Improving Fairness and Privacy in Selection Problems

Mohammad Mahdi Khalili,[1] **Xueru Zhang,** [2] **Mahed Abroshan,** [3] **Somayeh Sojoudi** [4]
[1] CIS Department, University of Delaware, Newark, DE, USA
[2] EECS Department, University of Michigan, Ann Arbor, MI, USA
[3] Alan Turing Institute, London, UK
[4] EECS Department, University of California, Berkeley, CA, USA
khalili@udel.edu, xueru@umich.edu, mabroshan@turing.ac.uk, sojoudi@berkeley.edu

Abstract

Supervised learning models have been increasingly used for making decisions about individuals in applications such as hiring, lending, and college admission. These models may inherit pre-existing biases from training datasets and discriminate against protected attributes (e.g., race or gender). In addition to unfairness, privacy concerns also arise when the use of models reveals sensitive personal information. Among various privacy notions, differential privacy has become popular in recent years. In this work, we study the possibility of using a differentially private exponential mechanism as a post-processing step to improve both fairness and privacy of supervised learning models. Unlike many existing works, we consider a scenario where a supervised model is used to select a limited number of applicants as the number of available positions is limited. This assumption is well-suited for various scenarios, such as job application and college admission. We use "equal opportunity" as the fairness notion and show that the exponential mechanisms can make the decision-making process perfectly fair. Moreover, the experiments on real-world datasets show that the exponential mechanism can improve both privacy and fairness, with a slight decrease in accuracy compared to the model without post-processing.

1 Introduction

Machine learning (ML) algorithms trained based on real-world datasets have been used in various decision-making applications (e.g., job applications and criminal justice). Due to the pre-existing bias in the datasets, the decision-making process can be biased against protected attributes (e.g., race and gender). For example, COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) recidivism prediction tool used as part of inmate parole decisions by courts in the United States has been shown to have a substantially higher false positive rate on African Americans compared to White people (Dressel and Farid 2018). In speech recognition, products such as Amazon's Alexa and Google Home can have accent bias, with Chinese-accented and Spanish-accented English hardest to understand (Harwell 2018). Amazon had been using automated software since 2014 to assess applicants resumes, which was found to be biased against women (Dastin 2018).

There are various potential causes for such discrimination. Bias can be introduced when the data is collected. For instance, if a group (i.e., majority group) contributes more to the dataset as compared to another group (i.e., minority group), then the model trained based on the dataset could be biased in favor of the majority group, and this group may experience a higher accuracy. Even if the data collection procedure is unbiased, the data itself may be biased, e.g., labels in the training dataset may exhibit bias if they are provided based on an agent's opinion (Jiang and Nachum 2019).

The fairness issue has been studied extensively in the literature. A variety of fairness notions have been proposed to measure the unfairness, and they can be roughly classified into two families: *individual fairness* and *group fairness*. Individual fairness is in pursuit of equity in the individual-level, that it requires any two similar individuals to be treated similarly (Biega, Gummadi, and Weikum 2018; Jung et al. 2019; Gupta and Kamble 2019). Group fairness aims to achieve a certain balance in the group-level, that the population is partitioned into a small number of protected groups and it requires a certain statistical measure (e.g., positive classification rates, true positive rates, etc.) to be approximately equalized across different protected groups (Zhang et al. 2019; Hardt, Price, and Srebro 2016; Conitzer et al. 2019; Zhang, Khalili, and Liu 2020; Zhang et al. 2020). There are mainly three approaches to improving fairness (Zhang and Liu 2020):

i) *Pre-processing:* modifying the training datasets to remove the discrimination before training an ML model (Kamiran and Calders 2012; Zemel et al. 2013);

ii) *In-processing:* imposing certain fairness criterion or modifying the loss function during the training process (Agarwal et al. 2018; Zafar et al. 2019);

iii) *Post-processing:* altering the output of an existing algorithm to satisfy the fairness requirements (Hardt, Price, and Srebro 2016; Pleiss et al. 2017).

In this work, we focus on group fairness and use the post-processing approach to improving the fairness of a supervised learning algorithm.

In addition to unfairness issues, privacy concerns may incur when making decisions based on individuals' sensitive data. Consider a lending scenario where the loan approval decision is made based on an applicant's credit score. The

decision-making outcome can reflect the applicants financial situation and hence compromise his/her privacy.

Various privacy-preserving techniques have emerged in recent years to protect individual privacy, such as randomizing or anonymizing sensitive data (Liu, Xie, and Wang 2017; Sweeney 2002; Zhu and Liu 2004; Wang et al. 2012). Among them, differential privacy (Dwork 2006) as a statistical notion of privacy has been extensively studied and deployed in practice. It ensures that no one, from an algorithm's outcome, can infer with substantial higher confidence than random guessing whether a particular individual's data was included in the data analysis. Various mechanisms have been developed to generate differentially private outputs such as *exponential mechanism* (McSherry and Talwar 2007) and *Laplace mechanism* (Dwork et al. 2006).

In this paper, we consider a scenario where a decision-maker (e.g., company) aims to accept $m \geq 1$ people from an applicant pool. This scenario can be referred to as a *selection problem*. Each applicant has a hidden qualification state (e.g., capability for tasks) and observable features (e.g., GPA, interview performance). Suppose there is a supervised learning model that has been trained in advance and can be used for assigning each applicant a qualification score based on the features and queried by the decision-maker. These qualification scores can represent the likelihood that applicants are qualified, and the decision-maker selects m applicants solely based on the qualification scores. During this process, privacy and fairness concerns may arise. As such, the decision maker's goal is to select m applicants among those that are most likely to be qualified, and at the same time, preserve individual privacy and satisfy a fairness constraint. Moreover, we allow the already trained supervised learning model to be a "black-box"; the decision-maker has no access to the model parameters and can only use it to observe an applicant's score. Within this context, we study the possibility of using an exponential mechanism as a post-processing scheme to improve both the privacy and fairness of the pre-trained supervised learning model. We consider "equal opportunity" as the notion of fairness[1] and examine the relationship between fairness, privacy, and accuracy.

Related work. The most related works of this paper are (Hardt, Price, and Srebro 2016; Jagielski et al. 2019; Cummings et al. 2019; Mozannar, Ohannessian, and Srebro 2020; Kleinberg and Raghavan 2018). (Kleinberg and Raghavan 2018) studies the effects of implicit bias on selection problems, and explore the role of the *Rooney Rule* in this process. They show that the Rooney Rule can improve both the decision maker's utility and the disadvantaged group's representation. However, neither the fairness notion nor the privacy issues are considered in this work. (Hardt, Price, and Srebro 2016) introduces the notion of equal opportunity and develop a post-processing algorithm to improve the fairness of a supervised learning model. This work solely focuses on the fairness issue and does not provide any privacy guarantee. (Cummings et al. 2019) studies fairness in supervised learning and its relationship to differential privacy. In particular, they show that it is *impossible* to train a differentially private classifier that satisfies exact (perfect) fairness and achieves a higher accuracy than a constant classifier. Therefore, many works in the literature have focused on developing approximately fair and differentially private algorithms. For instance, (Xu, Yuan, and Wu 2019) introduces an algorithm to train a differentially private logistic regression model that is approximately fair. (Jagielski et al. 2019) develops post-processing and in-processing algorithms to train a classifier that satisfies approximate fairness and protects the privacy of protected attributes (not the training data). (Mozannar, Ohannessian, and Srebro 2020) considers the notion of *local* differential privacy and aims to learn a fair supervised model from the data with the noisy and differentially private protected attributes. Similarly, (Wang et al. 2020; Kallus, Mao, and Zhou 2019; Awasthi, Kleindessner, and Morgenstern 2020) focus on fair learning using noisy protected attributes but without a privacy guarantee.

