Measuring Fairness of Rankings under Noisy Sensitive Information

Azin Ghazimatin*
Spotify
Berlin, Germany
azing@spotify.com

Matthäus Kleindessner
Amazon Web Services
Tübingen, Germany
matkle@amazon.de

Chris Russell
Amazon Web Services
Tübingen, Germany
cmruss@amazon.de

Ziawasch Abedjan
Leibniz Universität Hannover and
Amazon Search
Hannover, Germany
abedjan@dbs.uni-hannover.de

Jacek Golebiowski
Amazon Search
Berlin, Germany
jacekgo@amazon.de

ABSTRACT

Metrics commonly used to assess group fairness in ranking require the knowledge of group membership labels (e.g., whether a job applicant is male or female). Obtaining accurate group membership labels, however, may be costly, operationally difficult, or even infeasible. Where it is not possible to obtain these labels, one common solution is to use proxy labels in their place, which are typically predicted by machine learning models. Proxy labels are susceptible to systematic biases, and using them for fairness estimation can thus lead to unreliable assessments. We investigate the problem of measuring group fairness in ranking for a suite of divergence-based metrics in the presence of proxy labels. We show that under certain assumptions, fairness of a ranking can reliably be measured from the proxy labels. We formalize two assumptions and provide a theoretical analysis for each showing how the true metric values can be derived from the estimates based on proxy labels. We prove that without such assumptions fairness assessment based on proxy labels is impossible. Through extensive experiments on both synthetic and real datasets, we demonstrate the effectiveness of our proposed methods for recovering reliable fairness assessments in rankings.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; • **Computing methodologies** → **Machine learning**.

ACM Reference Format:
Azin Ghazimatin, Matthäus Kleindessner, Chris Russell, Ziawasch Abedjan, and Jacek Golebiowski. 2022. Measuring Fairness of Rankings under Noisy Sensitive Information. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3531146.3534641

*Work done while the author was an intern at Amazon Search.

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3534641

1 INTRODUCTION

Motivation. Ranking and retrieval systems aim to maximize the notion of relevance when retrieving objects for their typical users. Evidence suggests that these systems can introduce biases against certain groups of objects [23, 39]. One example is the unequal representation of different genders in image search results for a range of occupations [34]. The increasing concerns over the societal impact of the information retrieval (IR) systems have driven discussions over the importance of measuring and mitigating the bias of ranking algorithms [8, 9, 15, 36, 48, 54, 60, 63], and in this spirit, fairness has become one of the key criteria for evaluating ranking systems.[1]

To enhance fairness of a ranking with respect to certain demographic groups distinguished based on characteristics such as ethnicities or gender, we must first quantify the bias of the ranking. This can be done using one of the recently proposed fairness metrics [35, 54, 62]. Computing these metrics requires the knowledge of group membership, i.e., the ability to tell which demographic group an individual belongs to. Collecting the membership labels, however, may be costly, operationally challenging, or even illegal [12, 57].[2] In this case, the group membership labels may be inferred using machine learning models [19, 41] such as the ones used for missing data imputation [10, 11]. The estimated labels are usually referred to as proxy labels and the models that predict the labels based on observable variables are referred to as proxy models [16]. A common example of a proxy model is the Bayesian Improved Surname Geocoding method (BISG), which is often used in the domains of health, finance, and politics to predict race based on an individual's surname [3, 19, 29].

Unfortunately, evaluating fairness using proxy labels can lead to unreliable assessments. This problem has been shown in the context of classification [2, 3, 16, 20, 31], but it is also conceivable in the context of ranking: consider a recruiting tool that returns the best applicants for a given job. As an example output, suppose that the system returns four people with the following ordering: $Bob \succ Charles \succ Alicia \succ Clair$. To evaluate the group fairness of this ranking with respect to gender, we assume the population can be divided into two groups of "male" and "female" applicants.[3]

[1] We use the terms *bias measurement* and *fairness measurement* interchangeably.
[2] E.g.,the Equal Credit Opportunity Act prohibits a creditor from inquiring about an applicant's race, religion, etc. [43].
[3] We note that gender is not binary and our example is highly simplified.

Now, imagine a proxy model that correctly predicts the gender of *Bob* and *Alicia* as "male" and "female", respectively, but wrongly assigns the label "female" to *Charles* and the label "male" to *Clair*. Using these proxy labels to compute divergence-based metrics like pairwise demographic parity [8, 38, 49] (see Section 2 for the formal definition), we judge the ranking to be fair, as male and female applicants appear to be distributed evenly across the ranking. This conclusion, however, is wrong, as according to the actual gender labels, women are all ranked below men. This shows that fairness assessments obtained from proxy labels can be misleading, and hence can prevent us from taking proper actions in face of biased treatments.

Approach. In this work, we investigate the problem of group fairness assessment of rankings based on proxy labels. While the analogous problem has already been studied for classification tasks [2, 16, 31], there is no work providing solutions for reliable bias measurement in rankings when proxy labels are used. Note that we differentiate our study from those that assume other types of missing information (such as ground-truth scores [56] or the outcome of unadministered interventions [32]) and the ones that attempt to optimize fairness of classifiers or selected subsets when group membership labels are unknown or noisy [26, 47, 58].

We begin by outlining the challenges of this problem and show that bias estimation of rankings using proxy labels is impossible unless we make certain assumptions. These assumptions encode the dependency relationships between the random variables modeling the true group membership labels, their estimated values (i.e., the proxy labels), and the ordering of items in the ranking. We present a range of these assumptions, discuss their implications on the feasibility of bias estimation and specify the two feasible ones. We provide theoretical analysis for a suite of fairness metrics in ranking, showing that under each of the specified assumptions, it is possible to recover the true fairness value from the proxy measurement computed using the proxy labels. Each assumption identifies a specific type of conditional independence between the variables representing the underlying data model, which helps us derive a tractable relationship between the proxy measurement and the true fairness. These relationships are identified using aggregate statistics about the proxy model and can be used to recover the true measures from the proxy values in rankings where the specified assumption holds.

Contributions. The contributions of this work are as follows:

- We provide theoretical foundations for fairness measurement of rankings based on proxy group membership labels. We prove that without making any assumptions about the underlying data model, fairness measurement is impossible. We further specify two assumptions under which fairness measurement becomes feasible for a suite of divergence-based metrics.
- We provide theoretical analysis showing how the true fairness measures can be derived from the proxy values under the feasible assumptions.
- Through extensive experiments on both synthetic and real datasets, we demonstrate the effectiveness of the proposed methods for recovering the true fairness assessments in practice.

2 PROBLEM SETUP AND FAIRNESS METRICS

In this section, we present the problem setup and the fairness metrics considered in our study. Our main concern is to evaluate fairness of ranking processes, which we assume to be of the following form: there is a distribution $Q(S, A)$ over random variables $S \in \mathbb{R}$ and $A \in \{0, 1\}$, where S is an individual's ranking score and A denotes their sensitive attribute, such as their gender or race, according to which the demographic groups are defined (hence, A is also referred to as the group membership label). For simplicity and ease of presentation, we consider a binary-valued sensitive attribute A, and we refer to the individuals with $A = 0$ as the group G_0 and the ones with $A = 1$ as G_1. When generating a ranking $i_1 < i_2 < \ldots < i_n$ of n individuals, n independent and identically distributed (i.i.d.) samples $(S_1, A_1), \ldots, (S_n, A_n)$ are drawn from $Q(S, A)$ and ranked according to S_i; that is, we obtain the ranking $i_1 < i_2 < \ldots < i_n$ if $S_{i_1} < S_{i_2} < \ldots < S_{i_n}$. We assume that all S_i are different. This assumption holds with probability one if the marginal distribution over S has continuous support; otherwise, we can incorporate random tie-breaking into the process by considering slightly perturbed scores. Note that our formulation is fairly general and captures numerous examples of ranking processes where fairness is of concern (e.g., ranking university applicants by opaque scores [44], ranking job candidates by predicted job performance [27] or ranking houses by their estimated selling price [1]).

We measure fairness using group fairness notions that aim to ensure that none of the two groups G_0 or G_1 is disadvantaged. Intuitively, the notions that we declare a ranking to be fair if the individuals' ranks do not depend on their group membership. We do not consider notions such as xAUC [33] or Pairwise Equal Opportunity [49], where a certain "ground-truth" ranking of items is assumed and the ranking's error (i.e., the misalignment of the ground-truth and the observed ranking) is supposed to be independent of the sensitive attribute. Next, we present the formal definitions of the metrics considered in this work.

Fairness Metrics. In this paper, we focus on fairness metrics that measure the under-representation of certain groups among the top positions. These metrics are sometimes referred to as divergence-based metrics [35] since they measure the divergence of per-group statistics from a target value. They usually come in two flavors: pairwise and listwise. In the following, we present the general form of each kind along with some concrete examples.

- **Pairwise divergence-based metrics**: These metrics have the general form of

$$\begin{aligned}&\mathcal{M}_1 = \mathcal{M}_1(Q(S, A); n) = \\ &\mathbb{P}(T(\text{rank of } i \text{ and rank of } j \text{ in a ranking of } n \text{ individuals}) | A_i = 0, A_j = 1) - \\ &\mathbb{P}(T(\text{rank of } i \text{ and rank of } j \text{ in a ranking of } n \text{ individuals}) | A_i = 1, A_j = 0),\end{aligned} \tag{1}$$

  where i and j are two individuals with group membership labels A_i and A_j, respectively, sampled uniformly at random from a ranking with n individuals that are i.i.d. samples drawn from $Q(S, A)$. The variable T denotes an event that only depends on the ranks of individuals i and j. Throughout the paper, we use $\mathbb{P}$ as a generic symbol of probability that subsumes all the randomness in an event, such as sampling n individuals from $Q(S, A)$ and then subsampling two individuals i and j uniformly at random. The ranking process is fair if $\mathcal{M}_1 \approx 0$. A concrete

example comes from the notion of pairwise demographic parity (pairwise *DP*) [8, 38, 49], in which case the fairness metric is:

$$DP = DP(Q(S,A);n) = \mathbb{P}(S_i > S_j | A_i = 0, A_j = 1) - \mathbb{P}(S_i > S_j | A_i = 1, A_j = 0).$$

Pairwise *DP* compares the probability that an individual from group G_0 is ranked before an individual from group G_1 to the probability that an individual from G_1 is ranked before an individual from G_0.

