On Discrimination Discovery and Removal in Ranked Data using Causal Graph

Yongkai Wu
University of Arkansas
yw009@uark.edu

Lu Zhang
University of Arkansas
lz006@uark.edu

Xintao Wu
University of Arkansas
xintaowu@uark.edu

ABSTRACT

Predictive models learned from historical data are widely used to help companies and organizations make decisions. However, they may digitally unfairly treat unwanted groups, raising concerns about fairness and discrimination. In this paper, we study the fairness-aware ranking problem which aims to discover discrimination in ranked datasets and reconstruct the fair ranking. Existing methods in fairness-aware ranking are mainly based on statistical parity that cannot measure the true discriminatory effect since discrimination is causal. On the other hand, existing methods in causal-based anti-discrimination learning focus on classification problems and cannot be directly applied to handle the ranked data. To address these limitations, we propose to map the rank position to a continuous score variable that represents the qualification of the candidates. Then, we build a causal graph that consists of both the discrete profile attributes and the continuous score. The path-specific effect technique is extended to the mixed-variable causal graph to identify both direct and indirect discrimination. The relationship between the path-specific effects for the ranked data and those for the binary decision is theoretically analyzed. Finally, algorithms for discovering and removing discrimination from a ranked dataset are developed. Experiments using the real-world dataset show the effectiveness of our approaches.

CCS CONCEPTS

• **Mathematics of computing** → **Causal networks**; • **Computing methodologies** → **Ranking**; • **Applied computing** → *Law, social and behavioral sciences*;

KEYWORDS

Discrimination-aware machine learning; fair ranking; causal graph; direct and indirect discrimination

1 INTRODUCTION

Discrimination-aware machine learning which aims to construct discrimination-free machine learning models has been an active research area in the recent years. Many works have been conducted to achieve this goal by detecting and removing discrimination/biases from the historical training data [8, 10, 15, 24, 36], or from the constructed machine learning models [7, 16, 17, 31]. However, most works focus on the classification models built for categorical decisions, especially binary decisions. In this paper, we investigate discrimination in ranking models, which are another widely used machine learning models adopted by search engines, recommendation systems, and auction systems, etc. To be more specific, we study the discrimination discovery and removal from the ranked data. A ranked dataset is a combination of the candidate profiles with the permutation of the candidates as the decision. Fairness concerns are raised for the ranking models since biases and discrimination can also be introduced into the ranking.

ID	u1	u2	u3	u4	u5	u6	u7	u8	u9	u10
Race	1	1	1	1	1	0	0	0	0	0
Zip Code	1	1	1	1	1	1	0	0	1	0
Interview	1	2	2	4	2	5	4	4	3	2
Edu	1	2	1	2	4	5	4	5	3	5

Ranker#1	●	●	●	●	●	■	■	■	■	■
Ranker#2	■	●	●	■	■	●	■	●	■	●
ω	1	2	3	4	5	6	7	8	9	10

Figure 1: A toy example of dataset and ranking results produced by two rankers. Blue squares represent the favorable group and red circles represent unfavorable group.

Existing methods [30, 32] for studying the discrimination discovery and removal from ranked data are mainly based on statistical parity, which means that the demographics of individuals in any prefix of the ranking are identical to the demographics of the whole population. However, it has already been shown in classification that statistical parity does not take into account the fact that part of discrimination is explainable by some non-protected attributes and hence cannot accurately measure discrimination [9]. We believe that this observation also holds in the ranked data. Let's consider a toy example of ranked data for a company recruiting system shown in Figure 1. The data contains four profile attributes: *race* (C), *zip code* (Z), *education* (E), *interview result* (I), where *race* is the protected attribute with a favorable group ($C = 1$) and an unfavorable group ($C = 0$), and *education* and *interview result* are the objective requirements of getting the job. Assume that there are two rankers, both of which compute the qualification scores to produce the rankings. The first ranker, denoted by Ranker#1, produces qualification

scores as an equal-weighted linear combination of two attributes *education* and *interview result*. Intuitively, Ranker#1 produces a fair ranking since it purely depends on two objective attributes. However, as can be seen, the ranking results do not satisfy statistical parity. On the basis of Ranker#1, the second ranker, Ranker#2, further gives a bonus score of 2 for the favorable group (i.e., $C = 1$). The usage of the protected attribute explicitly results in the unfavorable treatment to the protected group ($C = 0$). Nevertheless, the ranking results satisfy statistical parity as two race groups are well-mixed in equal proportion. This example shows that statistical parity may produce misleading conclusions regarding discrimination.

To address the limitation of the statistical parity-based methods, the causal graph-based discrimination detection and removal methods have been recently proposed by Zhang et al. [36]. It shows that the correlation between the protected attribute and the decision is a nonlinear combination of the direct discrimination, the indirect discrimination, as well as the explainable effect. The path-specific effect technique has been used to capture the causal effects passing through different paths. However, this work focuses on binary classification. In ranking systems, the decisions are given in term of a permutation of a series of unique, concatenating integers which cannot be treated as regular random variables. This means that causal graphs cannot be built in traditional ways. Thus, the methods in [36] cannot be applied directly to deal with ranked data.

