 6. *Because the Laplace and Gaussian mechanisms have infinite support $(-\infty, \infty)$, query results can sit outside the range of our fairness metrics ($[-1, 1]$, or $[0, 1]$ for absolute value metrics). In general, these mechanisms do not automatically deal with bounding constraints. Some choose to ignore them and release the raw outputs of the mechanisms since it still satisfies DP's privacy and accuracy guarantees. In our case, probabilities of out-of-bounds values are often small unless ϵ is chosen to be very small. If one insists on having bounded outputs, there are recent approaches [52], such as the truncated and boundary-inflated truncation approaches. Other approaches map out-of-bounds outputs to the boundaries of the metric.*

4.2 Attribute-Conceal: Our Proposed Technique Using Smooth Sensitivity

We have focused on adding noise to the query calibrated to its global sensitivity. However, this might be excessive in many cases, that is, the frameworks would add so much noise that the output would be meaningless. Since we are interested in a particular test dataset $\mathcal{S}$, we define the local sensitivity of a query function f and test dataset $\mathcal{S}$ in l_1 as:

$$\Delta_f^{local}(\mathcal{S}) = \max_{\mathcal{S}': d(S,S')=1} \|f(\mathcal{S}) - f(\mathcal{S}')\|_1. \tag{8}$$

The challenge of calibrating noise to the local sensitivity $\Delta_f^{local}(\mathcal{S})$ is that it might leak information about the test dataset and therefore not sufficient to guarantee DP [56]. To address this, we investigate the idea of smooth sensitivity introduced in [56]. This is an intermediate notion between local and global sensitivity that allows dataset-specific additive noise to be added to achieve DP.

Definition 4.3 (Smooth sensitivity [56]). *For $\beta > 0$, the β-smooth sensitivity of f is*

$$\Delta_{f,\beta}^{smooth}(\mathcal{S}) = \max_{\bar{\mathcal{S}} \in \mathcal{D}} \left(\Delta_f^{local}(\bar{\mathcal{S}}) \cdot e^{-\beta d(\mathcal{S}, \bar{\mathcal{S}})} \right), \tag{9}$$

where $d(\mathcal{S}, \bar{\mathcal{S}})$ denotes the number of entries in which protected attribute vectors A and $\bar{A}$ disagree.

Algorithm 3 Attribute-Conceal for Compliance Team

Input: Model predictions $h_1(\cdot), h_2(\cdot), \ldots, h_m(\cdot)$, N_1, N_0, ϵ
Compute statistical parity gap for each model and store in $\overline{SP}$
Compute Smooth Sensitivity:

$$\Delta_{SP,\beta}^{smooth}(\mathcal{S}) = \max\left(\frac{m}{N_1+1} + \frac{m}{N_0},\ e^{\frac{-\epsilon(N_0-2)}{6m}} \left(\frac{m}{n-1} + \frac{m}{2} \right) \right)$$

for $i = 1, 2, \ldots, m$ **do**
Sample Z_i from a standard Cauchy distribution
$\overline{SP}_i \leftarrow \overline{SP}_i + \frac{6\Delta_{SP,\beta}^{smooth}}{\epsilon} Z_i$ ▷ This algorithm is based on Theorem 7
Return: $\overline{SP}$

Theorem 7 (Dataset Specific ϵ-DP Statistical Parity Query). *Let Z be an m-dimensional random noise with entries independently sampled from a Cauchy distribution $P(z) = \prod_{i=1}^{m} \frac{1}{1+z_i^2}$. For statistical parity gap query (SP), the mechanism $\mathcal{M}(\mathcal{S}) = SP(\mathcal{S}) + \frac{6\Delta_{SP,\beta}^{smooth}(\mathcal{S})}{\epsilon} Z$ is ϵ-differentially private, where the smooth sensitivity is given by:*

$$\Delta_{SP,\beta}^{smooth}(\mathcal{S}) = \max\left(\frac{m}{N_1+1} + \frac{m}{N_0},\ e^{\frac{-\epsilon(N_0-2)}{6m}} \left(\frac{m}{n-1} + \frac{m}{2} \right) \right).$$

Here, N_0 and N_1 are sizes of disadvantaged and advantaged groups in dataset $\mathcal{S}$, such that $N_0 \le N_1$, $N_0 + N_1 = n$, and $\beta = \frac{\epsilon}{6m}$.