- **Listwise divergence-based metrics**: These metrics measure aggregate statistics over slices of the ranking. They typically use attention scores to account for the decaying exposure of items as the rank decreases [30]. The core of these metrics comes in either of the following two forms:

$$\mathcal{M}_2 = \mathcal{M}_2(Q(S,A);n) = \mathbb{P}(A_i = 1 | T(\text{rank of } i)) - \mathbb{P}(A_i = 0 | T(\text{rank of } i)), \quad \text{or} \quad (2)$$

$$\mathcal{M}_3 = \mathcal{M}_3(Q(S,A);n) = \mathbb{P}(A_i = 1 | T(\text{rank of } i)) - \mathbb{P}(A_i = 1), \quad (3)$$

where i is an individual with group membership label A_i drawn uniformly at random from a ranking of n individuals that are i.i.d. samples from $Q(S,A)$. Like before, T is an event only depending on the rank of i. The listwise metrics take T to be the event "rank of individual $i = j$" or "rank of individual $i \le j$", and sum $\mathcal{M}_2$ or $\mathcal{M}_3$ (or their absolute value) over all $j \in [n]$, weighted by an attention score.[4] Concretely, the exposure parity metric *Exp* [54] is given by

$$Exp = Exp(Q(S,A);n) = \frac{1}{\sum_{j=1}^n v_j} \sum_{j=1}^n v_j \cdot (\mathbb{P}(A_i = 1 | \text{rank of } i = j) - \mathbb{P}(A_i = 0 | \text{rank of } i = j)), \quad (4)$$

where $v_1, \ldots, v_n$ are the attention scores. A common choice is $v_i = \frac{1}{\log_2^{i+1}}$ (sometimes with $v_i = 0$ if $i > K$ for some $K < n$), reflecting the higher importance of the top positions in the ranking. Summing the absolute value of $\mathcal{M}_3$, we obtain the normalized discounted difference metric *rND* [62]:

$$rND = rND(Q(S,A);n) = \frac{1}{\sum_{j=1}^n v_j} \sum_{j=1}^n v_j \cdot |\mathbb{P}(A_i = 1 | \text{rank of } i \le j) - \mathbb{P}(A_i = 1)|. \quad (5)$$

In this work, our goal is to evaluate the fairness of a ranking process by estimating one of the fairness metrics defined in (1) - (5). Given a ranking of length n, generated by the considered process, it would be straightforward to do so if all the group membership labels $A_1, \ldots, A_n$ were observed (see the discussion below). However, this is not the case in our problem. We only get to see proxy attributes $\hat{A}_i$ that are noisy versions of A_i. Formally, we now consider a distribution $Q(S, A, \hat{A})$, and when generating a ranking of n individuals, we draw n i.i.d. samples $(S_1, A_1, \hat{A}_1), \ldots, (S_n, A_n, \hat{A}_n)$ from $Q(S, A, \hat{A})$. Observing the ranking of the scores and the proxy attributes $\hat{A}_1, \ldots, \hat{A}_n$, our goal is to estimate the fairness metrics that are defined with respect to the ground-truth group membership label A. In the next section, we show how this is possible (under certain assumptions) by relating $M_l(Q(S,A);n)$ to $M_l(Q(S,\hat{A});n)$, for $l \in [3]$; here and in the following, $Q(S,A)$ is the marginal distribution of $Q(S,A,\hat{A})$ over S and A, and $Q(S,\hat{A})$ is the marginal distribution of $Q(S,A,\hat{A})$ over S and $\hat{A}$, that is, $M_l(Q(S,\hat{A});n)$ is the fairness measurement using metric M_l based on $\hat{A}$. We refer to $M_l(Q(S,\hat{A});n)$ as the proxy measurement hereinafter.

[4] For $l \in \mathbb{N}$, we write $[l] = \{1, \ldots, l\}$.

Fairness of the ranking process vs fairness of a single ranking. Our analysis applies to a population setting where we aim to evaluate the fairness of a ranking process and we relate $M_l(Q(S,A);n)$ to $M_l(Q(S,\hat{A});n)$. However, in practice one estimates $M_l(Q(S,A);n)$ or $M_l(Q(S,\hat{A});n)$ from a single ranking (or, if available, multiple rankings) generated from the process; we denote such estimates by $\widehat{M}_l(Q(S,A);n)$ and $\widehat{M}_l(Q(S,\hat{A});n)$, respectively. Clearly, our findings relating $M_l(Q(S,A);n)$ to $M_l(Q(S,\hat{A});n)$ imply a certain relationship between $\widehat{M}_l(Q(S,A);n)$ and $\widehat{M}_l(Q(S,\hat{A});n)$: if $\widehat{M}_l(Q(S,A);n) \approx M_l(Q(S,A);n)$, $\widehat{M}_l(Q(S,\hat{A});n) \approx M_l(Q(S,\hat{A});n)$, and $M_l(Q(S,A);n) = F(M_l(Q(S,\hat{A});n))$ for a continuous and invertible $F : \mathbb{R} \to \mathbb{R}$, then $\widehat{M}_l(Q(S,A);n) \approx F(\widehat{M}_l(Q(S,\hat{A});n))$.

More concretely, given the ranking $i_1 < i_2 < \ldots < i_n$, we can estimate $DP(Q(S,A);n)$ via

$$\widehat{DP}(Q(S,A);n) = \frac{\sum_{l,k\in[n]} \mathbb{1}[\text{rank of } l > \text{rank of } k, A_l = 0, A_k = 1]}{\sum_{l,k\in[n]} \mathbb{1}[A_l = 0, A_k = 1]} - \frac{\sum_{l,k\in[n]} \mathbb{1}[\text{rank of } l > \text{rank of } k, A_l = 1, A_k = 0]}{\sum_{l,k\in[n]} \mathbb{1}[A_l = 1, A_k = 0]},$$

and similarly we estimate $DP(Q(S,\hat{A});n)$ via $\widehat{DP}(Q(S,\hat{A});n)$. Considering the definition of pairwise demographic parity in Eq. (2), it is straightforward to see that $DP(Q(S,A);n)$ and $DP(Q(S,\hat{A});n)$ are independent of n: we have

$$DP(Q(S,A);n) = \mathbb{P}[S_i > S_j, A_i = 0, A_j = 1] - \mathbb{P}[S_i > S_j, A_i = 1, A_j = 0],$$

$$DP(Q(S,\hat{A});n) = \mathbb{P}[S_i > S_j, \hat{A}_i = 0, \hat{A}_j = 1] - \mathbb{P}[S_i > S_j, \hat{A}_i = 1, \hat{A}_j = 0],$$

where the probability is over (S_i, A_i) and (S_j, A_j), or $(S_i, \hat{A}_i)$ and $(S_j, \hat{A}_j)$, being independent samples from $Q(S,A)$ or $Q(S,\hat{A})$, respectively. Therefore, we may write $DP(Q(S,A))$ instead of $DP(Q(S,A);n)$, and $DP(Q(S,\hat{A}))$ instead of $DP(Q(S,\hat{A});n)$. We have $\widehat{DP}(Q(S,A);n) \to DP(Q(S,A))$ and $\widehat{DP}(Q(S,\hat{A});n) \to DP(Q(S,\hat{A}))$ as $n \to \infty$ almost surely (cf. Appendix A). In the next section, we show that $DP(Q(S,A)) = \gamma_1 \cdot DP(Q(S,\hat{A}))$ for some $\gamma_1 \in \mathbb{R}$. This implies that $\widehat{DP}(Q(S,A);n) \approx \gamma_1 \cdot \widehat{DP}(Q(S,\hat{A});n)$ when n is sufficiently large.

For the exposure metric, we can estimate $Exp(Q(S,A);n)$ via

$$\widehat{Exp}(Q(S,A);n) = \frac{1}{\sum_{j=1}^n v_j} \sum_{j=1}^n v_j \cdot (\mathbb{1}[i_j \in G_1] - \mathbb{1}[i_j \in G_0]) = \frac{1}{\sum_{j=1}^n v_j} \sum_{j=1}^n v_j \cdot (\mathbb{1}[A_{i_j} = 1] - \mathbb{1}[A_{i_j} = 0]),$$

and similarly estimate $Exp(Q(S,\hat{A});n)$ via $\widehat{Exp}(Q(S,\hat{A});n)$. Here we can derive a relationship in expectation: we have $\mathbb{E}[\widehat{Exp}(Q(S,A);n)] = Exp(Q(S,A);n)$, where the expectation is over the random rankings generated from the ranking process. Similarly, we have $\mathbb{E}[\widehat{Exp}(Q(S,\hat{A});n)] = Exp(Q(S,\hat{A});n)$. In the next section, we show $Exp(Q(S,A);n) = \gamma_2 \cdot Exp(Q(S,\hat{A});n) + \delta_2$ for some $\gamma_2, \delta_2 \in \mathbb{R}$. This implies $\mathbb{E}[\widehat{Exp}(Q(S,A);n)] = \gamma_2 \cdot \mathbb{E}[\widehat{Exp}(Q(S,\hat{A});n)] + \delta_2$.

Finally, we estimate $rND(Q(S,A);n)$ via

$$\widehat{rND}(Q(S,A);n) = \frac{1}{\sum_{j=1}^n v_j} \sum_{j=1}^n v_j \cdot \left| \frac{\sum_{k=1}^j \mathbb{1}[i_k \in G_1]}{j} - \widehat{\mathbb{P}}(A = 1) \right|,$$

where $\widehat{\mathbb{P}}(A = 1)$ can be $\frac{\sum_{j\in[n]} \mathbb{1}[i_j \in G_1]}{|n|}$ or some other estimate of $\mathbb{P}(A = 1)$. Similarly, we estimate $rND(Q(S,\hat{A});n)$ via

$\widehat{rND}(Q(S,\hat{A});n)$. While analytically proving $|rND(Q(S,A);n) - \widehat{rND}(Q(S,A);n)| \rightarrow 0$ or $\mathbb{E}[\widehat{rND}(Q(S,A);n)] = rND(Q(S,A);n)$ seems to be challenging, we will see in our experiments that the relationship between $rND(Q(S,A);n)$ and $rND(Q(S,\hat{A});n)$ that we derive in the next section approximately also holds between $\widehat{rND}(Q(S,A);n)$ and $\widehat{rND}(Q(S,\hat{A});n)$.

Table 1 summarizes the notations used in this work.

3 EVALUATING FAIRNESS USING PROXY LABELS

In this section, we study the feasibility of measuring fairness in ranking when proxy labels are used. More formally, we investigate the relationship between the true fairness assessment $\mathcal{M}_l(Q(S,A);n)$ and the proxy measurement $\mathcal{M}_l(Q(S,\hat{A});n)$ obtained from the proxy labels, for $l \in [3]$. We first provide a theoretical proof in Section 3.1 showing that without any assumptions about the underlying distribution $Q(S,A,\hat{A})$, it is impossible to estimate bias of a ranking by only observing the proxy attribute $\hat{A}$. We lay out examples showing that by only observing the marginal distribution $Q(S,\hat{A})$, it is impossible to decide whether a ranking is fair or maximally unfair. In Section 3.2, we specify two assumptions under which it is possible to recover fairness assessment $\mathcal{M}_l(Q(S,A);n)$ from the proxy measure $\mathcal{M}_l(Q(S,\hat{A});n)$ for all the divergence-based metrics introduced in Section 2.

3.1 Feasibility of Fairness Measurement using Proxy Labels

To investigate the feasibility of bias estimation using proxy labels, we first present several graphical models that encode all conceivable relationships between variables S, A, and $\hat{A}$. Each model encapsulates our belief about the joint distribution $Q(S,A,\hat{A})$ and how it factorizes. These models are illustrated in Fig. 1.

In all the models depicted in Fig. 1, there is an edge from A to $\hat{A}$, as any reasonable estimate $\hat{A}$ of A should be caused by the variable A. Furthermore, A can cause the score S, but A cannot be caused by S. For example, a person's gender may affect their credit score, but not the other way round. Therefore, we exclude all the cases where there is an edge from S to A from our analysis. Considering all the combinations of the connections, we end up with six possible graphical models (see Fig. 1) that possibly encode the underlying distribution Q.

In the following theorem, we prove that without making any assumption, it is impossible to decide whether a ranking is fair or maximally unfair by only observing the marginal distribution $Q(S,\hat{A})$. Therefore, the proxy fairness measure obtained from the proxy labels can be arbitrarily different from the true value, and hence is unreliable.