Main contributions. Most of the existing work aims to learn a model that minimizes the expected loss (e.g., classification error) over the entire population under certain fairness constraints. In settings such as hiring, lending, and college admission, it means the decision-maker should accept all the applicants as long as they are likely to be qualified. However, this may not be realistic for many real-world applications, when only a fixed number of positions are available, and only a limited number of applicants can be selected. In this paper, we shall consider this scenario where only a fixed number of people are selected among all applicants. Using equal opportunity as the fairness notion and differential privacy as the privacy notion, we identify sufficient conditions under which the exponential mechanism, in addition to a differential privacy guarantee, can also achieve *perfect* fairness. Our results show that the negative result shown in (Cummings et al. 2019) (i.e., it is impossible to attain a perfectly fair classifier under differential privacy) does not apply to our setting when the number of acceptance is limited. In summary, our main contributions are as follows:

- We show that although the exponential mechanism has been designed and used mainly for preserving individual privacy, it is also effective in improving fairness. The sufficient conditions under which the exponential mechanism can achieve *perfect* fairness are identified.
- We show that the accuracy of a supervised learning model after using the exponential mechanism is monotonic in privacy leakage, which implies that the improvement of fairness and privacy is at the cost of accuracy.
- Unlike (Cummings et al. 2019), in our setting, we show that compared to other trivial algorithms (e.g., uniform random selection) that are perfectly fair, the exponential mechanism can achieve a higher accuracy while maintaining perfect fairness.

The remainder of the paper is organized as follows. We present our model in Section 2. The relation between fairness, privacy and accuracy is examined in Section 3. The generalization to a scenario with m available positions is discussed in Section 4. We present the numerical experiments in Section 5 and conclude the paper in Section 6.

[1]We discuss the generalization of our results to demographic parity in the appendix.

2 Model

Consider a scenario where n individuals indexed by $\mathcal{N} = \{1, 2, \cdots, n\}$ apply for some jobs/tasks. Each individual i can be characterized by a tuple (X_i, A_i, Y_i), where $Y_i \in \{0, 1\}$ is the hidden qualification state representing whether i is qualified ($Y_i = 1$) for the position or not ($Y_i = 0$), $X_i \in \mathcal{X}$ is the observable features and $A_i \in \{0, 1\}$ is the protected attribute (e.g., race, gender) indicating the group membership of individual i. Tuples $\{(X_i, A_i, Y_i)|i = 1, \ldots, n\}$ are i.i.d. random variables following some distribution F. We allow X_i to be correlated with A_i, and it may include A_i as well. The decision-maker observes the applicants' features and aims to select m people that are most likely to be qualified and satisfy certain privacy and fairness constraints.

Pre-trained model and qualification scores. We assume there is a supervised learning model $r : \mathcal{X} \to \mathcal{R}$ that has been trained in advance and can be queried by the decision-maker. It takes features of each applicant as input, and outputs a qualification score indicating the likelihood that the applicant is qualified. Let $\mathcal{R} \coloneqq \{\rho_1, \ldots, \rho_{n'}\} \subset [0, 1]$ be a set of all possible values for the qualification score, and define $R_i = r(X_i)$ as individual i's qualification score. The higher R_i implies individual i is more likely to be qualified, i.e., $Y_i = 1$. Note that R_i depends on A_i through X_i since X_i and A_i are correlated, and X_i may include A_i. Without loss of generality, let $\rho_1 = 0$ and $\rho_{n'} = 1$.

Selection procedure. Let $\mathbf{D} = (X_1, \ldots, X_n)$ be a database that includes all the applicants' features, and D be its realization. D is the only information that the decision-maker can observe about applicants. The decision-maker first generates all applicants' qualification scores using pre-trained model $r(\cdot)$, and then uses these scores $(R_1, \ldots, R_n)$ to select m individuals. We first focus on a case when $m = 1$, i.e., only one applicant is selected, even though there could be more than one qualified applicant in the applicant pool. The generalization to $m > 1$ is studied in Section 4.

For notational convenience, we further define tuple (X, A, Y) as a random variable that also follows distribution F, and $R = r(X)$. Denote (x, a, y) as a realization of (X, A, Y). Similar to (Hardt, Price, and Srebro 2016), we assume F can be learned during the training process and is known to the decision-maker.

2.1 Differential Privacy

Let $x_i \in \mathcal{X}$ be the observable features of individual i, and $D = (x_1, x_2, \ldots, x_n)$ be a database which includes all individuals' data. Moreover, $\mathcal{D} = \{(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)|\hat{x}_i \in \mathcal{X}\}$ denotes the set of all possible databases.

Definition 1 (Neighboring Databases)**.** *Two databases $D = (x_1, \ldots, x_n)$ and $D' = (x'_1, \ldots, x'_n)$ are neighboring databases if they differ only in one data point, noted as $D \sim D'$, i.e.,*

$$\exists i \in \mathcal{N} \text{ s.t. } x_i \neq x'_i \text{ and } x_j = x'_j \ \forall j \neq i.$$

Definition 2 (Differential Privacy (Dwork 2006))**.** *A randomized algorithm $\mathscr{M}$ is ϵ-differentially private if for any two neighboring databases D and D' and for any possible set of output $\mathcal{W} \subseteq$ Range($\mathscr{M}$), it holds that $\frac{\Pr\{\mathscr{M}(D)\in\mathcal{W}\}}{\Pr\{\mathscr{M}(D')\in\mathcal{W}\}} \leq \exp\{\epsilon\}$.*

Privacy parameter $\epsilon \in [0, \infty)$ can be used to measure privacy leakage; the smaller ϵ corresponds to the stronger privacy guarantee. For sufficiently small ϵ, the distribution of output remains almost the same as a single data point in the database changes. It suggests that an attacker cannot infer the input data with high confidence after observing the output; thus, individual privacy is preserved. Next, we introduce a notable mechanism that can achieve differential privacy.