In this paper, we employ the causal graph to solve the fair ranking problem by adopting a continuous variable called *score* instead of the ranking positions to represent the qualifications of individuals in the rank. We use the Bradley-Terry model [6] to obtain a reasonable mapping from ranking positions to scores. We then construct the causal graph from the individuals' profiles and scores, a mix of categorical and continuous data. Traditional causal graph construction and inference are limited to the single data-type situations where the variables are all discrete (e.g., causal Bayesian networks) or all continuous (e.g., linear Gaussian models). To address this challenge, we associate the score with a set of Conditional Gaussian (CG) distributions instead of the Conditional Probability Table (CPT). Then, we extend the path-specific effect technique to our mixed-variable causal graph for capturing direct and indirect discrimination. We also derive a relationship between the path-specific effects for the ranked data and those for the binary decision, assuming that binary decision is obtained based on certain cut-off point imposed on the ranking. Algorithms for detecting discrimination in the causal graph, as well as for removing rank biases from the data are developed. Finally, we conduct experiments using the real-world dataset to show the effectiveness of our methods. The results show that our methods can correctly detect and remove both direct and indirect discrimination with relatively small data utility loss, while the statistical parity based methods neither correctly identify discrimination nor successfully mitigate discrimination.

2 PRELIMINARIES

Throughout the paper, attributes are denoted by an uppercase letter, e.g.: X; sets of attributes are denoted by a bold uppercase letter, e.g.: $\mathbf{X}$. A value of a certain attribute is denoted by a lowercase letter, e.g.: x; a set of values of a set attributes is denoted by a bold lowercase letter, e.g.: $\mathbf{x}$. The domain space of an attribute is denoted by $\mathfrak{X}_X$.

The domain space of an attribute set is a Cartesian product of the domain spaces of its elements, denoted by $\mathfrak{X}_{\mathbf{X}} = \prod_{X \in \mathbf{X}} \mathfrak{X}_X$.

A causal graph [25] is a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{A})$ where $\mathbf{V}$ is a set of nodes and $\mathbf{A}$ is a set of edges. Each node in the graph represents an attribute. Each edge, denoted by an arrow "→" pointing from the cause to the effect, represents the direct causal relationship. The path that traces arrows directed from one node X to another node Y is called the causal path from X to Y. For any node X, its parents are denoted by $Pa(X)$, and its children are denoted by $Ch(X)$. Usually, the *local Markov condition* is assumed to be satisfied, which means that each node is independent of its non-descendants conditional on all its parents. Each child-parent family in the graph is associated with a deterministic function

$$x = f_X(Pa(X), \varepsilon_X), \quad X \in \mathbf{V},$$

where ε_X is an arbitrarily distributed random disturbance. This functional characterization of the child-parent relationship leads to the conditional probability distribution that characterizes the graph, i.e., $P(x|Pa(X))$. When all variables are discrete, $P(x|Pa(X))$ is denoted by a conditional probability table (CPT).

Inferring causal effects in the causal graph is performed through interventions, which fixes the values a set of variables $\mathbf{X} \subseteq \mathbf{V}$ to constants $\mathbf{x}$. Symbolically, it is formalized as $do(\mathbf{X} = \mathbf{x})$ or simply $do(\mathbf{x})$. The post-intervention distribution of all other attributes $\mathbf{Y} = \mathbf{V}\backslash\mathbf{X}$, denoted by $P(\mathbf{Y} = \mathbf{y}|do(\mathbf{X} = \mathbf{x}))$ or simply $P(\mathbf{y}|do(\mathbf{x}))$, can be calculated using the truncated factorization formula [18]

$$P(\mathbf{y}|do(\mathbf{x})) = \prod_{Y \in \mathbf{Y}} P(y|Pa(Y))\delta_{\mathbf{X}=\mathbf{x}}, \tag{1}$$

where $\delta_{\mathbf{X}=\mathbf{x}}$ means assigning attributes in $\mathbf{X}$ involved in the term ahead with the corresponding values in $\mathbf{x}$. As a result, the *total causal effect* of $\mathbf{X}$ on $\mathbf{Y}$ is assessed by comparing the difference between the post-intervention distributions under two different interventions $do(\mathbf{x}')$ and $do(\mathbf{x}'')$. A common measure of the total causal effect is the expected difference as shown in Definition 1. Note that the total causal effect measures the effect of the intervention that is transmitted along all causal paths from $\mathbf{X}$ to $\mathbf{Y}$.

DEFINITION 1 (TOTAL CAUSAL EFFECT). *Given a causal graph $\mathcal{G} = (\mathbf{V}, \mathbf{A})$ and two disjoint sets of variables $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$, the total causal effect of $\mathbf{X}$ on $\mathbf{Y}$ in terms of two interventions $do(\mathbf{x}')$ and $do(\mathbf{x}'')$, denoted by $TE(\mathbf{x}', \mathbf{x}'')$, is given by*

$$TE(\mathbf{x}', \mathbf{x}'') = \mathbb{E}\left[\mathbf{Y}|do(\mathbf{x}')\right] - \mathbb{E}\left[\mathbf{Y}|do(\mathbf{x}'')\right],$$

where $\mathbb{E}[\cdot]$ is the expectation.