Theorem 3.1. *Without making any assumption about the underlying data model, it is impossible to decide whether a ranking is perfectly fair or maximally unfair, and hence the proxy fairness measure can be arbitrarily different from the actual value.*

Proof. To prove this theorem, we show that there exist two data models that yield the same marginal distribution over $\hat{A}$ and S. However, in one model A and S are independent, but in the other model, any sample point with $A = 1$ has a higher score than any point with $A = 0$. Hence, according to the fairness metrics introduced in Section 2, the first model is a perfectly fair ranking process, and the second one produces maximally unfair rankings. Let the first model be

$$A \sim \text{Bernoulli}(0.5),\ S = N_U \text{ with } N_U \sim \text{Unif}([0,2)),\ \text{and } \hat{A} = \mathbb{1}[S \geq 1]$$

and the second model be

$$A \sim \text{Bernoulli}(0.5),\ S = A + N_U \text{ with } N_U \sim \text{Unif}([0,1)),\ \text{and } \hat{A} = A.$$

In both models, we have $S \sim \text{Unif}([0,2))$ and $\hat{A} = 1 \Leftrightarrow S \geq 1$, which implies that both models yield the same marginal distribution over $\hat{A}$ and S. In the first model, A and S are independent; in the second model, $S \geq 1 \Leftrightarrow A = 1$. While the rankings generated by the first model are expected to be fair, the ones following the second model are maximally unfair according to notions such as demographic parity. If we only consider the marginal distribution over $\hat{A}$ and S, it is impossible to decide whether a ranking is perfectly fair or maximally unfair, and hence the bias estimation is impossible. □

Following Theorem 3.1, it is necessary to make assumptions about the underlying data model for estimating fairness using proxy labels. Examining the models shown in Fig. 1, we exclude models 1(a) and 1(c) from our analysis, as variables A and S are independent, and hence the bias of the induced ranking process is zero. Also, the proof of Theorem 3.1 implies that estimation of bias in models shown in Fig. 1(e) and 1(f) is impossible: the two models used in the proof are both in accordance with the graphical model 1(f), and hence the bias estimation is not possible. Similar construction works for the model shown in Fig. 1(e).

This leaves us with models shown in figures 1(b) and 1(d). In the rest of our study, we will focus on these two models and refer to them as Assumption I and Assumption II, respectively. These assumptions are specified as follows:

Assumption I: Illustrated in Fig. 1(b), this assumption encodes the conditional independence between variables $\hat{A}$ and S given A, i.e., $\hat{A} \perp S|A$. We expect this assumption to hold if the proxy model is trained on features that are conditionally independent of S, given A [2]. One example is when we predict a person's race or gender from their name and compute their ranking score based on some cognitive tests.

Assumption II: This assumption encodes the conditional independence between variables A and S given $\hat{A}$, i.e., $A \perp S|\hat{A}$. The corresponding model is shown in Fig. 1(d). We expect the assumption to hold if the dependence between A and S is mediated through $\hat{A}$. One practical example is when machine learning models for predicting S are trained on engineered features derived from the sensitive attributes to break their causal effect on the outcome of the model [53].

In the rest of this section, we discuss the implications of these two assumptions and present the relationship between $\mathcal{M}(Q(S,A);n)$ and $\mathcal{M}(Q(S,\hat{A});n)$ for each metric that helps us recover the true assessment from the proxy value under each assumption.

Table 1: Important mathematical notations and their meaning.

Notation	Concept	Notation	Concept
A, $\hat{A}$	Group membership label and its proxy value	DP	Pairwise demographic parity
S	Ranking score	Exp	Exposure parity
$Q(S, A, \hat{A})$	Distribution that governs the ranking process and the generation of ranking scores, group membership labels and proxy group membership labels	rND	Normalized discounted difference
$Q(S, A)$, $Q(S, \hat{A})$	Marginal distributions of $Q(S, A, \hat{A})$	G_0	Group of individuals with $A = 0$
$\mathcal{M}(Q(S, A); n)$	Fairness measurement of ranking process w.r.t. A	G_1	Group of individuals with $A = 1$
$\mathcal{M}(Q(S, \hat{A}); n)$	Proxy fairness measurement of ranking process w.r.t. $\hat{A}$		
$\widehat{\mathcal{M}}(Q(S, A); n)$	Estimate of fairness measurement based on a single ranking (computed using A)	β	Population fraction of G_1
		p	Proxy model's error rate for G_0: $\mathbb{P}(\hat{A} = 1 \mid A = 0)$
$\widehat{\mathcal{M}}(Q(S, \hat{A}); n)$	Estimate of proxy fairness measurement based on a single ranking (computed using $\hat{A}$)	q	Proxy model's error rate for G_1: $\mathbb{P}(\hat{A} = 0 \mid A = 1)$
		r	rND ratio

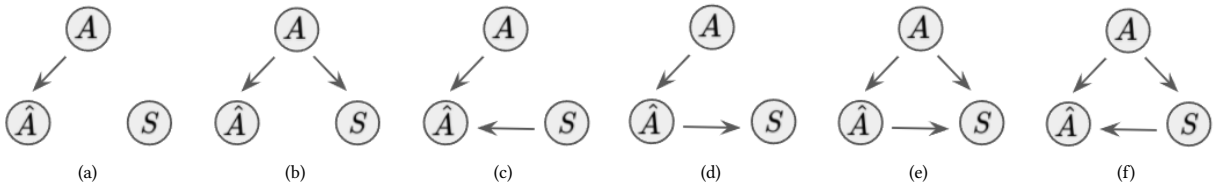


Figure 1: Graphical models encoding different assumptions about the underlying data.

3.2 Fairness Measurement Correction

Given the specified assumptions, we aim to identify the relationship between $\mathcal{M}_l(Q(S, A); n)$ and $\mathcal{M}_l(Q(S, \hat{A}); n)$ for all the metrics formally defined in Eq. (1), (2) and (3).

Measurement correction under Assumption I. We can show that under this assumption, the following relationships hold:

$$\mathcal{M}_1(Q(S, A); n) = \frac{\mathcal{M}_1(Q(S, \hat{A}); n) \cdot x \cdot y}{\beta \cdot (1 - \beta) \cdot (1 - p - q)}, \tag{6}$$

$$\mathcal{M}_2(Q(S, A); n) = \frac{\mathcal{M}_2(Q(S, \hat{A}); n) - p + q}{1 - p - q}, \quad \text{and} \tag{7}$$

$$\mathcal{M}_3(Q(S, A); n) = \frac{\mathcal{M}_3(Q(S, \hat{A}); n)}{1 - p - q} \tag{8}$$

where $\beta = \mathbb{P}(A = 1)$ which can be estimated on the training data available to the proxy model that predicts $\hat{A}$. The variables p and q are group-conditional error rates of the proxy model, which are reported on some test data where the ground-truth group membership labels, A, are observed. These variables are defined as $p = \mathbb{P}(\hat{A} = 1 | A = 0)$, and $q = \mathbb{P}(\hat{A} = 0 | A = 1)$. Also variables x and y are defined as follows:

$$x = (1 - q) \cdot \beta + p \cdot (1 - \beta), \quad y = q \cdot \beta + (1 - p) \cdot (1 - \beta)$$

For the proof, we refer the reader to Appendix B.

Measurement correction under Assumption II. Similarly, we derive the following relationships under the second assumption:

$$\mathcal{M}_1(Q(S, A); n) = \mathcal{M}_1(Q(S, \hat{A}); n) \cdot (1 - p - q), \tag{9}$$

$$\mathcal{M}_2(Q(S, A); n) = (\mathcal{M}_2(Q(S, \hat{A}); n) + 1) \cdot \left(\frac{(1 - q) \cdot \beta}{x} - \frac{q \cdot \beta}{y} \right) + \frac{2q \cdot \beta}{y} - 1, \quad \text{and} \tag{10}$$

$$\mathcal{M}_3(Q(S, A); n) = \mathcal{M}_3(Q(S, \hat{A}); n) \cdot \left(\frac{(1 - q) \cdot \beta}{x} - \frac{q \cdot \beta}{y} \right) \tag{11}$$

We present the proof of the above relationships in Appendix C.

According to the described relationships, we can estimate $\mathcal{M}(Q(S, A); n)$ under each assumption using the estimation of $\mathcal{M}(Q(S, \hat{A}); n)$ and the quantities β, p, and q.

4 EVALUATION SETUP

We now examine the performance of the proposed bias correction methods described in Section 3.2 in practice. Using proxy labels, we first compute $\widehat{\mathcal{M}}_l(Q(S, \hat{A}); n)$ for all $l \in [3]$. As concrete examples of $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$, we use demographic parity (DP, Eq. (2)), exposure parity (Exp, Eq. (4)), and normalized discounted difference (rND, Eq. (5)), respectively. In Section 3.2, we showed that under each assumption, $\mathcal{M}_l(Q(S, A); n)$ can be recovered from the proxy value $\mathcal{M}_l(Q(S, \hat{A}); n)$ using some correction factor γ_l and in some cases an additive correction term δ_l. Similarly, we correct the measurement $\widehat{\mathcal{M}}_l(Q(S, \hat{A}); n)$ and refer to the corrected value as $\widehat{\mathcal{M}}_l^{rec}(Q(S, \hat{A}); n)$. As a concrete example for $l = 1$ in Assumption II using DP, we will get:

$$DP^{rec}(Q(S, \hat{A}); n) = \gamma_1 \cdot \widehat{DP}(Q(S, \hat{A}); n)$$

where $\gamma_1 = 1 - p - q$ according to Eq. (9).
We measure the accuracy of the corrected estimates using the following metrics:

$$error_{rec} = \widehat{\mathcal{M}}_l(Q(S,A);n) - \widehat{\mathcal{M}}_l^{rec}(Q(S,\hat{A});n)\ ,$$
$$error_{ratio} = \frac{|error_{rec}|}{|error_{proxy}|}$$

where $error_{proxy}$ is the error of the proxy bias estimate computed using proxy labels. It is defined as:

$$error_{proxy} = \widehat{\mathcal{M}}_l(Q(S,A);n) - \widehat{\mathcal{M}}_l(Q(S,\hat{A});n).$$

The $error_{rec} \in (-\infty, \infty)$ computes the error of the corrected bias, and the $error_{ratio} \in [0, \infty)$ is the ratio of the corrected bias error to that of the proxy value $\widehat{\mathcal{M}}(Q(S,\hat{A});n)$. Ideally, we expect $error_{rec} \approx 0$ and $error_{ratio} \ll 1$.

4.1 Synthetic Datasets

We first evaluate the bias estimate correction methods on synthetic datasets. This enables us to examine the behavior of bias estimate correction with respect to different parameters such as the size of the dataset or the proxy model's group-conditional error rates (denoted by p and q). In what follows, we describe the synthetic data generation process for both the assumptions specified in Section 3.1.

Generating synthetic data under Assumption I. To create synthetic rankings, we generate a two-dimensional dataset where each data point is associated with two features: (i) a binary-valued group membership label, and (ii) a real-valued score used for ranking items. We assume that the ranking scores of the members in group G_0 (with $A = 0$) and the ones in G_1 (with $A = 1$) are generated according to a one-dimensional distribution D_0 and D_1, respectively. Given the total number of samples, n, and G_1's population fraction, β, we sample $n \times \beta$ numbers from D_1 and treat them as the ranking scores of the data points in G_1. Similarly, we sample $n \times (1-\beta)$ numbers from D_0 and assign them to the members of G_0. Next, we generate a ranking by sorting the pairs of scores and labels in descending order of their score values. Finally, we generate the proxy labels $\hat{A}$ using a classifier that randomly flips the label $A = 1$ with probability q to $A = 0$ and the label $A = 0$ with probability p to $A = 1$. We refer to this model as a flip classifier and denote it by $flip(p,q)$, hereinafter.