Definition 3 (Exponential mechanism (McSherry and Talwar 2007))**.** *Denote $\mathcal{O} = \{o_1, \ldots, o_{\hat{n}}\}$ as the set of all possible outputs of algorithm $\mathscr{M}$, and $v : \mathcal{O} \times \mathcal{D} \to \mathbb{R}$ as a score function, where a higher value of $v(o_i, D)$ implies that output o_i is more appealing under database D. Let $\Delta = \max_{i,D\sim D'} |v(o_i, D) - v(o_i, D')|$ be defined as the sensitivity of score function. Then, exponential mechanism $\mathscr{M} : \mathcal{D} \to \mathcal{O}$ that satisfies ϵ-differential privacy selects $o_i \in \mathcal{O}$ with probability $\Pr\{\mathscr{M}(D) = o_i\} = \frac{\exp\{\epsilon\cdot\frac{v(o_i,D)}{2\Delta}\}}{\sum_{j=1}^{\hat{n}}\exp\{\epsilon\cdot\frac{v(o_j,D)}{2\Delta}\}}$.*

2.2 Selection Using Exponential Mechanism

Given a set of qualification scores $(R_1, \ldots, R_n)$ generated from a pre-trained model $r(\cdot)$, the decision-maker selects an individual based on them, and meanwhile tries to preserve privacy with respect to database $\mathbf{D} = (X_1, \ldots, X_n)$. To this end, the decision-maker makes a selection using the exponential mechanism with score function $v : \mathcal{N} \times \mathcal{D} \to [0, 1]$, where $\mathcal{D} = \mathcal{X}^n$ is the set of all possible databases.

One natural choice of the score function would be $v(i, D) = r(x_i)$, i.e., an applicant with a higher qualification score is more likely to be selected. Because $0 \leq r(x) \leq 1$, for all $x \in \mathcal{X}$, the sensitivity of score function $v(i, D)$ is $\Delta = \max_{i,D\sim D'} |v(i, D) - v(i, D')| = 1$.

Let $\mathscr{A}_\epsilon : \mathcal{D} \to \mathcal{N}$ be an ϵ-differentially private exponential mechanism used by the decision-maker to select one individual. Using $\mathscr{A}_\epsilon(\cdot)$, after observing realizations of $(R_1, R_2, \ldots, R_n)$, individual $i \in \mathcal{N}$ is selected with probability $\Pr\{\mathscr{A}_\epsilon(D) = i\} = \frac{\exp\{\epsilon\cdot\frac{r_i}{2}\}}{\sum_{j\in\mathcal{N}}\exp\{\epsilon\cdot\frac{r_j}{2}\}}$, where r_i is the realization of random variable R_i. For each individual i, define a Bernoulli random variable $I_{i,\epsilon} \in \{0, 1\}$ indicating whether i is selected ($I_{i,\epsilon} = 1$) under algorithm $\mathscr{A}_\epsilon(\cdot)$ or not ($I_{i,\epsilon} = 0$). We have,

$$\begin{aligned}
&\Pr\{I_{i,\epsilon} = 1\} \\
=& \sum_{(r_1,\ldots,r_n)\in\mathcal{R}^n} \Pr\left\{I_{i,\epsilon} = 1|\cap_{j=1}^n \{R_j = r_j\}\right\} \cdot \prod_{j=1}^n f_R(r_j) \\
=& \sum_{(r_1,\ldots,r_n)\in\mathcal{R}^n} \frac{\exp\{\epsilon\cdot\frac{r_i}{2}\}}{\sum_{j=1}^n\exp\{\epsilon\cdot\frac{r_j}{2}\}} \cdot \prod_{j=1}^n f_R(r_j),
\end{aligned}$$

where $f_R(\cdot)$ is the probability mass function (PMF) of random variable R. If we further define random variable $Z_{i,\epsilon} = \frac{\exp\{\epsilon\cdot\frac{R_i}{2}\}}{\sum_{j\in\mathcal{N}}\exp\{\epsilon\cdot\frac{R_j}{2}\}}$ and denote the expectation of $Z_{i,\epsilon}$ by $E\{Z_{i,\epsilon}\}$, then $E\{Z_{i,\epsilon}\} = \Pr\{I_{i,\epsilon} = 1\}$ holds.

2.3 Fairness Metric

Based on the protected attribute A, n applicants can be partitioned into two groups. To measure the unfairness between two groups resulted from using algorithm $\mathscr{A}_\epsilon(\cdot)$, we shall adopt a group fairness notion. For the purpose of exposition, we focus on one of the most commonly used notions called *equal opportunity* (Hardt, Price, and Srebro 2016). The generalization to *demographic parity* fairness (Dwork et al. 2012) is discussed in the appendix.

In binary classification, *equal opportunity* fairness requires that the true positive rates experienced by different groups to be equalized, i.e., $\Pr\{\widehat{Y}=1|A=0,Y=1\} = \Pr\{\widehat{Y}=1|A=1,Y=1\}$, where $\widehat{Y}$ is the predicted label by the classifier. In our problem when the number of acceptance is $m=1$, this definition can be adjusted as follows,

Definition 4 (Fairness metric)**.** *Consider an algorithm $\mathscr{M}: \mathcal{D} \to \mathcal{N}$ that selects one individual from n applicants. Given database $\mathbf{D}$ and $\mathscr{M}(\cdot)$, for all $i \in \mathcal{N}$ define a Bernoulli random variable K_i such that $K_i = 1$ if $\mathscr{M}(\mathbf{D}) = i$ and $K_i = 0$ otherwise. Then algorithm $\mathscr{M}(\cdot)$ is γ-fair if*

$$\Pr\{K_i=1|A_i=0,Y_i=1\} - \Pr\{K_i=1|A_i=1,Y_i=1\} = \gamma.$$

Note that $-1 \le \gamma \le 1$, and negative (resp. positive) γ implies that algorithm $\mathscr{M}(\cdot)$ is biased in favor of the group with protected attribute $A=1$ (resp. $A=0$). In particular, we say $\mathscr{M}(\cdot)$ is ***perfectly fair*** if $\gamma = 0$.[2]

For algorithm $\mathscr{A}_\epsilon(\cdot)$ in Section 2.2 that selects an individual using the exponential mechanism, γ in above definition can be equivalently written as

$$E\{Z_{i,\epsilon}|A_i=0,Y_i=1\} - E\{Z_{i,\epsilon}|A_i=1,Y_i=1\} = \gamma. \quad (1)$$

2.4 Accuracy Metric

It is easy to develop trivial algorithms that are both 0-differentially private and 0-fair. For example, $\mathscr{A}_0(\cdot)$ which selects one individual randomly and uniformly from $\mathcal{N}$, or a deterministic algorithm that always selects a particular individual. However, both algorithms do not use qualification scores to make decisions. The primary goal of the decision-maker that selecting the most qualified individuals is thus undermined. We need to introduce another metric to evaluate the ability of $\mathscr{A}_\epsilon(\cdot)$ to select qualified individuals.