Proof. We first find the β-smooth sensitivity of statistical parity gap $\Delta_{SP,\beta}^{smooth}(\mathcal{S})$.

$$\begin{aligned}
\Delta_{SP,\beta}^{smooth}(\mathcal{S}) &\overset{(a)}{=} \max_{\bar{\mathcal{S}} \in \mathcal{D}} \left(\Delta LS_{SP}(\bar{\mathcal{S}}) \cdot e^{-\beta d(\mathcal{S}, \bar{\mathcal{S}})} \right) \\
&\overset{(b)}{=} \max_{k=0,1,\ldots,N_0-2} e^{-k\beta} \left(\max_{\bar{S}: d(S,\bar{S})=k} \Delta LS_{SP}(\bar{S}) \right) \\
&\overset{(c)}{=} \max_{k=0,\ldots,N_0-2} e^{-k\beta} \left(\frac{m}{N_1+k+1} + \frac{m}{N_0-k} \right) \\
&\overset{(d)}{=} \max\left(\frac{m}{N_1+1} + \frac{m}{N_0}, e^{-(N_0-2)\beta} \left(\frac{m}{n-1} + \frac{m}{2} \right) \right).
\end{aligned}$$

Here, (a) is from the definition of smooth sensitivity in (4.3). Next, (b) is the expression of the smooth sensitivity when looking at datasets at distance k (for details see [56, Def 3.1]). To obtain (c) we find the maximum local sensitivity over datasets that are at distance k (see Lemma 5 in Appendix D.3). Finally, (d) holds since the function is convex and hence its maximum occurs at the boundaries $k \in \{0, N_0 - 2\}$ (see Lemma 6 in Appendix D.3). The rest of the proof follows directly from Lemma 2 and Lemma 3 in Appendix D with $\gamma = 2$ and $\beta = \frac{\epsilon}{6m}$. □

To achieve pure differential privacy, noise is introduced following a Cauchy distribution. This motivates Algorithm 3 (Attribute-Conceal), a differentially private technique to answer statistical parity gap queries based on Theorem 7.

We also note that the behavior of the Cauchy distribution can sometimes be unusual, as it does not have an expected value and has heavy tails that decay polynomially, compared to the exponential decay observed in Laplace and Gaussian distributions. In Theorem

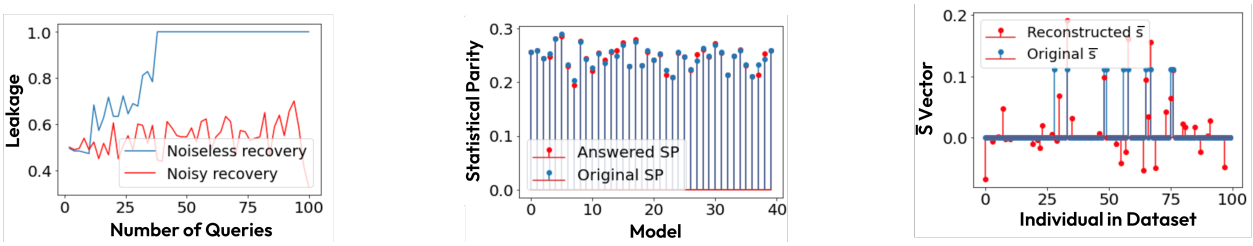

Figure 2: Experimental Results on Adult dataset for test size $n = 100$: (a) Leakage as a function of No. of queries in noisy ($\epsilon = 100$) and noiseless case ($\epsilon = \infty$). Attribute-Conceal prevents Leakage even with an increase in the number of queries. Note that random guessing achieves a Leakage of about 50%, meaning no individual's protected attribute is recovered with certainty. (b) Answered SP queries for $m = 40$ achieves a low Avg. SP err of 7.41×10^{-4}. (c) Reconstructed $\bar{s}$ vector with Attribute-Conceal varies a lot from the original ($\bar{s}$ vector reveals an individual's protected attribute, see equation (7)). Trained models have an Avg. accuracy of 86.23%, and std of 0.2583.