Generating synthetic data under Assumption II. Similar to the procedure described for Assumption I, we first generate $n \times \beta$ instances with label $A = 1$ and $n \times (1-\beta)$ ones with $A = 0$. Next, we compute proxy labels $\hat{A}$ for all the generated instances using the flip classifier $flip(p,q)$. Following the definition of Assumption II, we finally sample the ranking scores S according to the proxy labels $\hat{A}$: for instances with $\hat{A} = 0$ and the ones with $\hat{A} = 1$, we sample score values from the distributions D_0 and D_1, respectively, and subsequently sort the instances according to their scores.

Default setting: In all our experiments with synthetic datasets, we assume that distributions D_0 and D_1 are both normal and are specified with (μ_0, σ_0) and (μ_1, σ_1), respectively. We set the parameters as follows unless otherwise stated: $n = 10\text{k}, \beta = 0.5, \mu_0 = 2, \sigma_0 = 2, \mu_1 = 1$, and $\sigma_1 = 0.5$. Therefore, we enforce bias against the group G_1 in the resulting rankings by deliberately choosing $\mu_0 > \mu_1$.

4.2 Real-world Datasets

In addition to the synthetic data, we test the performance of the proposed bias estimate correction methods on the following real-world datasets:

Goodreads Authors dataset. This dataset has been recently crawled from the Goodreads website and used for author analysis.[5] It contains information, such as name, bio, and work-count of over 209k authors. Note that we filtered out authors whose names contained non-English letters. We use the binary-valued feature gender as the sensitive attribute (male: G_0 and female: G_1)[6], and sort the records based on author's number of fans.

FIFA Players dataset. We use the FIFA 20 player dataset[7] that contains information about over 18k FIFA players including their nationality and their value in Euro. Following the setup introduced in [2], we treat the nationality as the sensitive attribute and restrict the dataset to players of English or German nationality. This reduces the size of data to 2.9k players. We use the real-valued feature "value in Euro" as the ranking score to sort the players.

IBM HR Analytics dataset. This is a fictional dataset created by IBM scientists[8] to study employees' attrition rate and performance level. It has about 2k records, each containing information such as education, monthly income, and experience years about one employee. We sort the employees based on the number of years predicted for them to stay in their current role, and use their marital status as the sensitive attribute. We consider two values for marital status: single and other. The latter describes both married and divorced people.

COMPAS Dataset. This dataset contains the criminal history of ~61k defendants from Broward County along with their recidivism risk score computed by COMPAS tool.[9] In our experiments, we treat the feature race as the sensitive attribute and binarize its values to labels "African-American" and "Others". We cast the recidivism risk prediction to a ranking problem and sort the instances according to the predicted recidivism scores[10]. The ground truth scores have three different values of *low*, *medium*, and *high*. We broke the ties arbitrarily. In Appendix D, we provide a summary of the real-world datasets used in our study.

4.3 Initialization

Parameter initialization for metrics: In our experiments, we compute the metric rND up until the ranking position $K = r \cdot n$, where we set $r = 0.1$. For rND and Exp metrics, we define the attention distribution according to the logarithmic decaying factor used in the normalized Discounted Cumulative Gain (nDCG) [59]: $\mathbb{P}(i) = \frac{1}{\log_2^{i+1}} / \sum_{j=1}^{n} \frac{1}{\log_2^{j+1}}$.

Parameter initialization for proxy models. We used sklearn Python library to implement standard classifiers such as Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and

[5] https://www.kaggle.com/choobani/goodread-authors
[6] While gender is non-binary, one limitation of the datasets is that only binary data is reported.
[7] https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset
[8] https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset
[9] https://github.com/propublica/compas-analysis
[10] We acknowledge that our experiments with COMPAS dataset do not address the complexities of working with risk assessment instrument (RAI) datasets [5]. Our results on this dataset should not lead to any actionable, real-world insights, especially in criminal justice.

Multilayer Perceptron (*MLP*). For *SVM*, we chose the linear kernel and set the maximum number of iterations to 4000, and the regularization coefficient C to 0.1. For *MLP*, we set the maximum number of iterations to 1000 to ensure convergence. The rest of the parameters were set to their default value. For categorical features, we used one-hot encoding. Inspired by [2], we also considered LSTM as a proxy model in some of our experiments. We used the Tensorflow implementation of this model with 64 units and a dropout rate of 0.25. We set the batch size to 64 and the number of epochs to 30.

4.4 Experimental Setup

To verify the effectiveness of the correction methods on both synthetic and real datasets, we designed an experiment with the following pipeline: (1) we first split the dataset into train set (60%), which is used for training the proxy model and estimating β, test set (20%) used for estimating the correction terms p and q, and evaluation set (20%) used for validating the correction methods, (2) next, we estimate the actual fairness metric value, $\widehat{\mathcal{M}}_l(Q(S,A);n)$, on the evaluation set, (3) train the proxy model on the training data, and estimate its group-conditional error rates on the test set, (4) predict the labels of items in the evaluation set using the learned proxy model, (5) estimate $\widehat{\mathcal{M}}_l(Q(S,\hat{A});n)$, the bias of the evaluation data using the proxy labels, and (6) finally compute the corrected estimation, $\widehat{\mathcal{M}}_l^{rec}(Q(S,\hat{A});n)$. To break the dependency of the estimations from one specific data split, we repeat each experiment 10 times for each metric and show the boxplots of $error_{rec}$ and $error_{ratio}$ values for all datasets. These plots illustrate the range of errors (minimum, maximum, and the median values) along with the inter-quantile ranges.

5 RESULTS AND INSIGHTS

5.1 Evaluating Fairness on Synthetic Datasets

Investigating the effect of different parameters, the following observations are made on the synthetic data:

The corrected bias estimate converges to the true value for sufficiently large datasets. To explore the effect of the dataset size on the reliability of bias estimate correction, we vary the parameter n from 100 to 100k on a logarithmic scale. Fig. 2 shows the results of this experiment for measurement correction under both assumptions. As evident, for all metrics, $error_{rec}$ centers towards zero with small variance as n grows larger. This observation is explained by the fact that our estimations for variables p, q and β become more accurate with the increase in n, and hence the corrected estimates become closer to the actual bias.

Bias estimate correction is robust against variation of group-conditional error rates. To understand the effect of group-conditional error rates on the accuracy of the corrected values, we kept the error rate of group G_0, denoted by p, fixed ($p = 0.3$) and varied error rate q from 0.1 to 0.9. Fig. 3 shows that the bias estimate correction under Assumption I is stable regardless of the value of q. Note that the variance of the corrected values increases as q approaches 0.7. The reason for this observation is that in equations (6)-(8), the denominator approaches zero as $p + q$ becomes closer to 1. For $q = 0.7$, these relationships become undefined, and hence we return the proxy estimates. We omitted the corresponding plot for Assumption II as the trends were overall similar to those observed for Assumption I.

5.2 Evaluating Fairness on Real-world Datasets

Empirical validation of Assumption I. We verified Assumption I on two datasets: Goodreads Authors and FIFA Players. To obtain proxy labels, we trained *LSTM* models for predicting the gender of authors using their first names, and the nationality of players using their last names. This way, we expect Assumption I to hold as it is safe to assume that given the gender or nationality of a person, their names are conditionally independent of the attributes that determine their number of fans or their Euro value (for soccer players). Fig. 4 summarizes the results of our experiments on these two datasets. We observe that for all metrics $error_{rec}$ is significantly smaller than $error_{proxy}$, suggesting that our proposed bias estimate correction methods yield more accurate estimates.

Empirical validation of Assumption II. We tested Assumption II on two datasets: COMPAS and IBM HR Analytics. To obtain proxy labels, we used the standard classifiers *LR*, *SVM*, *RF*, and *MLP*. In the COMPAS dataset, we used the predicted recidivism risk as the ranking score. In the IBM HR Analytics, we sorted the records according to the predicted value of the "Years In Current Role". In both datasets, we break the ties arbitrarily.

To enforce Assumption II, we take the following strategy: we first sort features according to their importance for predicting the group membership labels A_i. For this, we first train an *RF* classifier for predicting A using the entire set of features and then sort features according to their importance for predicting A computed using the permutation technique [21]. Next, we use the first half of the features for learning the proxy labels $\hat{A}$. Then, we use the predicted $\hat{A}$ together with the second half of the features to predict the ranking scores S. For this, we use the learning-to-rank model, LambdaMART [13], implemented in the Python library, pyltr.[11] This way, we mediate the causal effect of A on predicted S, by using $\hat{A}$ along with other features least correlated with A.

Figures 5 and 6 show the comparison of the proxy and the corrected estimates for all the selected metrics. For both datasets, the average $error_{ratio}$ for *DP* and *Exp* metrics is consistently below one, demonstrating the success of bias estimate correction when different proxy models are used. For the *rND* metric, we observe that the bias estimate correction performs poorly in the IBM HR Analytics dataset. This observation can be attributed to the small size of the dataset which makes the estimation of group-conditional error rates in the top 10% of the ranking less reliable. Another exception is the low accuracy of corrected estimates in COMPAS dataset when *RF* is the proxy model. This might be due to the violation of Assumption II which will be discussed in Section 5.4.

5.3 Comparison with Weighted Sampling

We compared the accuracy of the corrected measurements against that of the estimates computed using a recently proposed method for bias estimation in settings with incomplete group annotations [35]. To measure fairness, this method relies on crowdsourcing group membership labels for a small subset of individuals in the ranking. To select individuals, it employs a top-heavy sampling

[11] https://github.com/jma127/pyltr

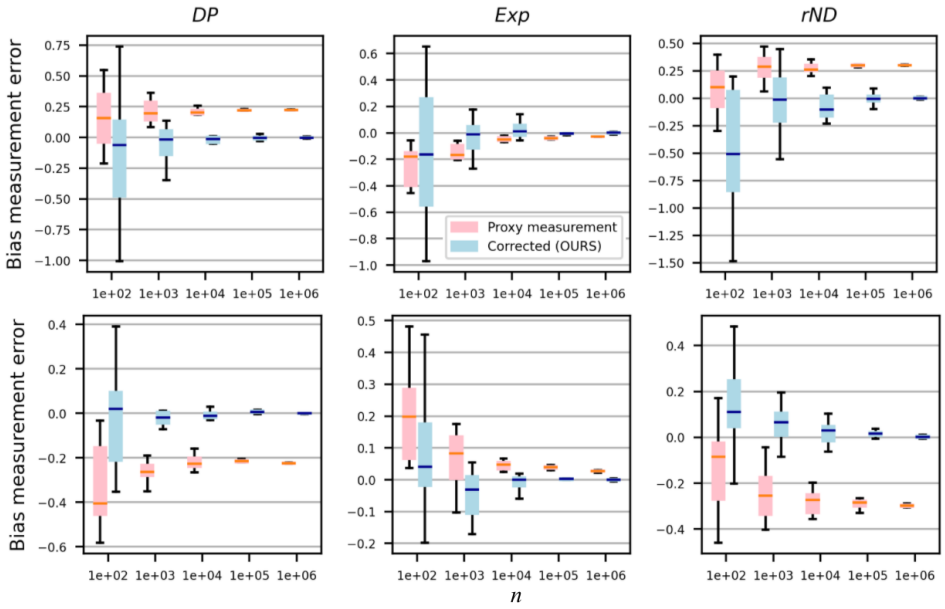


Figure 2: Effect of n on the error of the proxy and corrected estimates. First row: Assumption I, second row: Assumption II. The horizontal line in each box represents the median.