Definition 5 (Accuracy)**.** *An algorithm $\mathscr{M}: \mathcal{D} \to \mathcal{N}$ that selects one individual from n applicants is θ-accurate if*

$$\Pr\{Y_{\mathscr{M}(\mathbf{D})}=1\} = \theta. \quad (2)$$

As an example, for algorithm $\mathscr{A}_0(\cdot)$ that selects one individual uniformly at random from $\mathcal{N}$, it is θ-accurate with $\theta = \Pr\{Y_{\mathscr{A}_0(\mathbf{D})}=1\} = \Pr\{Y=1\}$. The goal of this work is to examine whether it is possible to use exponential mechanism to achieve both differential privacy and the *perfect* fairness, while maintaining a sufficient level of accuracy. In the next section, we study the relation between fairness γ, privacy ϵ, and accuracy θ under $\mathscr{A}_\epsilon(\cdot)$.

[2]Note that if algorithm $\mathscr{M}(\cdot)$ selects an individual solely based on i.i.d qualification scores $(R_1,\ldots,R_n)$ and does not differentiate individuals based on their indexes, then $\Pr\{K_i=1|A_i=0,Y_i=1\} - \Pr\{K_i=1|A_i=1,Y_i=1\} = \Pr\{K_j=1|A_j=0,Y_j=1\} - \Pr\{K_j=1|A_j=1,Y_j=1\}, \forall i,j$.

3 Analysis

3.1 Fairness-Privacy Trade-off

To study the relation between the fairness and privacy, let $\gamma(\epsilon) = E\{Z_{i,\epsilon}|A_i=0,Y_i=1\} - E\{Z_{i,\epsilon}|A_i=1,Y_i=1\}$. Note that $\gamma(\epsilon)$ does not depend on i as individuals are i.i.d. Let $\mathscr{A}(\cdot)$ be the algorithm that selects an individual with the highest qualification score and breaks ties randomly and uniformly if more than one individual have the highest qualification score. Define a set of Bernoulli random variables $\{I_i\}_{i=1}^n$ indicating whether individual i is selected $(I_i=1)$ or not $(I_i=0)$ under algorithm $\mathscr{A}(\cdot)$. Let $N_{\max} = |\{i \in \mathcal{N} | R_i = \max_j R_j\}|$ be the number of individuals who have the highest qualification score, then

$$\Pr\{I_i=1\} = \sum_{k=1}^{n} \frac{1}{k} \cdot \Pr\{R_i = \max_j R_j, N_{\max}=k\}. \quad (3)$$

Define

$$Z_i = \begin{cases} 0 & \text{if } R_i \neq \max_j R_j \\ \frac{1}{N_{\max}} & o.w. \end{cases}. \quad (4)$$

Then it holds that $E\{Z_i\} = \Pr\{I_i=1\}$. The following lemma characterizes the relation between random variables Z_i and $Z_{i,\epsilon}$, which is essential to prove the next theorem.

Lemma 1 (Sure convergence)**.** *Consider two algorithms $\mathscr{A}_\epsilon(\cdot)$ and $\mathscr{A}(\cdot)$ and the corresponding random variables $Z_{i,\epsilon}$ and Z_i. The following statements are true,*

1. *$Z_{i,\epsilon}$ converges surely towards Z_i as $\epsilon \to +\infty$.*
2. *$\mathscr{A}(\cdot)$ is γ_∞-fair with $\gamma_\infty = \lim_{\epsilon\to+\infty}\gamma(\epsilon)$, i.e.,*

$$\begin{aligned} & \lim_{\epsilon\to+\infty} E\{Z_{i,\epsilon}|A_i=0,Y_i=1\} - E\{Z_{i,\epsilon}|A_i=1,Y_i=1\} \\ = \; & E\left\{\lim_{\epsilon\to+\infty} Z_{i,\epsilon}|A_i=0,Y_i=1\right\} - E\left\{\lim_{\epsilon\to+\infty} Z_{i,\epsilon}|A_i=1,Y_i=1\right\} \\ = \; & E\{Z_i|A_i=0,Y_i=1\} - E\{Z_i|A_i=1,Y_i=1\}. \end{aligned}$$

Lemma 1 implies that $\lim_{\epsilon\to+\infty}\gamma(\epsilon) = \gamma_\infty$ exists. It shows that $\mathscr{A}_\epsilon(\cdot)$ using exponential mechanism is equivalent to algorithm $\mathscr{A}(\cdot)$ as $\epsilon \to \infty$. In the next theorem, we identify a sufficient condition under which the exponential mechanism can achieve *perfect* fairness with non-zero privacy leakage.

Theorem 1. *There exists $\epsilon_o > 0$ such that $\gamma(\epsilon_o) = 0$ under $\mathscr{A}_{\epsilon_o}(.)$ if both of the following constraints are satisfied:*

(1) $E\{Z_i|A_i=a,Y_i=1\} < E\{Z_i|A_i=\neg a,Y_i=1\}$,
(2) $E\{R_i|A_i=a,Y_i=1\} > E\{R_i|A_i=\neg a,Y_i=1\}$,

where $a \in \{0,1\}$ and $\neg a = \{0,1\} \setminus a$.

Constraint (1) above suggests that the applicants with protected attribute $A=\neg a$ are more likely to be selected than those with $A=a$ under algorithm $\mathscr{A}(\cdot)$; Constraint (2) implies that on average the applicants with protected attribute $A=a$ have a higher qualification score than those with $A=\neg a$. These constraints may be satisfied when the applicants with $A=\neg a$, as compared to those with $A=a$, have the smaller mean but much larger variance in their qualification cores. In Sec. 5, we will show FICO dataset satisfies those constraints for certain social groups.

It is also worth noting that perfect fairness is not always attainable under the exponential mechanism. In the next theorem, we identify sufficient conditions under which it is impossible to achieve the *perfect* fairness using exponential mechanism unless the privacy guarantee is trivial ($\epsilon = 0$).

Theorem 2. *Let $f^a(\rho) \coloneqq \Pr\{R = \rho | A = a, Y = 1\}$. If $f^0(\rho) - f^1(\rho) > f^0(\rho') - f^1(\rho')$ and $f_R(\rho) < f_R(\rho')$ for all $\rho > \rho'$, and $f^0(\rho) - f^1(\rho) \geq 0$ for $\rho = \rho_2, \ldots, \rho_{n'}$, then we have,*
1. $\gamma(\epsilon) > 0$ for $\epsilon > 0$, i.e., $\mathscr{A}_\epsilon(\cdot)$ is always biased in favor of individuals with protected attribute $A = 0$.
2. $\gamma(\epsilon) < \gamma_\infty, \forall \epsilon \geq 0$, i.e., $\mathscr{A}_\epsilon(\cdot)$ is always fairer than $\mathscr{A}(\cdot)$.

The first condition implies that among applicants who are qualified, individuals with $A = 0$ are more likely to have higher qualification scores as compared to individuals with protected attribute $A = 1$. The second condition implies that most of the applicants have small qualification scores. Under these conditions, an exponential mechanism with $\epsilon > 0$ can never achieve perfect fairness. Moreover, it shows that the exponential mechanism can improve fairness compared to the non-private algorithm $\mathscr{A}(\cdot)$ selecting an individual with the highest score.