Table 1: Detailed Experimental Results on Adult dataset for test size $n = 100$ and $n = 1000$: Attribute-Conceal (Ours) has much lower Avg. SP err (query error) than Laplace Mechanism (Lap.) for the same privacy parameter ϵ (and similar Leakage).

m	ϵ	$n = 100$				m	ϵ	$n = 1000$			
		Avg. SP err ($\times 10^{-3}$)		Leakage (%)				Avg. SP err ($\times 10^{-3}$)		Leakage (%)	
		Ours	Lap.	Ours	Lap.			Ours	Lap.	Ours	Lap.
25	5	10.9	200.9	50	54	300	5	122.2	874.5	52	49
	10	5.1	112.6	48	49		10	35.7	160.8	50	51
	100	0.05	12.1	58	44		100	2.0	11.5	53	51
	∞	0	0	63	63		∞	0	0	79	79
40	5	77.9	660.2	55	57	400	5	140.0	893.4	52	49
	10	12.3	236.8	49	43		10	42.2	216.3	52	52
	100	0.7	27.6	67	47		100	0.2	54.4	55	49
	∞	0	0	100	100		∞	0	0	100	100

8, we therefore also provide a relaxed (ϵ, δ)-differentially private mechanism that introduces noise from a Laplace distribution.

Theorem 8. *Let Z be random noise samples from a m-dimensional Laplace distribution $P(z) = \frac{1}{2^m} e^{-\|z\|_1}$. For statistical parity gap query (SP), and $\epsilon, \delta \in (0, 1)$, the mechanism $\mathcal{M}(\mathcal{S}) = SP(\mathcal{S}) + \frac{2\Delta^{smooth}_{SP,\beta}}{\epsilon} Z$ is (ϵ, δ)-differentially private, where the smooth sensitivity is given by:*

$$\Delta^{smooth}_{SP,\beta}(\mathcal{S}) = \max\left(\frac{m}{N_1 + 1} + \frac{m}{N_0},\ e^{-\frac{(N_0-2)\epsilon}{4(m+\ln(2/\delta))}}\left(\frac{m}{n-1} + \frac{m}{2}\right)\right).$$

Here, N_0 and N_1 are sizes of disadvantaged and advantaged groups in dataset $\mathcal{S}$, such that $N_0 \leq N_1$, $N_0 + N_1 = n$, and $\beta = \frac{\epsilon}{4(m+\ln(2/\delta))}$.

Proof. The proof follows from Lemma 2 and 4 in Appendix D. Sensitivity analysis follow from proof of Theorem 7 with $\beta = \frac{\epsilon}{4(m+\ln(2/\delta))}$ □

These results can be extended to the absolute statistical parity gap with β-smooth sensitivity (see Appendix D.4), i.e.,

$$\Delta^{smooth}_{|SP|,\beta}(\mathcal{S}) = \max\left(\frac{m}{N_0}, \frac{me^{-(N_0-2)\beta}}{2}\right).$$

5 EXPERIMENTS

We include experimental results on the Adult dataset (see Table 1 and Figure 2) and simulations on synthetic dataset (see Figure 3, Table 2 and Table 3). For the Adult dataset, the protected attribute is race (assumed binary). We restrict ourselves to only White and Black with the latter being relatively sparse (10.4%). We first demonstrate how querying using Attribute-Reveal can leak the protected attributes. Then, we show that Attribute-Conceal effectively prevents this leakage (also outperforming the naive Laplace mechanism). Our performance metrics of interest are (1) Average error in answering the Statistical Parity query (Avg. SP Err); and (2) Accuracy of correctly recovering (essentially leaking) the protected attribute balanced across both races (Leakage, formally defined in Definition 5.1). To observe the tradeoff between Avg. SP Err (query error) and Leakage over a broader range of parameters (privacy parameter ϵ, sparsity N_0, and test size n), we also perform simulations on a synthetic dataset. We provide additional experimental results on the German Credit dataset [25] in Appendix E.