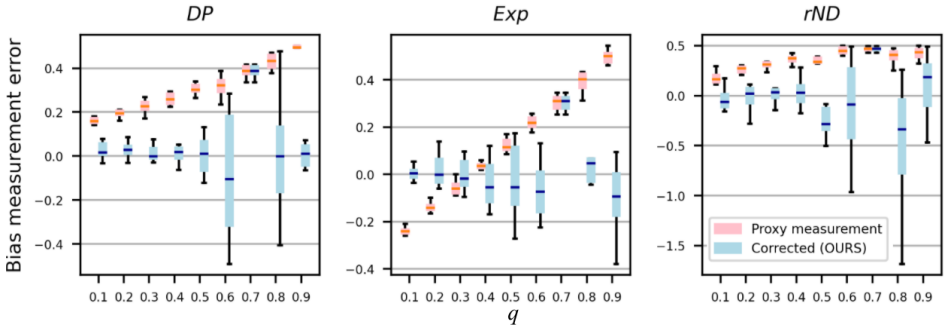


Figure 3: Effect of q on mean and variance error of the proxy and corrected estimates under Assumption I with $n = 10$k. The horizontal line in each box represents the median.

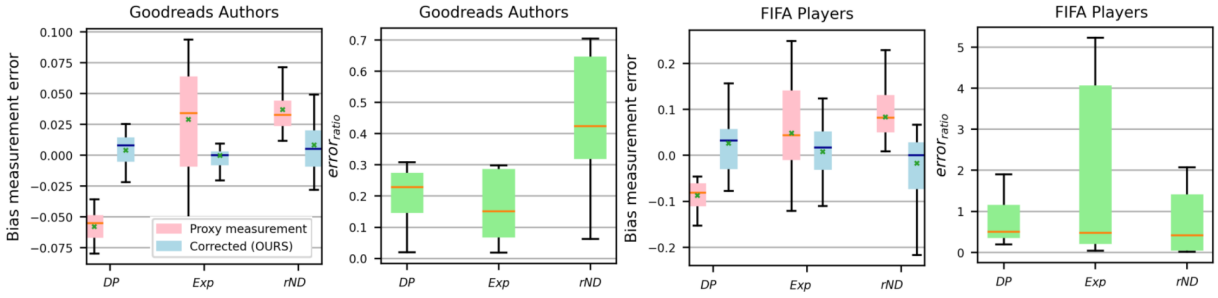


Figure 4: Empirical validation of Assumption I on Goodreads Authors (first two plots) and FIFA Players dataset (last two plots) for different metrics. The horizontal line and the green "x" symbol in each box represent the median and mean, respectively.

strategy, i.e., the individuals at the top of the ranking have a higher chance of being chosen. Using the collected labels, it then measures fairness using unbiased estimators. For more details, we refer the reader to [35]. We used our own implementation of this method

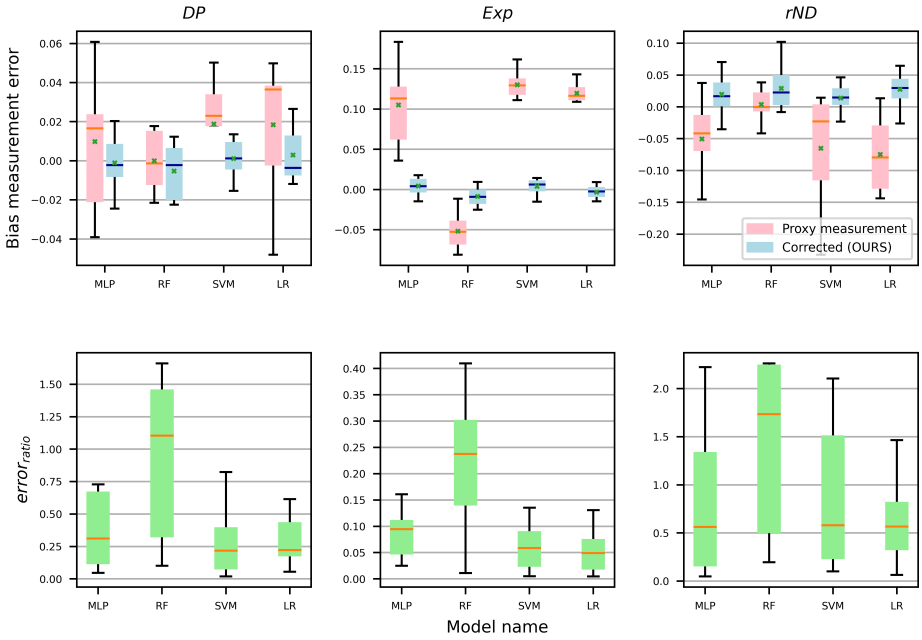


Figure 5: Empirical validation of Assumption II on COMPAS dataset. The green "x" symbols represent the mean values.

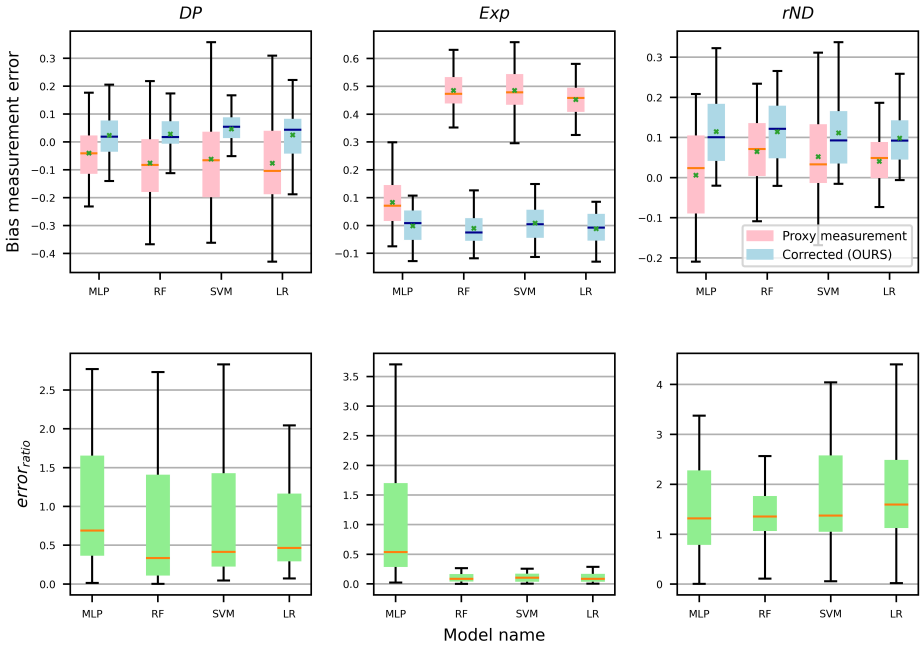
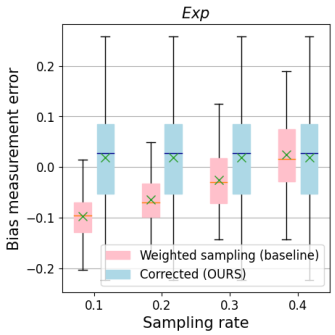
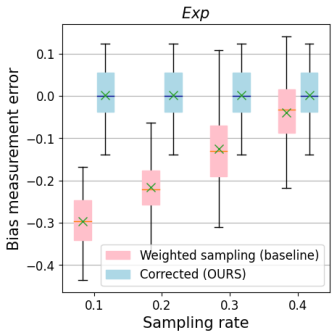


Figure 6: Empirical validation of Assumption II on IBM HR Analytics dataset. The green "x" symbols represent the mean values.

and compared it against our method for bias estimate correction at different sampling rates for the *Exp* metric, which is explored in both their study and our work. The sampling rates determine the fraction of the crowdsourced items. Fig. 7 summarizes the results on two real-world datasets: FIFA Players and IBM HR Analytics. Note that with the current choice of datasets, we can compare the



(a) FIFA Players dataset



(b) IBM HR Analytics dataset

Figure 7: Comparison of our method with the weighted sampling [35]. The green "x" symbols represent the mean values.

Table 2: Analysis of the proxy models w.r.t. their error rates and their extent of assumption violation for each dataset.

	Goodreads Authors	FIFA Players	COMPAS				IBM HR Analytics			
	LSTM	LSTM	MLP	RF	SVM	LR	MLP	RF	SVM	LR
Avg. p	0.09	0.16	0.30	0.35	0.29	0.29	0.26	0.05	0.06	0.06
Avg. q	0.12	0.33	0.48	0.39	0.50	0.50	0.68	0.85	0.87	0.86
Avg. accuracy	0.90	0.77	0.62	0.63	0.62	0.62	0.60	0.70	0.68	0.68
$\%\rho > 0.05$	100	100	97	97	100	97	100	100	100	99
$\%\rho > 0.1$	97	100	96	87	93	93	99	99	99	99

bias estimate correction methods against the baseline under both Assumption I and Assumption II. The results indicate that the baseline is comparable with our methods for bias estimate correction only at higher sampling rates. This highlights the benefit of our method when obtaining group membership annotations is costly, but proxy labels are available.

5.4 Analysis

We further examined the proxy models learned in our experiments to answer the following two questions: (i) is there any trend in the error rates of the trained proxy models that possibly contribute to the effectiveness of bias estimate correction?, and (ii) to what extent are the specified assumptions violated in the chosen datasets?

To answer the first question, we computed the average group-conditional error rates of all the trained classifiers as presented in the first two rows of Table 2. The results indicate that bias estimate correction is effective not only when the error rates are low and balanced (e.g., LSTM in Goodreads Authors), but also when the models have relatively high (e.g., MLP in COMPAS) or imbalanced error rates (e.g., RF in IBM HR Analytics). This shows that regardless of the proxy model's error rates, corrected bias values are more accurate than the proxy values.

As for the second question, we performed statistical tests on the chosen datasets to ensure that the assumptions are not violated. Note that performing such tests in real situations may not be possible as it requires common data where all the three variables A, $\hat{A}$ and S are observed. We used Kruskal Wallies H-test [45], a non-parametric version of one-way ANOVA [28], to test the independence between the categorical variable A (or $\hat{A}$) and the continuous variable S. As we are interested in testing conditional independence between two variables given a third one, we perform Kruskal Wallies H-test twice: once for $A = 1$ in Assumption I (or $\hat{A} = 1$ in Assumption II), and once for $A = 0$ (or $\hat{A} = 0$ in Assumption II). We reject the conditional independence between two variables if both the tests are rejected. The last two rows in Table 2 show the percentage of experiments where we could not reject the conditional independence for either of the assumptions at $\rho\text{-}value = 0.05$ and ρ value 0.05, respectively. As expected, in the majority of our experiments the established assumptions could not be rejected. The only exception is when we use *RF* for learning proxy labels in COMPAS dataset. This justifies the low accuracy of the corrected measurements for *RF* classifier as illustrated in Fig. 5.

6 RELATED WORK

Fairness notions. Evidence of algorithmic bias in various domains [4, 6, 7, 37, 51, 52] has prompted efforts to define notions and metrics for evaluating fairness of the systems' outcomes [46, 65]. The fairness notions studied in the literature can be classified into two groups: (i) *individual fairness* which emphasizes on consistent treatment of individuals [9, 17], and (ii) *group fairness* which ensures fair treatment across different groups of entities [50]. In this work, we focus on evaluating the group fairness of the ranked outputs with respect to their constituent items [14]. For instance, for a ranked set of job applicants, we would like to know if there is bias against certain genders. Metrics commonly used for measuring group fairness are divergence-based, i.e., they compare proportion

or aggregate ranking positions of a certain group against a target value [35, 54]. To account for the order of items, these metrics are either pairwise, such as pairwise demographic parity [8, 36, 49] or listwise, such as rND or its equivalents [23, 62, 63] and disparate exposure [54, 64]. In our study, we mainly focus on divergence-based metrics for computing statistical parity and leave metrics such as equal opportunity [25, 36] for future work.