Theorems 1 and 2 together show that the exponential mechanism may or may not achieve *perfect* fairness. Nevertheless, we can show that always there exists privacy parameter $\bar{\epsilon}$ such that $\mathscr{A}_{\bar{\epsilon}}(\cdot)$ is fairer than non-private $\mathscr{A}(\cdot)$.

Theorem 3. *If $\mathscr{A}(\cdot)$ is not 0-fair, then there exists $\hat{\epsilon} \in (0, +\infty)$ such that $|\gamma(\epsilon)| < |\gamma_\infty|, \forall \epsilon \in (0, \hat{\epsilon})$.*

This section has studied the possibility of using exponential mechanism to improve both fairness and privacy. Note that even when perfect fairness is attainable, the outcome may not be desirable to the decision-maker if it is not accurate enough. In the next section, we shall take accuracy into account and examine its relation with privacy and fairness.

3.2 Accuracy-Privacy Trade-off

Let $\theta(\epsilon) = \Pr\{Y_{\mathscr{A}_\epsilon(\mathbf{D})} = 1\}$ be the accuracy of $\mathscr{A}_\epsilon(\cdot)$. We have, $\theta(\epsilon) = \sum_{i=1}^n \Pr\{Y_i = 1, I_{i,\epsilon} = 1\} = \sum_{i=1}^n \Pr\{I_{i,\epsilon} = 1 | Y_i = 1\} \cdot \Pr\{Y_i = 1\} = \Pr\{Y = 1\} \sum_{i=1}^n E\{Z_{i,\epsilon} | Y_i = 1\}$. Therefore, maximizing accuracy equals to maximizing $E\{Z_{i,\epsilon} | Y_i = 1\}$. Since $Z_{i,\epsilon}$ converges surely to Z_i, similar to Lemma 1, we can show that $\lim_{\epsilon \to +\infty} \theta(\epsilon)$ exists and is equal to the accuracy of non-private algorithm $\mathscr{A}(\cdot)$. We further make the following assumption that has been widely used in literature (Jung et al. 2020; Barman and Rathi 2020).

Assumption 1. $\frac{\Pr\{R=\rho|Y=1\}}{\Pr\{R=\rho|Y=0\}} \geq \frac{\Pr\{R=\rho'|Y=1\}}{\Pr\{R=\rho'|Y=0\}}, \forall \rho > \rho'$.

Assumption 1, also known as the monotone likelihood ratio property of two PMFs $\Pr\{R = \rho | Y = 1\}$ and $\Pr\{R = \rho | Y = 0\}$, is relatively mild and can be satisfied by various probability distributions including Binomial and Poisson distributions. It implies that a qualified individual is more likely to have a high qualification score. The next theorem characterizes the effect of privacy parameter ϵ on the accuracy of $\mathscr{A}_\epsilon(\cdot)$.

Theorem 4. *Under Assumption 1, $\theta(\epsilon)$ is increasing in ϵ.*

Suppose that the task is to make a selection such that unfairness and privacy leakage are less than or equal to $\gamma_{\max}$ and $\epsilon_{\max}$, respectively. Then, exponential mechanism $\mathscr{A}_{\epsilon^*}(\cdot)$ has the highest accuracy, where ϵ^* is the solution to (5).

$$\epsilon^* = \arg \max_{\epsilon \leq \epsilon_{\max}} \theta(\epsilon), \quad s.t. \ |\gamma(\epsilon)| \leq \gamma_{\max}. \tag{5}$$

Based on Theorem 4, we have the following corollary.

Corollary 1. *Under Assumption 1,*

$$\epsilon^* = \begin{cases} \epsilon_{\max}, & \text{if } |\gamma(\epsilon_{\max})| \leq \gamma_{\max} \\ \max\{\epsilon \leq \epsilon_{\max} | \gamma_{\max} = |\gamma(\epsilon)|\}, & o.w. \end{cases}$$

We conclude this section by comparing our results with (Cummings et al. 2019). Consider a constant algorithm that always selects the first individual. The accuracy of this algorithm is given by $\Pr(Y_1 = 1)$, which is equal to the accuracy of algorithm $\mathscr{A}_0(\cdot)$, i.e., $\Pr(Y = 1) = \theta(0)$. Moreover, the constant algorithm is perfectly fair. If there exists $\epsilon_o > 0$ such that $\mathscr{A}_{\epsilon_o}(\cdot)$ is perfectly fair, then according to Theorem 4, the accuracy of $\mathscr{A}_{\epsilon_o}(.)$ would be larger than that of $\mathscr{A}_0(\cdot)$, implying that the perfect fair $\mathscr{A}_{\epsilon_o}(\cdot)$ is also more accurate than the constant algorithm. In contrast, Cummings *et al.* in (Cummings et al. 2019) conclude that perfect (exact) fairness is not compatible with differential privacy. Specifically, they show that any differentially private classifier which is perfectly fair would have lower accuracy than a constant classifier. The reason that our conclusion differs from (Cummings et al. 2019) is as follows. (Cummings et al. 2019) studies a classification problem where there is a hypothesis class $\mathcal{H}$ (i.e., a set of possible classifiers) and it aims to select a *perfect* fair classifier randomly from $\mathcal{H}$ with high accuracy using a differentially private algorithm. In particular, they show that if h is perfectly fair and more accurate than a constant classifier under database D, then there exists another database D' with $D' \sim D$ such that h violates perfect fairness under D'. Because h is selected with zero probability under D', differential privacy is violated, which implies their negative results. In contrast, we focus on a selection problem with a fixed number of approvals using *fixed* supervised learning model $r(\cdot)$. In this case, privacy and perfect fairness can be compatible with each other.

4 Choosing More Than One Applicant

Our results so far are concluded under the assumption that only one applicant is selected. In this section, we extend our results to a scenario where $m > 1$ applicants are selected.

To preserve individual privacy, an exponential mechanism is adopted to select m individuals from n applicants based on their qualification scores. Let $\mathcal{S} = \{\mathcal{G} | \mathcal{G} \subseteq \mathcal{N}, |\mathcal{G}| = m\}$ be the set of all possible selections, and we index the elements in $\mathcal{S}$ by $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_{\binom{n}{m}}$. Let $\mathscr{B}_\epsilon(\cdot)$ be the exponential mechanism that selects m individuals from $\mathcal{N}$ and satisfies ϵ-differential privacy. One choice of score function $v : \mathcal{S} \times \mathcal{D} \to [0, 1]$ is $v(\mathcal{G}_i, \mathbf{D}) = \frac{1}{m} \sum_{j \in \mathcal{G}_i} R_j$, representing the averaged qualification of selected individuals in $\mathcal{G}_i$.[3] The sensitivity of $v(\cdot, \cdot)$ is $\max_{\mathcal{G} \in \mathcal{S}, D \sim D'} |v(\mathcal{G}, D) -$

[3]The generalization of our results to other types of score function will be discussed in the appendix.