Definition 5.1 (Leakage(%)). *Let N_A be number of individuals in the advantaged group whose protected attribute was correctly predicted and N_B be number of individuals in the disadvantaged group whose protected attribute was correctly predicted.*

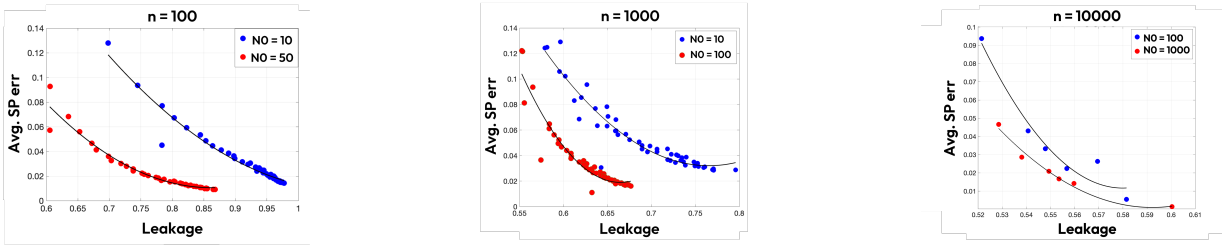

Figure 3: Experimental Results on Synthetic data for test size $n = 100, 1000,$ and 10000: Avg. SP err and Leakage trade-off with Attribute-Conceal. Each point represents an $\epsilon \in [10, 500]$ averaged over 50 runs. Results for varying sparsity N_0 and $m = O(N_0 \log n/N_0)$.

Table 2: Experimental Results on Synthetic data: Avg. SP err and Leakage for test dataset size $n = 1000$ and $n = 10000$ for Attribute-Conceal varying test dataset sparsity and model number m with $\epsilon = 100$.

N_0	$n = 1000$ $\epsilon = 100$				N_0	$n = 10000$ $\epsilon = 100$	
	Avg. SP err ($\times 10^{-3}$)		Leakage (%)			Avg. SP err ($\times 10^{-3}$)	Leakage (%)
	$m = 500$	$m = 900$	$m = 500$	$m = 900$		$m = 1000$	$m = 1000$
10	128.8	191.4	65	64	50	96.9	64
50	90.8	156.3	67	61	100	76.9	61
100	87.4	102.9	62	55	500	38.3	59
200	51.9	96.7	62	47	1000	26.9	59
300	38.9	59.9	48	47	3000	12.8	55
400	41.6	54.5	47	45	5000	9.9	48

Table 3: Experimental Results on Synthetic data: Avg. SP err and Leakage for test size $n = 1000$ and $n = 1000$ for Attribute-Conceal varying N_0 and privacy parameter ϵ.

ϵ	$n = 1000$				$n = 10000$			
	Avg. SP err ($\times 10^{-3}$)		Leakage (%)		Avg. SP err ($\times 10^{-3}$)		Leakage (%)	
	$N_0 = 10$	$N_0 = 100$	$N_0 = 10$	$N_0 = 100$	$N_0 = 100$	$N_0 = 10^3$	$N_0 = 100$	$N_0 = 10^3$
10	929	838	49	51	832.3	678.4	51	50
50	149.3	44.8	65	55	46.5	85.3	55	52
100	83	30.5	71	56	33.1	47.4	55	53
500	26.6	9.7	75	53	31	13.6	57	56

The leakage is defined as:

$$Leakage = \frac{1}{2}\left(\frac{N_A}{N_1} + \frac{N_B}{N_0}\right) \times 100.$$

The leakage is the balanced accuracy of recovery. This is used to deal with imbalanced data, i.e., when one target class appears a lot more than the other.