Fairness measurement under uncertainty assumptions. Evaluating fairness using proxy labels is a fundamentally challenging problem [2, 16, 31]. Recent studies have provided theoretical evidence showing that fairness estimates obtained from the proxy labels can be arbitrarily different from the true assessments unless strong assumptions are satisfied [2, 31]. These studies, however, are primarily concerned with metrics used to assess fairness of classifiers. In fact, there has been very little work addressing similar problems for the ranking tasks. Ghosh et al. [24] present empirical evidence showing the negative impact of using proxy labels on the output of fair ranking algorithms. In another work, Kirnap et al. [35] study a similar problem where group membership labels are available but at a cost. They propose unbiased estimators for several ranking fairness metrics that rely on crowdsourcing ground truth labels for a subset of data. In our work, we eliminate the need for crowdsourcing by proposing methods for rectifying the fairness measurements when proxy labels are available.

Bias mitigation in ranking. To reduce the influence of algorithmic bias on society, there has been a growing body of literature on operationalizing diversity in ranking. Existing techniques for bias mitigation fall into three categories of pre-processing, in-processing, and post-processing methods [65]. The pre-processing methods attempt to reduce bias in downstream tasks through learning fair representations [40]. The in-processing techniques, however, take an end-to-end approach and directly optimize for fair ranking by imposing (soft) constraints that enforce fairness of the output [48, 55, 60, 64]. The last group of methods relies on fairness-aware re-ranking to introduce diversity among the results without too much sacrificing the accuracy [15, 23, 42, 61, 63]. For a broad survey, we refer the reader to [18, 65].

7 CONCLUSIONS AND FUTURE WORK

In this work, we investigated the reliability of group fairness assessment in ranking when proxy group membership labels are used. We provided a theoretical analysis showing that in this situation bias estimation is impossible unless certain assumptions about the underlying data model hold. We specified two feasible assumptions and presented their implications on bias estimation for a suite of divergence-based metrics. We showed that under each assumption, the true bias can be recovered from the estimates obtained from the proxy labels, and presented empirical validation of the proposed bias estimate correction methods on synthetic as well as real-world datasets. Our findings provide valuable insights for bias estimation in IR systems that rely on proxy labels (e.g., blind recruitment tools that do not ask for the race or gender of the applicants).

Our formalism is currently focused on a single binary sensitive attributes. It can be easily extended to support multiple, multi-value attributes by computing bias for each value separately in a one-vs-all approach. With that in mind, a true extension of our work to support many multi values attributes, is a natural direction for future work.

REFERENCES

[1] Dana Anderson. 2021. The Price of Racial Bias: Homes in Black Neighborhoods Are Valued at an Average of $46,000 Less Than Similar Homes in White Neighborhoods. https://www.redfin.com/news/undervaluation-homes-black-versus-white-neighborhoods/.

[2] Pranjal Awasthi, Alex Beutel, Matthäus Kleindessner, Jamie Morgenstern, and Xuezhi Wang. 2021. Evaluating Fairness of Machine Learning Models Under Uncertain and Incomplete Information. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 206–214.

[3] Arthur P Baines and Marsha J Courchane. 2014. Fair lending: Implications for the indirect auto finance market. *study prepared for the American Financial Services Association* (2014).

[4] Ryan S Baker and Aaron Hawn. 2021. Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education* (2021), 1–41.

[5] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. 2021. It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. *arXiv preprint arXiv:2106.05498* (2021).

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness in machine learning. *NeurIPS tutorial* 1 (2017), 2017.

[7] Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. 2018. Consumer-lending discrimination in the era of fintech. *Unpublished working paper. University of California, Berkeley* (2018).

[8] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. 2019. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2212–2220.

[9] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 405–414.

[10] Felix Biessmann, Tammo Rukat, Philipp Schmidt, Prathik Naidu, Sebastian Schelter, Andrey Taptunov, Dustin Lange, and David Salinas. 2019. DataWig: Missing Value Imputation for Tables. *Journal of Machine Learning Research* 20, 175 (2019), 1–6.

[11] Felix Biessmann, David Salinas, Sebastian Schelter, Philipp Schmidt, and Dustin Lange. 2018. " Deep" Learning for Missing Value Imputationin Tables with Non-Numerical Data. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 2017–2025.

[12] Miranda Bogen, Aaron Rieke, and Shazeda Ahmed. 2020. Awareness in practice: tensions in access to sensitive attribute data for antidiscrimination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 492–500.

[13] Christopher JC Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning* 11, 23-581 (2010), 81.

[14] Robin Burke. 2017. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093* (2017).

[15] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. 2018. Ranking with Fairness Constraints. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[16] 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*. 339–348.

[17] 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*. 214–226.

[18] Michael D Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2021. Fairness in Information Access Systems. *arXiv preprint arXiv:2105.05779* (2021).

[19] Marc N Elliott, Peter A Morrison, Allen Fremont, Daniel F McCaffrey, Philip Pantoja, and Nicole Lurie. 2009. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology* 9, 2 (2009), 69–83.

[20] Alessandro Fabris, Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani. 2021. Measuring Fairness under Unawareness via Quantification. *arXiv preprint arXiv:2109.08549* (2021).

[21] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2019. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *J. Mach. Learn. Res.* 20, 177 (2019), 1–81.

[22] Dmitry Gavinsky, Shachar Lovett, Michael Saks, and Srikanth Srinivasan. 2014. A tail bound for read-k families of functions. *Random Structures and Algorithms* 47, 1 (2014), 99–108.

[23] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin

talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2221–2231.
[24] Avijit Ghosh, Ritam Dutt, and Christo Wilson. 2021. When fair ranking meets uncertain inference. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1033–1043.
[25] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems* 29 (2016), 3315–3323.
[26] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*. PMLR, 1929–1938.
[27] Rebecca Heilweil. 2019. Artificial intelligence will help determine if you get your next job. https://www.vox.com/recode/2019/12/12/20993665/artificial-intelligence-ai-job-screen.
[28] David C Howell. 2012. *Statistical methods for psychology*. Cengage Learning.
[29] Kosuke Imai and Kabir Khanna. 2016. Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis* 24, 2 (2016), 263–272.
[30] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. 2007. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)* 25, 2 (2007), 7–es.
[31] Nathan Kallus, Xiaojie Mao, and Angela Zhou. 2021. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science* (2021).
[32] Nathan Kallus and Angela Zhou. 2019. Assessing disparate impact of personalized interventions: identifiability and bounds. *Advances in Neural Information Processing Systems* 32 (2019).
[33] Nathan Kallus and Angela Zhou. 2019. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. *Advances in Neural Information Processing Systems* 32 (2019).
[34] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3819–3828.
[35] Ömer Kırnap, Fernando Diaz, Asia Biega, Michael Ekstrand, Ben Carterette, and Emine Yilmaz. 2021. Estimation of Fair Ranking Metrics with Incomplete Judgments. In *Proceedings of the Web Conference 2021*. 1065–1075.
[36] Matthäus Kleindessner, Samira Samadi, Muhammad Bilal Zafar, Krishnaram Kenthapadi, and Chris Russell. 2021. Pairwise Fairness for Ordinal Regression. *arXiv preprint arXiv:2105.03153* (2021).
[37] Alina Köchling and Marius Claus Wehner. 2020. Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research* (2020), 1–54.
[38] Caitlin Kuhlman, MaryAnn VanValkenburg, and Elke Rundensteiner. 2019. FARE: Diagnostics for Fair Ranking using Pairwise Error Metrics. In *International Conference on World Wide Web (WWW)*.
[39] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. 2019. Search bias quantification: investigating political bias in social media and web search. *Information Retrieval Journal* 22, 1 (2019), 188–227.
[40] 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 (ICDE)*. IEEE, 1334–1345.
[41] Huyen Le, Raven Maragh, Brian Ekdale, Andrew High, Timothy Havens, and Zubair Shafiq. 2019. Measuring political personalization of Google news search. In *The World Wide Web Conference*. 2957–2963.
[42] Weiwen Liu, Jun Guo, Nasim Sonboli, Robin Burke, and Shengyu Zhang. 2019. Personalized fairness-aware re-ranking for microlending. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 467–471.
[43] Mark MacCarthy. 2018. Standards of fairness for disparate impact assessment of big data. *Cumberland Law Review* 48, 1 (2018), 67–145.
[44] Douglas MacMillan and Nick Anderson. 2019. Student tracking, secret scores: How college admissions offices rank prospects before they apply. https://www.washingtonpost.com/business/2019/10/14/colleges-quietly-rank-prospective-students-based-their-personal-data/.
[45] Patrick E McKight and Julius Najab. 2010. Kruskal-wallis test. *The Corsini Encyclopedia of Psychology* (2010), 1–1.
[46] 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.
[47] Anay Mehrotra and L Elisa Celis. 2021. Mitigating bias in set selection with noisy protected attributes. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 237–248.
[48] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. 2020. Controlling fairness and bias in dynamic learning-to-rank. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 429–438.
[49] Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Serena Wang. 2020. Pairwise fairness for ranking and regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5248–5255.
[50] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 560–568.
[51] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 469–481.
[52] Eva Rosen, Philip ME Garboden, and Jennifer E Cossyleon. 2021. Racial discrimination in housing: how landlords use algorithms and home visits to screen tenants. *American Sociological Review* 86, 5 (2021), 787–822.
[53] Ricardo Salazar, Felix Neutatz, and Ziawasch Abedjan. 2021. Automated feature engineering for algorithmic fairness. *Proceedings of the VLDB Endowment* 14, 9 (2021), 1694–1702.
[54] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2219–2228.
[55] Ashudeep Singh and Thorsten Joachims. 2019. Policy Learning for fairness in ranking. *Advances in Neural Information Processing Systems* 32 (2019).
[56] Ashudeep Singh, David Kempe, and Thorsten Joachims. 2021. Fairness in ranking under uncertainty. In *Thirty-Fifth Conference on Neural Information Processing Systems*.
[57] Nenad Tomasev, Kevin R McKee, Jackie Kay, and Shakir Mohamed. 2021. Fairness for unobserved characteristics: Insights from technological impacts on queer communities. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 254–265.
[58] Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael Jordan. 2020. Robust optimization for fairness with noisy protected groups. *Advances in Neural Information Processing Systems* 33 (2020), 5190–5203.
[59] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A theoretical analysis of NDCG type ranking measures. In *Conference on Learning Theory*. 25–54.
[60] Himank Yadav, Zhengxiao Du, and Thorsten Joachims. 2021. Policy-Gradient Training of Fair and Unbiased Ranking Functions. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1044–1053.
[61] Ke Yang, Vasilis Gkatzelis, and Julia Stoyanovich. 2019. Balanced Ranking with Diversity Constraints. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI*.
[62] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. 1–6.
[63] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *ACM Conference on Information and Knowledge Management (CIKM)*. 1569–1578.
[64] Meike Zehlike and Carlos Castillo. 2020. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of The Web Conference 2020*. 2849–2855.
[65] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2021. Fairness in Ranking: A Survey. *arXiv preprint arXiv:2103.14000* (2021).