$v(\mathcal{G}, D')| = \frac{1}{m}$. That is, under algorithm $\mathscr{B}_\epsilon(\cdot)$, $\mathcal{G}_i$ is selected according to probability

$$\Pr\{\mathscr{B}_\epsilon(D) = \mathcal{G}_i\} = \frac{\exp\{\epsilon \cdot \frac{\sum_{j\in\mathcal{G}_i} r_j}{2}\}}{\sum_{\mathcal{G}\in\mathcal{S}} \exp\{\epsilon \cdot \frac{\sum_{j\in\mathcal{G}} r_j}{2}\}}. \quad (6)$$

Further define $\mathcal{S}_i = \{\mathcal{G}|\mathcal{G} \in \mathcal{S}, i \in \mathcal{G}\}$ as the set of all selections that contain individual i. Define random variable

$$W_{i,\epsilon} = \sum_{\mathcal{G}\in\mathcal{S}_i} \frac{\exp\{\epsilon \cdot \frac{\sum_{j\in\mathcal{G}} R_j}{2}\}}{\sum_{\mathcal{G}'\in\mathcal{S}} \exp\{\epsilon \cdot \frac{\sum_{j\in\mathcal{G}'} R_j}{2}\}} \quad (7)$$

and Bernoulli random variable $J_{i,\epsilon}$ indicating whether i is selected ($J_{i,\epsilon} = 1$) under $\mathscr{B}_\epsilon(\cdot)$ or not ($J_{i,\epsilon} = 0$). We have $\Pr\{J_{i,\epsilon} = 1\} = E\{W_{i,\epsilon}\}$. Similar to Section 2, we further introduce algorithm $\mathscr{B}(\cdot)$ which selects a set of m individuals with the highest average qualification score. If there are more than one set with the highest average qualification score, $\mathscr{B}(\cdot)$ selects one set among them uniformly at random. Let $\mathcal{S}_{\max} = \{\mathcal{G}' \in \mathcal{S} | \sum_{j\in\mathcal{G}'} R_j = \max_{\mathcal{G}\in\mathcal{S}} \sum_{i\in\mathcal{G}} R_i\}$. Each element in $\mathcal{S}_{\max}$ is a set of m individuals who have the highest qualification scores in total. Define random variable

$$W_i = \begin{cases} 0 & \text{if } \mathcal{S}_i \cap \mathcal{S}_{\max} = \emptyset \\ \frac{1}{|\mathcal{S}_{\max}|} & o.w. \end{cases}$$

and Bernoulli random variable J_i indicating whether i is selected ($J_i = 1$) under $\mathscr{B}(D)$ or not ($J_i = 0$). We have $\Pr\{J_i = 1\} = E\{W_i\}$. Similar to the fairness metric defined in Definition 4, we say algorithm $\mathscr{B}_\epsilon(\cdot)$ is γ-fair if the following holds,

$$E\{W_{i,\epsilon}|A_i = 0, Y_i = 1\} - E\{W_{i,\epsilon}|A_i = 1, Y_i = 1\} = \gamma\ .$$

Re-write γ above as $\gamma(\epsilon)$, a function of ϵ. Similar to Lemma 1, we can show that $W_{i,\epsilon}$ converges surely to W_i as $\epsilon \to +\infty$, and that $\lim_{\epsilon\to+\infty} \gamma(\epsilon)$ exists. Moreover, $\mathscr{B}(D)$ is γ_∞-fair with $\gamma_\infty = \lim_{\epsilon\to+\infty} \gamma(\epsilon)$. Next theorem identifies sufficient conditions under which exponential mechanism $\mathscr{B}_\epsilon(.)$ can be perfectly fair.

Theorem 5. *There exists $\epsilon_o > 0$ such that $\gamma(\epsilon_o) = 0$ under $\mathscr{B}_\epsilon(.)$ if both of the following constraints are satisfied:*

(1) $E\{W_i|A_i = a, Y_i = 1\} < E\{W_i|A_i = \neg a, Y_i = 1\}$;

(2) $E\{R_i|A_i = a, Y_i = 1\} > E\{R_i|A_i = \neg a, Y_i = 1\}$.

To measure the accuracy in this scenario, we adjust Definition 5 accordingly. For each individual i, we define random variable $U_i = \begin{cases} 1, & \text{if } i \in \mathscr{B}_\epsilon(\mathbf{D}) \ \&\ Y_i = 1 \\ 0, & o.w. \end{cases}$ as the utility gained by the decision-maker from individual i, i.e., the decision-maker receives benefit $+1$ if accepting a qualified applicant and 0 otherwise. Then the accuracy of $\mathscr{B}_\epsilon(\cdot)$ can be defined as the expected utility received by decision-maker, i.e.,

$$\theta(\epsilon) = E\Big\{\frac{1}{m}\sum_{i\in\mathcal{N}} U_i\Big\} = \frac{1}{m}\sum_{i\in\mathcal{N}} \Pr\{J_{i,\epsilon} = 1, Y_i = 1\}. \quad (8)$$

Note that $(1/m)$ in (8) is a normalization factor to make sure $\theta(\epsilon) \in [0, 1]$. Also, it is worth mentioning that $\theta(\epsilon)$ reduces to Definition 5 when $m = 1$. Under Assumption 1, we can show that $\theta(\epsilon)$ is increasing in ϵ, and an optimization problem similar to optimization (5) can be formulated given $\epsilon_{\max}$ and $\gamma_{\max}$ to find appropriate ϵ^*.

		ϵ_o	$\theta(\epsilon_o)$	$\theta(\infty)$	Acc. Red.
Synthetic	$m = 1$	2.76	0.50	0.96	47.92%
	$m = 2$	7.78	0.64	0.88	27.27%
	$m = 3$	21.11	0.73	0.77	5.19%
	$m = 4$	0	0.4	0.71	44.66%
FICO Fig. 8 & 9	$m = 1$	10.35	0.94	0.97	3.09%
	$m = 2$	22.47	0.94	0.95	1.05%
	$m = 3$	0	0.70	0.93	24.73%
	$m = 4$	0	0.70	0.90	22.22%

Table 1: Accuracy and privacy under perfect fairness. ϵ_o is a privacy parameter at which the exponential mechanism is perfectly fair. If $\epsilon_o = 0$ in this table, then the exponential mechanism with a non-zero privacy parameter cannot achieve perfect fairness.