5.1 Experiments with Adult Dataset

The Adult dataset has 14 attributes for 48842 loan applicants. The classification task is to predict whether an individual's income is more or less that 50K [25]. The feature "race" is chosen as the protected attribute. This feature is excluded from training and only used for statistical parity evaluation. We restrict ourselves to only White and Black (binary) with the latter being relatively sparse (10.4%). We compare Attribute-Conceal with a naive differential privacy technique, Laplace mechanism. We experiment with different test sizes and show our results in Figure 2 and Table 1.

Given an input, our base model $h_0(\cdot)$ outputs a probability value between 0 and 1. For the other m models, we add a small noise sampled from Uniform$(-0.1, 0.1)$ distribution to each output of the base model.

We observe that the accuracy of the other models is quite close to the original. We created 40 models from the base model: they had a mean accuracy of 86.23% and a standard deviation of 0.2583.

Interestingly, our experiments demonstrate that with the uniform noise, we can still recover the protected attributes with far fewer models than the full-rank case. As shown in Figure 2, we are able to recover all the protected attributes using $m = 40$ models. Notice that, this is roughly $O(N_0 \log(n/N_0))$.

Our recovery of protected attributes is based on the values of the $\bar{s}$ vector in Algorithm 2. Ideally, it should be 0 if $a_j = 1$ and $1/N_1 + 1/N_0$ if $a_j = 0$. In our practical implementation, the compressed sensing solution is not always exact but still good enough to infer the protected attribute. Due to this, we use a threshold between 0 and $1/N_1 + 1/N_0$ to identify the protected attribute.

5.2 Experiments with Synthetic Dataset

We perform simulations on synthetic data to observe the trade-off between Avg. SP Err and Leakage over a broader range of parameters (privacy parameter ϵ, sparsity N_0, and test size n). In Figure 3, we show this trade-off with Attribute-Conceal for test size $n = 100, 1000$, and 10000. Each point represents an $\epsilon \in [10, 500]$ averaged over 50 runs. We show results for varying sparsity N_0 and $m = O(N_0 \log n/N_0)$. Table 2 and Table 3 provide additional experimental results highlighting the Avg. SP err and Leakage for test size $n = 1000$ and $n = 10000$ for different sparsity N_0, the model number m, and the privacy parameter ϵ. A clear trend observed is that Attribute-Conceal results in a significantly lower Avg. SP error compared to the Laplace mechanism, for a similar level of protected attribute leakage.

6 CONCLUSION AND FUTURE WORK

This work highlights a major concern with fairness assessments in scenarios where protected attributes such as gender or race cannot be accessed during model training. Showing that simply querying for fairness metrics can leak sensitive information to model developers raises important questions about the ethical implications of these assessments. As a remedy, we also propose a novel technique, Attribute-Conceal, which achieves differential privacy by calibrating noise to the smooth sensitivity of our bias query.

The results of this study have important implications for regulations and privacy in the field of algorithmic fairness and provide a new approach to protect the sensitive information of individuals in fairness assessments. This also provides a potential resolution to the continuing debate about whether protected attributes should be used in training. Future research could look into expanding the framework to include other fairness metrics or incorporating these techniques into training or post-processing to directly reduce bias without leaking protected attributes.

Our current approach assumes that both model developers and the compliance (or auditing) team work with the same test set. However, this might not hold true in every context. The compliance/auditing team may choose to use a different test set. However, note that a different test set may not adequately represent the true training distribution, which could potentially affect generalization.

We note that while our focus is on leakage from bias queries, future work could also look into *inferring* the protected attributes from the other available features using alternate techniques [7, 8]. For example, if one has prior knowledge that a feature such as hours-worked-per-week is strongly correlated with gender, one might just be able to *infer* gender with reasonable accuracy from that feature. However, it remains debatable if such *indirect inferring* of protected attribute from correlated features would legally constitute a violation of disparate treatment (or privacy). On the other hand, asking the compliance team for bias assessments actually accesses the protected attributes using queries. We do make a distinction between *leaking* and *inferring* protected attributes here. An interesting scenario would arise if one exploits a synergy of both bias queries as well as inference mechanisms to obtain even more accurate predictions of protected attributes than using either of them individually, and if such techniques would constitute a violation of anti-discrimination and privacy.