APPENDIX

A Almost sure convergence of $\widehat{DP}(Q(S,A);n)$ to $DP(Q(S,A))$

We have

$$DP(Q(S,A);n) = \mathbb{P}(S_i > S_j | A_i = 0, A_j = 1) - \mathbb{P}(S_i > S_j | A_i = 1, A_j = 0)$$

with (S_i, A_i) and (S_j, A_j) being independent samples from $Q(S,A)$ and

$$\widehat{DP}(Q(S,A);n) = \frac{\sum_{l,k\in[n]} \mathbb{1}[S_l > S_k, A_l = 0, A_k = 1]}{\sum_{l,k\in[n]} \mathbb{1}[A_l = 0, A_k = 1]} - \frac{\sum_{l,k\in[n]} \mathbb{1}[S_l > S_k, A_l = 1, A_k = 0]}{\sum_{l,k\in[n]} \mathbb{1}[A_l = 1, A_k = 0]},$$

where $(S_1, A_1), \ldots, (S_n, A_n)$ is an i.i.d. sample from $Q(S,A)$.

The random variables $\mathbb{1}[S_l > S_k, A_l = 0, A_k = 1]_{l,k\in[n]}$ are a read-n family [22, Definition 1]. It is $\mathbb{P}[\mathbb{1}[S_l > S_k, A_l = 0, A_k = 1] = 1] = \mathbb{P}[S_l > S_k, A_l = 0, A_k = 1]$ for $l \neq k$ and $\mathbb{P}[\mathbb{1}[S_l > S_k, A_l = 0, A_k = 1] = 1] = 0$ for $l = k$. It follows from [22, Theorem 1.1] that for any $\varepsilon > 0$, with probability $1 - 2e^{-2\varepsilon^2 n}$ over the sample $(S_i, A_i)_{i=1}^n$ we have

$$\frac{\sum_{l=1}^n \sum_{k=1}^n \mathbb{1}[S_l > S_k, A_l = 0, A_k = 1]}{n^2} \leq \frac{n(n-1)}{n^2}\mathbb{P}[S_l > S_k, A_l = 0, A_k = 1] + \varepsilon \leq \mathbb{P}[S_l > S_k, A_l = 0, A_k = 1] + \varepsilon$$

as well as

$$\frac{\sum_{l=1}^n \sum_{k=1}^n \mathbb{1}[S_l > S_k, A_l = 0, A_k = 1]}{n^2} \geq \frac{n(n-1)}{n^2}\mathbb{P}[S_l > S_k, A_l = 0, A_k = 1] - \varepsilon \geq \mathbb{P}[S_l > S_k, A_l = 0, A_k = 1] - \frac{1}{n} - \varepsilon.$$

Hence, for any $\varepsilon > 0$, with $N_0 > 1/\varepsilon$, we have

$$\sum_{n=N_0}^{\infty} \mathbb{P}\left[\left|\frac{\sum_{l=1}^n \sum_{k=1}^n \mathbb{1}[S_l > S_k, A_l = 0, A_k = 1]}{n^2} - \mathbb{P}[S_l > S_k, A_l = 0, A_k = 1]\right| > 2\varepsilon\right] \leq \sum_{n=N_0}^{\infty} 2e^{-2\varepsilon^2 n} < \infty,$$

and by the Borel-Cantelli lemma $\frac{\sum_{l=1}^n \sum_{k=1}^n \mathbb{1}[S_l > S_k, A_l = 0, A_k = 1]}{n^2}$ converges to $\mathbb{P}[S_l > S_k, A_l = 0, A_k = 1]$ almost surely.

Similarly, we can show that $\frac{\sum_{l=1}^n \sum_{k=1}^n \mathbb{1}[A_l = 0, A_k = 1]}{n^2}$ converges to $\mathbb{P}[A_l = 0, A_k = 1]$ almost surely, which implies that $\frac{\sum_{l,k\in[n]} \mathbb{1}[S_l > S_k, A_l = 0, A_k = 1]}{\sum_{l,k\in[n]} \mathbb{1}[A_l = 0, A_k = 1]}$ converges to $\mathbb{P}(S_i > S_j | A_i = 0, A_j = 1)$. In the same way, we can show that $\frac{\sum_{l,k\in[n]} \mathbb{1}[S_l > S_k, A_l = 1, A_k = 0]}{\sum_{l,k\in[n]} \mathbb{1}[A_l = 1, A_k = 0]}$ converges to $\mathbb{P}(S_i > S_j | A_i = 1, A_j = 0)$ almost surely, and hence $\widehat{DP}(Q(S,A);n)$ converges to $DP(Q(S,A))$ almost surely.

B Proof of bias estimate correction under Assumption I

Proof of Eq. (6)*:*

$$\begin{aligned}\mathbb{P}(T|\hat{A}_i = a_i, \hat{A}_j = a_j) &= \sum_{s,t\in\{0,1\}} \mathbb{P}(T|A_i = s, A_j = t) \cdot \mathbb{P}(A_i = s, A_j = t|\hat{A}_i = a_i, \hat{A}_j = a_j) \\ &= \sum_{s,t\in\{0,1\}} \mathbb{P}(T|A_i = s, A_j = t) \cdot \mathbb{P}(A_i = s|\hat{A}_i = a_i) \cdot \mathbb{P}(A_j = t|\hat{A}_j = a_j)\end{aligned} \tag{12}$$

and

$$\begin{aligned}&\mathbb{P}(T|A_i = 0, A_j = 0) \cdot \mathbb{P}(A_i = 0|\hat{A}_i = 0) \cdot \mathbb{P}(A_j = 0|\hat{A}_j = 1) \\ &\quad + \mathbb{P}(T|A_i = 1, A_j = 1) \cdot \mathbb{P}(A_i = 1|\hat{A}_i = 0) \cdot \mathbb{P}(A_j = 1|\hat{A}_j = 1) \\ &\quad - \mathbb{P}(T|A_i = 0, A_j = 0) \cdot \mathbb{P}(A_i = 0|\hat{A}_i = 1) \cdot \mathbb{P}(A_j = 0|\hat{A}_j = 0) \\ &\quad - \mathbb{P}(T|A_i = 1, A_j = 1) \cdot \mathbb{P}(A_i = 1|\hat{A}_i = 1) \cdot \mathbb{P}(A_j = 1|\hat{A}_j = 0) = 0\end{aligned} \tag{13}$$

and hence

$$
\begin{aligned}
&\mathbb{P}(T|\hat{A}_i=0,\hat{A}_j=1)-\mathbb{P}(T|\hat{A}_i=1,\hat{A}_j=0)= \\
&\quad=\sum_{s\neq t\in\{0,1\}}\mathbb{P}(T|A_i=s,A_j=t)\cdot\mathbb{P}(A_i=s|\hat{A}_i=0)\cdot\mathbb{P}(A_j=t|\hat{A}_j=1) \\
&\qquad-\sum_{s\neq t\in\{0,1\}}\mathbb{P}(T|A_i=s,A_j=t)\cdot\mathbb{P}(A_i=s|\hat{A}_i=1)\cdot\mathbb{P}(A_j=t|\hat{A}_j=0) \\
&\quad=\mathbb{P}(T|A_i=0,A_j=1)\cdot\mathbb{P}(A_i=0|\hat{A}_i=0)\cdot\mathbb{P}(A_j=1|\hat{A}_j=1) \\
&\qquad+\mathbb{P}(T|A_i=1,A_j=0)\cdot\mathbb{P}(A_i=1|\hat{A}_i=0)\cdot\mathbb{P}(A_j=0|\hat{A}_j=1) \\
&\qquad-\mathbb{P}(T|A_i=0,A_j=1)\cdot\mathbb{P}(A_i=0|\hat{A}_i=1)\cdot\mathbb{P}(A_j=1|\hat{A}_j=0) \\
&\qquad-\mathbb{P}(T|A_i=1,A_j=0)\cdot\mathbb{P}(A_i=1|\hat{A}_i=1)\cdot\mathbb{P}(A_j=0|\hat{A}_j=0) \\
&\quad=(\mathbb{P}(T|A_i=0,A_j=1)-\mathbb{P}(T|A_i=1,A_j=0))\cdot\mathbb{P}(A_i=0|\hat{A}_i=0)\cdot\mathbb{P}(A_j=1|\hat{A}_j=1) \\
&\qquad+(\mathbb{P}(T|A_i=1,A_j=0)-\mathbb{P}(T|A_i=0,A_j=1))\cdot\mathbb{P}(A_i=1|\hat{A}_i=0)\cdot\mathbb{P}(A_j=0|\hat{A}_j=1) \\
&\quad=(\mathbb{P}(T|A_i=0,A_j=1)-\mathbb{P}(T|A_i=1,A_j=0))\cdot \\
&\qquad(\mathbb{P}(A_i=0|\hat{A}_i=0)\cdot\mathbb{P}(A_j=1|\hat{A}_j=1)-\mathbb{P}(A_i=1|\hat{A}_i=0)\cdot\mathbb{P}(A_j=0|\hat{A}_j=1)) \\
&\quad=(\mathbb{P}(T|A_i=0,A_j=1)-\mathbb{P}(T|A_i=1,A_j=0))\cdot \\
&\qquad\Big(\frac{\mathbb{P}(\hat{A}_i=0|A_i=0)\cdot\mathbb{P}(A_i=0)}{\mathbb{P}(\hat{A}_i=0)}\cdot\frac{\mathbb{P}(\hat{A}_j=1|A_j=1)\cdot\mathbb{P}(A_j=1)}{\mathbb{P}(\hat{A}_j=1)} \\
&\qquad\quad-\frac{\mathbb{P}(\hat{A}_i=0|A_i=1)\cdot\mathbb{P}(A_i=1)}{\mathbb{P}(\hat{A}_i=0)}\cdot\frac{\mathbb{P}(\hat{A}_j=1|A_j=0)\cdot\mathbb{P}(A_j=0)}{\mathbb{P}(\hat{A}_j=1)}\Big) \\
&\quad=(\mathbb{P}(T|A_i=0,A_j=1)-\mathbb{P}(T|A_i=1,A_j=0))\cdot\frac{m\cdot(1-m)\cdot(1-p-q)}{\mathbb{P}(\hat{A}_i=0)\cdot\mathbb{P}(\hat{A}_j=1)}.
\end{aligned}
$$

Using Bayes theorem, it is straightforward to show that

$$
\begin{aligned}
\mathbb{P}(\hat{A}_i=0)&=(1-p)\cdot(1-\beta)+q\cdot\beta, \\
\mathbb{P}(\hat{A}_j=1)&=p\cdot(1-\beta)+(1-q)\cdot\beta.
\end{aligned}
$$

This implies that

$$
\begin{aligned}
&\mathbb{P}(T|A_i=0,A_j=1)-\mathbb{P}(T|A_i=1,A_j=0)= \\
&\qquad\Big(\mathbb{P}(T|\hat{A}_i=0,\hat{A}_j=1)-\mathbb{P}(T|\hat{A}_i=1,\hat{A}_j=0)\Big)\cdot\frac{\Big((1-p)\cdot(1-\beta)+q\cdot\beta\Big)\cdot\Big(p\cdot(1-\beta)+(1-q)\cdot\beta\Big)}{\beta\cdot(1-\beta)\cdot(1-p-q)}.
\end{aligned}
$$