5 Numerical Experiments

Case study 1: Synthetic data. To evaluate the fairness of algorithms $\mathscr{A}_\epsilon(\cdot)$ and $\mathscr{B}_\epsilon(\cdot)$, we consider a scenario where the qualification scores are generated randomly based on a distribution shown in Figure 1. In this scenario, $n = 10$, $\Pr\{A = 0\} = 0.1, \Pr\{Y = 0|A = 0\} = 0.7$, and $\Pr\{Y = 0|A = 1\} = 0.6$. Fig. 2 illustrates the fairness level $\gamma(\epsilon)$ of algorithms $\mathscr{A}_\epsilon(\cdot)$ and $\mathscr{B}_\epsilon(\cdot)$ as a function of ϵ. In this case, both conditions in Theorem 1 and Theorem 5 are satisfied when $m \in \{1, 2, 3\}$ (See the appendix for details). As a result, we can find privacy parameter ϵ_o at which the exponential mechanism is perfectly fair. Note that conditions of Theorem 5 do not hold for $m = 4$ (see the appendix) and $\mathscr{B}_\epsilon(\cdot)$ is not perfectly fair in this case. Fig. 3 illustrates accuracy of $\mathscr{A}_\epsilon(\cdot)$ and $\mathscr{B}_\epsilon(\cdot)$ as a function of privacy loss ϵ. As expected, accuracy $\theta(\epsilon)$ is increasing in ϵ. By comparing Fig. 2 and Fig. 3, we observe that even though improving privacy decreases accuracy, it can improve fairness. Lastly, privacy and accuracy of the exponential mechanism under perfect fairness have been provided in Table 1.

Case study 2: FICO score. We conduct two experiments using FICO credit score dataset.[4] FICO scores are widely used in the United States to predict how likely an applicant is to pay back a loan. FICO scores are ranging from 350 to 850. However, we normalize them to range from zero to one. The FICO credit score dataset has been processed by Hardt *et al.* (Hardt, Price, and Srebro 2016) to generate CDF and non-default rate (i.e., $\Pr(Y = 1|R = \rho)$) for different social groups (Asian, White, Hispanic, and Black).

First, we consider a setting where individuals from White and Black groups are selected based on their FICO scores using the exponential mechanism. Figure 4 illustrates the PMF of FICO score for White and Black groups. It shows PMF is (approximately) decreasing for Black group while is (approximately) increasing for White group. Moreover, the overall PMF for two groups is (approximately) uniform and remains constant. As shown in Fig. 5 and 6, both accuracy $\theta(\epsilon)$ and fairness $\gamma(\epsilon)$ are increasing in ϵ. Therefore, both algorithm $\mathscr{A}_\epsilon(\cdot)$ and $\mathscr{B}_\epsilon(\cdot)$ cannot be perfectly fair, and the

[4]Find the dataset here: https://bit.ly/3di5NOC

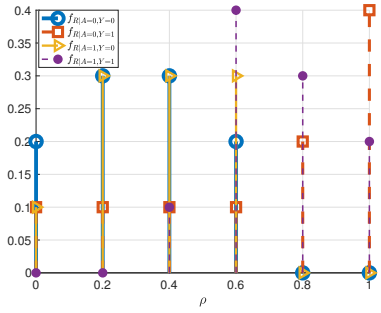

Fig. 1: PMF of score R conditional on Y and A.

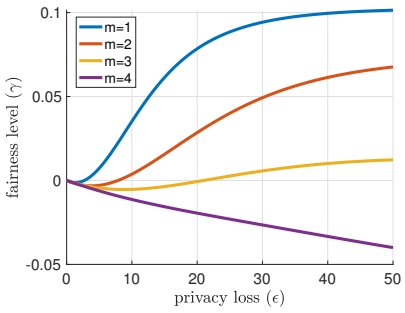

Fig. 2: Fairness attained under $\mathscr{A}_\epsilon(\cdot)$ and $\mathscr{B}_\epsilon(\cdot)$ as a function of privacy ϵ.

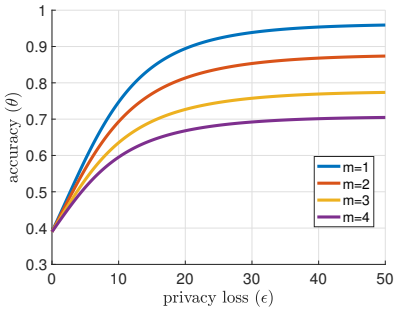

Fig. 3: Accuracy attained under $\mathscr{A}_\epsilon(\cdot)$ and $\mathscr{B}_\epsilon(\cdot)$ as a function of privacy ϵ.

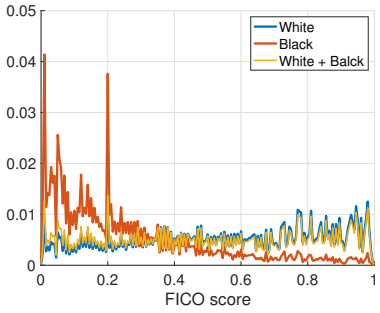

Fig. 4: PMF of FICO score for *Black* and *White* social groups.

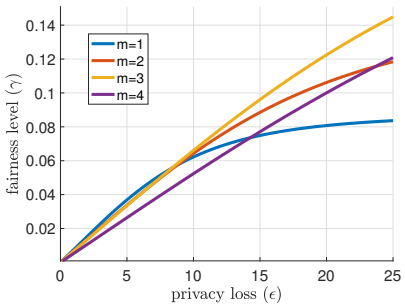

Fig. 5: Fairness $\gamma(\epsilon)$ when m people are selected from *White* and *Black*.

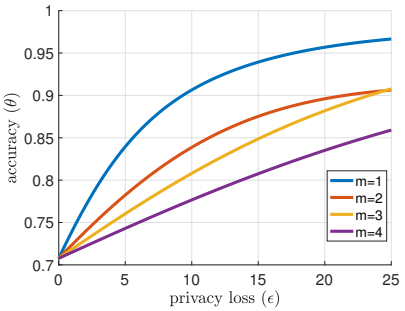

Fig. 6: Accuracy $\theta(\epsilon)$ when m people are selected from *White* and *Black*.

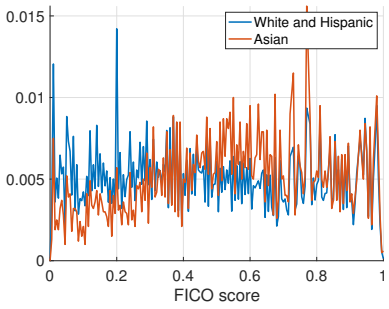

Fig. 7: PMF of FICO score for *White-Hispanic* and *Asian* groups.

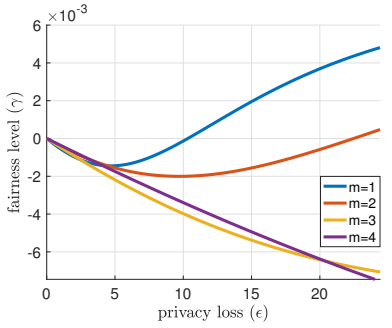

Fig. 8: Fairness when m people are selected from *White-Hispanic* & *Asian*.