REFERENCES

[1] Daniel Alabi. 2019. The cost of a reductions approach to private fair optimization. *arXiv preprint arXiv:1906.09613* (2019).

[2] Wael Alghamdi, Hsiang Hsu, Haewon Jeong, Hao Wang, Peter Michalak, Shahab Asoodeh, and Flavio Calmon. 2022. Beyond ADULT and COMPAS: Fair multi-class prediction via information projection. *Advances in Neural Information Processing Systems* 35 (2022), 38747–38760.

[3] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. *Differential Privacy Has Disparate Impact on Model Accuracy.* Proceedings of the 33rd International Conference on Neural Information Processing Systems, NY, USA.

[4] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. 2008. A simple proof of the restricted isometry property for random matrices. *Constructive approximation* 28 (2008), 253–263.

[5] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104 (2016), 671.

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

[7] Flavio Calmon and Nadia Fawaz. 2012. Privacy against statistical inference. In *2012 50th annual Allerton conference on communication, control, and computing (Allerton)*. IEEE, 1401–1408.

[8] Flavio Calmon, Ali Makhdoumi, and Muriel Médard. 2015. Fundamental limits of perfect privacy. In *2015 IEEE International Symposium on Information Theory (ISIT)*. 1796–1800. https://doi.org/10.1109/ISIT.2015.7282765

[9] E.J. Candes and T. Tao. 2005. Decoding by linear programming. *IEEE Transactions on Information Theory* 51, 12 (2005), 4203–4215.

[10] Emmanuel J. Candes. 2008. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique* 346, 9 (2008), 589–592.

[11] Emmanuel J Candes, Justin K Romberg, and Terence Tao. 2006. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 59, 8 (2006), 1207–1223.

[12] Emmanuel J. Candes and Terence Tao. 2006. Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies? *IEEE Transactions on Information Theory* 52, 12 (2006), 5406–5425.

[13] Emmanuel J. Candes and Michael B. Wakin. 2008. An Introduction To Compressive Sampling. *IEEE Signal Processing Magazine* 25, 2 (2008), 21–30.

[14] Jiahao Chen. 2018. Fair lending needs explainable models for responsible recommendation. *ArXiv* abs/1809.04684 (2018).

[15] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. 339–348.

[16] Albert Cohen, Wolfgang Dahmen, and Ronald Devore. 2009. Compressed sensing and best k-term approximation. *Journal of the American Mathematical Society* 22, 1 (2009), 211–231.

[17] Ethan Cohen-Cole. 2011. CREDIT CARD REDLINING. *The Review of Economics and Statistics* 93, 2 (2011), 700–713. http://www.jstor.org/stable/23015965

[18] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. 2019. Fair transfer learning with missing protected attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 91–98.

[19] Marsha Courchane, David Nebhut, and David Nickerson. 2000. Lessons Learned: Statistical Techniques and Fair Lending. *Journal of Housing Research* 11, 2 (2000), 277–295.

[20] Emiliano De Cristofaro. 2020. An Overview of Privacy in Machine Learning. *CoRR* abs/2005.08679 (2020). arXiv:2005.08679 https://arxiv.org/abs/2005.08679

[21] Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. 2019. On the Compatibility of Privacy and Fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization (UMAP'19 Adjunct)*. 309–315.

[22] Anupam Datta, Matthew Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. 2017. Use Privacy in Data-Driven Systems: Theory and Experiments with Machine Learnt Programs. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS*. ACM, 1193–1210.

[23] Leo de Castro, Jiahao Chen, and Antigoni Polychroniadou. 2020. CryptoCredit: Securely Training Fair Models. In *Proceedings of the First ACM International Conference on AI in Finance*. Article 17, 8 pages.