□

Proof of Eq. (7)*:*

$$\begin{aligned}
\mathbb{P}(\hat{A}=1|T) - P(\hat{A}=0|T) &= \frac{\mathbb{P}(T|\hat{A}=1)\cdot\mathbb{P}(\hat{A}=1) - \mathbb{P}(T|\hat{A}=0)\cdot\mathbb{P}(\hat{A}=0)}{\mathbb{P}(T)} \\
&= \frac{\mathbb{P}(\hat{A}=1)\cdot(\mathbb{P}(T|A=0)\cdot\mathbb{P}(A=0|\hat{A}=1) + \mathbb{P}(T|A=1)\cdot\mathbb{P}(A=1|\hat{A}=1))}{\mathbb{P}(T)} \\
&\quad - \frac{\mathbb{P}(\hat{A}=0)\cdot(\mathbb{P}(T|A=0)\cdot\mathbb{P}(A=0|\hat{A}=0) + \mathbb{P}(T|A=1)\cdot\mathbb{P}(A=1|\hat{A}=0)}{\mathbb{P}(T)} \\
&= \frac{\mathbb{P}(T|A=0)\cdot\mathbb{P}(\hat{A}=1|A=0)\cdot\mathbb{P}(A=0) + \mathbb{P}(T|A=1)\cdot\mathbb{P}(\hat{A}=1|A=1)\cdot\mathbb{P}(A=1)}{\mathbb{P}(T)} \\
&\quad - \frac{\mathbb{P}(T|A=0)\cdot\mathbb{P}(\hat{A}=0|A=0)\cdot\mathbb{P}(A=0) + \mathbb{P}(T|A=1)\cdot\mathbb{P}(\hat{A}=0|A=1)\cdot\mathbb{P}(A=1)}{\mathbb{P}(T)} \\
&= \mathbb{P}(A=0|T)\cdot\mathbb{P}(\hat{A}=1|A=0) + \mathbb{P}(A=1|T)\cdot\mathbb{P}(\hat{A}=1|A=1) \\
&\quad - \mathbb{P}(A=0|T)\cdot\mathbb{P}(\hat{A}=0|A=0) - \mathbb{P}(A=1|T)\cdot\mathbb{P}(\hat{A}=0|A=1) \\
&= (1-2q)\cdot\mathbb{P}(A=1|T) - (1-2p)\cdot P(A=0|T) \\
&= 2\mathbb{P}(A=1|T)(1-p-q) - 1 + 2p
\end{aligned}$$

This entails

$$\mathbb{P}(A=1|T) - \mathbb{P}(A=0|T) = 2\mathbb{P}(A=1|T) - 1 = \frac{\mathbb{P}(\hat{A}=1|T) - P(\hat{A}=0|T) - p + q}{1-p-q}.$$

□

Proof of Eq. (8)*:*

$$\begin{aligned}
\mathbb{P}(\hat{A}=1|T) - \mathbb{P}(\hat{A}=1) &= \mathbb{P}(\hat{A}=1|T,A=1)\cdot\mathbb{P}(A=1|T) + \mathbb{P}(\hat{A}=1|T,A=0)\cdot\mathbb{P}(A=0|T) - \mathbb{P}(\hat{A}=1) \\
&= \mathbb{P}(\hat{A}=1|A=1)\cdot\mathbb{P}(A=1|T) + \mathbb{P}(\hat{A}=1|A=0)\cdot\mathbb{P}(A=0|T) - \mathbb{P}(\hat{A}=1) \\
&= (1-q)\cdot\mathbb{P}(A=1|T) + p\cdot\mathbb{P}(A=0|T) - \mathbb{P}(\hat{A}=1|A=1)\cdot\mathbb{P}(A=1) - \mathbb{P}(\hat{A}=1|A=1)\cdot\mathbb{P}(A=0)
\end{aligned}$$

After replacing $\mathbb{P}(A=0|T)$ with $1-\mathbb{P}(A=1|T)$, we get

$$\mathbb{P}(A=1|T) - \mathbb{P}(A=1) = \frac{\mathbb{P}(\hat{A}=1|T) - \mathbb{P}(\hat{A}=1)}{1-p-q}.$$

□

C Proof of bias estimate correction under Assumption II

Proof of Eq. (9)*:*

Exchanging the role of A and $\hat{A}$, we obtain equalities analogous to (12) and (13), respectively. Then,

$$\begin{aligned}
&\mathbb{P}(T|A_i=0,A_j=1)-\mathbb{P}(T|A_i=1,A_j=0)= \\
&\quad=\sum_{s\neq t\in\{0,1\}}\mathbb{P}(T|\hat{A}_i=s,\hat{A}_j=t)\cdot\mathbb{P}[\hat{A}_i=s|A_i=0]\cdot\mathbb{P}(\hat{A}_j=t|A_j=1) \\
&\qquad-\sum_{s\neq t\in\{0,1\}}\mathbb{P}(T|\hat{A}_i=s,\hat{A}_j=t)\cdot\mathbb{P}(\hat{A}_i=s|A_i=1)\cdot\mathbb{P}(\hat{A}_j=t|A_j=0) \\
&\quad=\mathbb{P}(T|\hat{A}_i=0,\hat{A}_j=1)\cdot\mathbb{P}(\hat{A}_i=0|A_i=0)\cdot\mathbb{P}(\hat{A}_j=1|A_j=1) \\
&\qquad+\mathbb{P}(T|\hat{A}_i=1,\hat{A}_j=0)\cdot\mathbb{P}(\hat{A}_i=1|A_i=0)\cdot\mathbb{P}(\hat{A}_j=0|A_j=1) \\
&\qquad-\mathbb{P}(T|\hat{A}_i=0,\hat{A}_j=1)\cdot\mathbb{P}(\hat{A}_i=0|A_i=1)\cdot\mathbb{P}(\hat{A}_j=1|A_j=0) \\
&\qquad-\mathbb{P}(T|\hat{A}_i=1,\hat{A}_j=0)\cdot\mathbb{P}(\hat{A}_i=1|A_i=1)\cdot\mathbb{P}(\hat{A}_j=0|A_j=0) \\
&\quad=(\mathbb{P}(T|\hat{A}_i=0,\hat{A}_j=1)-\mathbb{P}(T|\hat{A}_i=1,\hat{A}_j=0))\cdot\mathbb{P}(\hat{A}_i=0|A_i=0)\cdot\mathbb{P}(\hat{A}_j=1|A_j=1) \\
&\qquad+(\mathbb{P}(T|\hat{A}_i=1,\hat{A}_j=0)-\mathbb{P}(T|\hat{A}_i=0,\hat{A}_j=1))\cdot\mathbb{P}(\hat{A}_i=1|A_i=0)\cdot\mathbb{P}(\hat{A}_j=0|A_j=1) \\
&\quad=(\mathbb{P}(T|\hat{A}_i=0,\hat{A}_j=1)-\mathbb{P}(T|\hat{A}_i=1,\hat{A}_j=0))\cdot \\
&\qquad(\mathbb{P}(\hat{A}_i=0|A_i=0)\cdot\mathbb{P}(\hat{A}_j=1|A_j=1)-\mathbb{P}(\hat{A}_i=1|A_i=0)\cdot\mathbb{P}(\hat{A}_j=0|A_j=1)) \\
&\quad=(\mathbb{P}(T|\hat{A}_i=0,\hat{A}_j=1)-\mathbb{P}(T|\hat{A}_i=1,\hat{A}_j=0))\cdot(1-p-q).
\end{aligned}$$

□

Proof of Eq. (10)*:*

$$\begin{aligned}
&\mathbb{P}(A=1|T)-\mathbb{P}(A=0|T)=2\cdot\mathbb{P}(A=1|T)-1 \\
&\quad=2\cdot\mathbb{P}(A=1|T,\hat{A}=1)\cdot\mathbb{P}(\hat{A}=1|T)+2\cdot\mathbb{P}(A=1|T,\hat{A}=0)\cdot\mathbb{P}(\hat{A}=0|T)-1 \\
&\quad=2\cdot\mathbb{P}(\hat{A}=1|T)\cdot\Big(\mathbb{P}(A=1|\hat{A}=1)-\mathbb{P}(A=1|\hat{A}=0)\Big)+2\cdot\mathbb{P}(A=1|\hat{A}=0)-1 \\
&\quad=\Big(\mathbb{P}(\hat{A}=1|T)-\mathbb{P}(\hat{A}=0|T)+1\Big)\cdot\left(\frac{\mathbb{P}(\hat{A}=1|A=1)\cdot\mathbb{P}(A=1)}{\mathbb{P}(\hat{A}=1)}-\frac{\mathbb{P}(\hat{A}=0|A=1)\cdot\mathbb{P}(A=1)}{\mathbb{P}(\hat{A}=0)}\right) \\
&\qquad+2\cdot\frac{\mathbb{P}(\hat{A}=0|A=1)\cdot\mathbb{P}(A=1)}{\mathbb{P}(\hat{A}=0)}-1 \\
&\quad=\Big(\mathbb{P}(\hat{A}=1|T)-\mathbb{P}(\hat{A}=0|T)+1\Big)\cdot\left(\frac{(1-q)\cdot\beta}{(1-q)\cdot\beta+p\cdot(1-\beta)}-\frac{q\cdot\beta}{q\cdot\beta+(1-p)\cdot(1-\beta)}\right) \\
&\qquad+\frac{2\cdot q\cdot\beta}{q\cdot\beta+(1-p)\cdot(1-\beta)}-1
\end{aligned}$$

□

Proof of Eq. (11)*:*

$$\begin{aligned}
&\mathbb{P}(A=1|T)-\mathbb{P}(A=1)= \\
&\quad\mathbb{P}(A=1|T,\hat{A}=1)\cdot\mathbb{P}(\hat{A}=1|T)+\mathbb{P}(A=1|T,\hat{A}=0)\cdot(1-\mathbb{P}(\hat{A}=1|T)) \\
&\qquad-\mathbb{P}(A=1|\hat{A}=1)\cdot\mathbb{P}(\hat{A}=1)-\mathbb{P}(A=1|\hat{A}=0)\cdot\mathbb{P}(\hat{A}=0) \\
&\quad=\mathbb{P}(A=1|\hat{A}=1)\cdot\mathbb{P}(\hat{A}=1|T)+\mathbb{P}(A=1|\hat{A}=0)\cdot(1-\mathbb{P}(\hat{A}=1|T)) \\
&\qquad-\mathbb{P}(A=1|\hat{A}=1)\cdot\mathbb{P}(\hat{A}=1)-\mathbb{P}(A=1|\hat{A}=0)\cdot\mathbb{P}(\hat{A}=0) \\
&\quad=(\mathbb{P}(\hat{A}=1|T)-\mathbb{P}(\hat{A}=1))\cdot(\mathbb{P}(A=1|\hat{A}=1)-\mathbb{P}(A=1|\hat{A}=0)) \\
&\quad=(\mathbb{P}(\hat{A}=1|T)-\mathbb{P}(\hat{A}=1))\cdot\left(\frac{\mathbb{P}(\hat{A}=1|A=1)\cdot\mathbb{P}(A=1)}{\mathbb{P}(\hat{A}=1)}-\frac{\mathbb{P}(\hat{A}=0|A=1)\cdot\mathbb{P}(A=1)}{1-\mathbb{P}(\hat{A}=1)}\right) \\
&\quad=(\mathbb{P}(\hat{A}=1|T)-\mathbb{P}(\hat{A}=1))\cdot\left(\frac{(1-q)\cdot\beta}{(1-q)\cdot\beta+p\cdot(1-\beta)}-\frac{q\cdot\beta}{q\cdot\beta+(1-p)\cdot(1-\beta)}\right)
\end{aligned}$$

□

D Summary of the real-world datasets used in our study

Table 3: Summary of the datasets used in our study.

Dataset	Size	Sensitive attribute	Sensitive value	%G1	Avg. DP	Avg. Exp	Avg. rND
Goodreads authors	~22k	Gender	Female	44%	−0.288	−0.083	0.144
FIFA Players	~2.9k	Nationality	Germany	42%	−0.237	−0.112	0.164
COMPAS	~61k	Ethnicity	African-American	36%	−0.004	−0.112	0.042
IBM HR Analytics	~2k	Marital status	Single	31%	0.037	−0.366	0.124