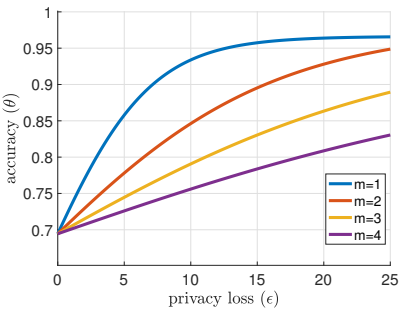

Fig. 9: Accuracy when m people are selected from *White-Hispanic* & *Asian*.

conditions in Theorem 1 and 5 do not hold in this example.

In the next experiment, we combine White and Hispanic applicants into one group and regard Asian applicants as the other social group. Fig. 7 illustrates the PMF of FICO score for these two groups. In this example, the conditions of Theorem 1 and Theorem 5 are satisfied for $m \in \{1, 2\}$. When $m \in \{1, 2\}$, perfect fairness is achievable for some $\epsilon_o > 0$ and leads to a slight decrease in accuracy compared to non-private algorithms $\mathscr{A}(\cdot)$ and $\mathscr{B}(\cdot)$ (see Table 1).

6 Conclusion

In this paper, we consider a common scenario in job/loan applications where a decision-maker selects a limited number of people from an applicant pool based on their qualification scores. These scores are generated by a pre-trained supervised learning model, which may be biased against certain social groups, and the use of such a model may violate applicants' privacy. Within this context, we investigated the possibility of using exponential mechanism to address both privacy and unfairness issues. We show that this mechanism can be used as a post-processing step to improve fairness and privacy of the pre-trained model. Moreover, we identified conditions under which the exponential mechanism is able to make the selection procedure *perfectly* fair.

References

Agarwal, A.; Beygelzimer, A.; Dudik, M.; Langford, J.; and Wallach, H. 2018. A Reductions Approach to Fair Classification. In *International Conference on Machine Learning*, 60–69.

Awasthi, P.; Kleindessner, M.; and Morgenstern, J. 2020. Equalized odds postprocessing under imperfect group information. In *International Conference on Artificial Intelligence and Statistics*, 1770–1780. PMLR.

Barman, S.; and Rathi, N. 2020. Fair Cake Division Under Monotone Likelihood Ratios. In *Proceedings of the 21st ACM Conference on Economics and Computation*, EC '20, 401437. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379755. doi:10.1145/3391403.3399512. URL https://doi.org/10.1145/3391403.3399512.

Biega, A. J.; Gummadi, K. P.; and Weikum, G. 2018. Equity of attention: Amortizing individual fairness in rankings. In *The 41st international acm sigir conference on research & development in information retrieval*, 405–414.

Conitzer, V.; Freeman, R.; Shah, N.; and Vaughan, J. W. 2019. Group fairness for the allocation of indivisible goods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 1853–1860.

Cummings, R.; Gupta, V.; Kimpara, D.; and Morgenstern, J. 2019. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, 309–315.

Dastin, J. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Dressel, J.; and Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4(1): eaao5580.

Dwork, C. 2006. Differential privacy. In *Proceedings of the 33rd international conference on Automata, Languages and Programming-Volume Part II*, 1–12. Springer-Verlag.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, 265–284. Springer.

Gupta, S.; and Kamble, V. 2019. Individual fairness in hindsight. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, 805–806.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 3315–3323.

Harwell, D. 2018. The accent gap. URL https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/?utm_term=.ca17667575d1. Accessed: 2018-01-07.

Jagielski, M.; Kearns, M.; Mao, J.; Oprea, A.; Roth, A.; Sharifi-Malvajerdi, S.; and Ullman, J. 2019. Differentially private fair learning. In *International Conference on Machine Learning*, 3000–3008. PMLR.

Jiang, H.; and Nachum, O. 2019. Identifying and correcting label bias in machine learning. *arXiv preprint arXiv:1901.04966* .

Jung, C.; Kannan, S.; Lee, C.; Pai, M. M.; Roth, A.; and Vohra, R. 2020. Fair prediction with endogenous behavior. *arXiv preprint arXiv:2002.07147* .

Jung, C.; Kearns, M.; Neel, S.; Roth, A.; Stapleton, L.; and Wu, Z. S. 2019. Eliciting and enforcing subjective individual fairness. *arXiv preprint arXiv:1905.10660* .

Kallus, N.; Mao, X.; and Zhou, A. 2019. Assessing algorithmic fairness with unobserved protected class using data combination. *arXiv preprint arXiv:1906.00285* .

Kamiran, F.; and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33(1): 1–33.

Kleinberg, J.; and Raghavan, M. 2018. Selection Problems in the Presence of Implicit Bias. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*, volume 94, 33. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Liu, X.; Xie, Q.; and Wang, L. 2017. Personalized extended (α, k)-anonymity model for privacy-preserving data publishing. *Concurrency and Computation: Practice and Experience* 29(6): e3886.

McSherry, F.; and Talwar, K. 2007. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, 94–103. IEEE.

Mozannar, H.; Ohannessian, M. I.; and Srebro, N. 2020. Fair Learning with Private Demographic Data. *arXiv preprint arXiv:2002.11651* .

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On Fairness and Calibration. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*, 5680–5689. Curran Associates, Inc. URL http://papers.nips.cc/paper/7151-on-fairness-and-calibration.pdf.

Sweeney, L. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10(05): 557–570.

Wang, S.; Cui, L.; Que, J.; Choi, D.-H.; Jiang, X.; Cheng, S.; and Xie, L. 2012. A randomized response model for privacy preserving smart metering. *IEEE transactions on smart grid* 3(3): 1317–1324.

Wang, S.; Guo, W.; Narasimhan, H.; Cotter, A.; Gupta, M.; and Jordan, M. I. 2020. Robust Optimization for Fairness with Noisy Protected Groups. *arXiv preprint arXiv:2002.09343* .

Xu, D.; Yuan, S.; and Wu, X. 2019. Achieving differential privacy and fairness in logistic regression. In *Companion Proceedings of The 2019 World Wide Web Conference*, 594–599.

Zafar, M. B.; Valera, I.; Gomez-Rodriguez, M.; and Gummadi, K. P. 2019. Fairness Constraints: A Flexible Approach for Fair Classification. *Journal of Machine Learning Research* 20(75): 1–42.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In *International Conference on Machine Learning*, 325–333.

Zhang, X.; Khalili, M. M.; and Liu, M. 2020. Long-term impacts of fair machine learning. *Ergonomics in Design* 28(3): 7–11.

Zhang, X.; Khalili, M. M.; Tekin, C.; and Liu, M. 2019. Group Retention when Using Machine Learning in Sequential Decision Making: the Interplay between User Dynamics and Fairness. In *Advances in Neural Information Processing Systems*, 15243–15252.

Zhang, X.; and Liu, M. 2020. Fairness in Learning-Based Sequential Decision Algorithms: A Survey. *arXiv preprint arXiv:2001.04861* .

Zhang, X.; Tu, R.; Liu, Y.; Liu, M.; Kjellstrom, H.; Zhang, K.; and Zhang, C. 2020. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems* 33.

Zhu, Y.; and Liu, L. 2004. Optimal randomization for privacy preserving data mining. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 761–766.