[24] D.L. Donoho. 2006. Compressed sensing. *IEEE Transactions on Information Theory* 52, 4 (2006), 1289–1306.

[25] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[26] Sanghamitra Dutta, Praveen Venkatesh, Piotr Mardziel, Anupam Datta, and Pulkit Grover. 2020. An information-theoretic quantification of discrimination with exempt features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 3825–3833.

[27] Sanghamitra Dutta, Praveen Venkatesh, Piotr Mardziel, Anupam Datta, and Pulkit Grover. 2021. Fairness Under Feature Exemptions: Counterfactual and Observational Measures. *IEEE Transactions on Information Theory* 67, 10 (2021), 6675–6710.

[28] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. 2020. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International Conference on Machine Learning*. PMLR, 2803–2813.

[29] Cynthia Dwork. 2006. Differential Privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II* (Venice, Italy) *(ICALP'06)*. Berlin, Heidelberg, 1–12.

[30] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. 214–226.

[31] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Proceedings of the Third Conference on Theory of Cryptography* (New York, NY) *(TCC'06)*. 265–284.

[32] Cynthia Dwork and Aaron Roth. 2014. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.* 9, 3–4 (aug 2014), 211–407.

[33] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. 259–268.

[34] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS '15)*. 1322–1333.

[35] Maya R. Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. 2018. Proxy Fairness. *CoRR* abs/1806.11212 (2018). arXiv:1806.11212 http://arxiv.org/abs/1806.11212

[36] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. 2125–2126.

[37] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, Vol. 29. 3315–3323.

[38] Hui Hu, Yijun Liu, Zhen Wang, and Chao Lan. 2019. A distributed fair machine learning framework with private demographic data protection. In *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1102–1107.

[39] Hongsheng Hu, Zoran Salcic, Gillian Dobbie, and Xuyun Zhang. 2021. Membership Inference Attacks on Machine Learning: A Survey. *CoRR* abs/2103.07853 (2021). arXiv:2103.07853 https://arxiv.org/abs/2103.07853

[40] Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi Malvajerdi, and Jonathan Ullman. 2019. Differentially Private Fair Learning. In *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97. PMLR, 3000–3008.

[41] Marc Juarez and Aleksandra Korolova. 2022. "You Can't Fix What You Can't Measure": Privately Measuring Demographic Performance Disparities in Federated Learning. *arXiv preprint arXiv:2206.12183* (2022).

[42] Nathan Kallus, Xiaojie Mao, and Angela Zhou. 2022. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science* 68, 3 (2022), 1959–1981.

[43] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, 643–650.

[44] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80. PMLR, 2564–2572.

[45] Niki Kilbertus, Adria Gascon, Matt Kusner, Michael Veale, Krishna Gummadi, and Adrian Weller. 2018. Blind Justice: Fairness with Encrypted Sensitive Attributes. In *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80. PMLR, 2630–2639.

[46] Preethi Lahoti, Krishna P. Gummadi, and Gerhard Weikum. 2019. iFair: Learning Individually Fair Data Representations for Algorithmic Decision Making. In *2019 IEEE 35th International Conference on Data Engineering*. 1334–1345.

[47] Preethi Lahoti, Krishna P. Gummadi, and Gerhard Weikum. 2019. Operationalizing individual fairness with pairwise fair representations. *Proceedings of the VLDB Endowment* 13, 4 (Dec 2019), 506–518.

[48] Anja Lambrecht and Catherine Tucker. 2019. Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science* 65, 7 (2019), 2966–2981.

[49] Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. 2018. Does mitigating ML's impact disparity require treatment disparity?. In *Advances in Neural Information Processing Systems*, Vol. 31. 8136–8146.

[50] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. 2021. When Machine Learning Meets Privacy: A Survey and Outlook. *ACM Comput. Surv.* 54, 2, Article 31 (mar 2021), 36 pages.

[51] Fang Liu. 2019. Generalized Gaussian Mechanism for Differential Privacy. *IEEE Transactions on Knowledge and Data Engineering* 31, 4 (Apr 2019), 747–756.

[52] Fang Liu. 2019. Statistical Properties of Sanitized Results from Differentially Private Laplace Mechanism with Univariate Bounding Constraints. *Trans. Data Priv.* 12 (2019), 169–195.

[53] Mahsa Lotfi and Mathukumalli Vidyasagar. 2020. Compressed Sensing Using Binary Matrices of Nearly Optimal Dimensions. *IEEE Transactions on Signal Processing* 68 (2020), 3008–3021.

[54] Julien Mairal, Francis R. Bach, and Jean Ponce. 2014. Sparse Modeling for Image and Vision Processing. *CoRR* abs/1411.3230 (2014). arXiv:1411.3230 http://arxiv.org/abs/1411.3230

[55] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[56] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2007. Smooth Sensitivity and Sampling in Private Data Analysis. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*. New York, NY, USA, 75–84.

[57] Jessica L Roberts. 2014. Protecting privacy to prevent discrimination. *Wm. & Mary L. Rev.* 56 (2014), 2097.

[58] George Rutherglen. 1987. Disparate Impact under Title VII: An Objective Theory of Discrimination. *Virginia Law Review* 73, 7 (1987), 1297–1345.

[59] Ali Shahin Shamsabadi, Mohammad Yaghini, Natalie Dullerud, Sierra Wyllie, Ulrich Aïvodji, Aisha Alaagib, Sébastien Gambs, and Nicolas Papernot. 2022. Washing The Unwashable: On The (Im) possibility of Fairwashing Detection. *Advances in Neural Information Processing Systems* 35 (2022), 14170–14182.

[60] Megan Smith, Cecilia Munoz, and D. J. Patil. 2016. Big Risks, Big Opportunities: the Intersection of Big Data and Civil Rights. https://obamawhitehouse.archives.gov/blog/2016/05/04/big-risks-big-opportunities-intersection-big-data-and-civil-rights

[61] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. 2020. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems* 33 (2020), 19339–19352.

[62] Kush R Varshney. 2019. Trustworthy machine learning and artificial intelligence. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 26–29.

[63] Kush R Varshney. 2022. *Trustworthy Machine Learning*. Independently Published.

[64] Michael Veale and Reuben Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society* 4, 2 (2017).

[65] Akshaj Kumar Veldanda, Ivan Brugere, Sanghamitra Dutta, Alan Mishler, and Siddharth Garg. 2023. Hyper-parameter Tuning for Fair Classification without Sensitive Attribute Access. *arXiv preprint arXiv:2302.01385* (2023).

[66] Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)* (Gothenburg, Sweden). IEEE, Piscataway, New Jersey, 1–7.

[67] Neil Vigdor. 2019. Apple Card Investigated After Gender Discrimination Complaints. https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html

[68] Meredith Whittaker, Meryl Alper, Cynthia L. Bennett, Sara Hendren, Liz Kaziunas, Mara Mills, Meredith Ringel Morris, Joy Rankin, Emily Rogers, Marcel Salas, and Sarah Myers West. 2019. *Disability, Bias, and AI.* Technical Report MSR-TR-2019-38. AI Now Institute.

[69] Depeng Xu, Shuhan Yuan, and Xintao Wu. 2019. Achieving Differential Privacy and Fairness in Logistic Regression. In *Companion Proceedings of The 2019 World Wide Web Conference (WWW '19)*. 594–599.

[70] Tom Yan and Chicheng Zhang. 2022. Active fairness auditing. In *International Conference on Machine Learning*. PMLR, 24929–24962.

[71] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. 1171–1180.

[72] Gesen Zhang, Shuhong Jiao, Xiaoli Xu, and Lan Wang. 2010. Compressed sensing and reconstruction with Bernoulli matrices. In *The 2010 IEEE International Conference on Information and Automation* (Harbin, China). Piscataway, New Jersey, 455–460.

[73] Michael J Zimmer. 1995. Emerging Uniform Structure of Disparate Treatment Discrimination Litigation. *Georgia Law Review* 30 (1995), 563.