As an extension to the total causal effect, Avin et al. [3] proposed the *path-specific effect* that measures the causal effect where the intervention's effect is transmitted along a subset of the causal paths from $\mathbf{X}$ to $\mathbf{Y}$. Denote a subset of causal paths by π, and denote by $P(\mathbf{Y}|do(\mathbf{x}'|_\pi))$ the post-intervention distribution of $\mathbf{Y}$ with the intervention's effect transmitted along π. Based on that, the π-specific effect is given by Definition 2.

DEFINITION 2 (PATH-SPECIFIC EFFECT). *Given a causal graph $\mathcal{G} = (\mathbf{V}, \mathbf{A})$, two disjoint sets of variables $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$, and a subset of causal paths π, the π-specific effect of $\mathbf{X}$ on $\mathbf{Y}$ in terms of two interventions $do(\mathbf{x}')$ and $do(\mathbf{x}'')$, denoted by $SE_\pi(\mathbf{x}', \mathbf{x}'')$, is given by*

$$SE_\pi(\mathbf{x}', \mathbf{x}'') = \mathbb{E}\left[\mathbf{Y}|do(\mathbf{x}'|_\pi)\right] - \mathbb{E}\left[\mathbf{Y}|do(\mathbf{x}'')\right].$$

In [3], it is pointed out the condition under which the path-specific effect can be estimated from the observed data, known as identifiability of the path-specific effects. In [27], Shpitser et al. gave the method for calculating the identifiable path-specific effect. How to deal with the unidentifiable situation is discussed in [34]. These strategies are readily to be applied to our methods.

3 MODELING DIRECT AND INDIRECT DISCRIMINATION IN RANKED DATA

In this section, we study how to model direct and indirect discrimination in a ranked dataset as the causal effect. We consider a ranked dataset $\mathcal{D}$ consisting of N individuals with a protected attribute C, several non-protected attributes $\mathbf{Z} = \{Z_1, \cdots, Z_j, \cdots\}$, and a rank permutation π as the decision. There is a subset of attributes $\mathbf{R} \subseteq \mathbf{Z}$ that may cause indirect discrimination, referred to as the *redlining attributes*. We assume all attributes are categorical. We further make two reasonable assumptions: 1) the protected attribute C has no parent; and 2) the score S has no child. The two assumptions are to make our theoretical results more concise and can be easily relaxed.

3.1 Building Causal Graph for Ranked Data

A rank permutation is a series of unique, concatenating integers that cannot be treated as normal categorical random variables. In data science, a number of models [23] are proposed to map the ranking positions in a ranked data to the continuous scores. In this paper we use the Bradley-Terry model [6] but the logic also applies to other models. The comparison of the performance of different models is beyond the scope of the paper and is left for future work.

A Bradley-Terry model $\mathcal{M}$ assigns each individual i a score s_i ($s_i \in \mathbb{R}$) to indicate the qualification preference of individual. Generally, a larger score represents a better qualification. The difference between the scores of two individuals i, j corresponds to the log-odds of the probability $p_{i,j}$ that individual i is ranked before individual j in the rank, i.e.,

$$s_i - s_j = \log \frac{p_{ij}}{1 - p_{ij}}.$$

Equivalently, solving for p_{ij} yields

$$p_{ij} = \frac{e^{s_i}}{e^{s_i} + e^{s_j}}.$$

On the other hand, the probability of any rank permutation ω given a Bradley-Terry model $\mathcal{M}$ is proportional to the product of the probability $p_{i,j}$ of all preference pairs subject to ω, i.e.,

$$P(\omega|\mathcal{M}) \propto \prod_{(i,j):\omega_i < \omega_j} p_{ij},$$

where ω_i, ω_j are the ranking positions of individuals i, j. Thus, the logarithm likelihood of the Bradley-Terry model $\mathcal{M}$ given the observed rank permutation ω is given by $\mathcal{L}(\mathcal{M}|\omega) = -\log P(\omega|\mathcal{M})$. As a result, the optimal Bradley-Terry model that best fits the observed rank permutation ω can be obtained by minimizing $\mathcal{L}(\mathcal{M}|\omega)$ as the loss function. Wu et al. [29] proved that the loss function is convex and could be efficiently optimized with gradient descent.

After obtaining the score S using the Bradley-Terry model, we build a causal graph for variables C, $\mathbf{Z}$ and S. We first adopt the PC-algorithm for learning the structure of the causal graph. Since

there exist both discrete and continuous variables, different conditional independence testing methods can be adopted, such as chi-square test for discrete variables, partial correlation matrix for continuous variables, and conditional Gaussian likelihood ratio test for mixed variables. Then, for parameterizing the causal graph, we treat discrete and continuous variables in different ways. For discrete variables C and $\mathbf{Z}$ (we can extend our method to the situation where some profile attributes are continuous), each of them is associated with a Conditional Probability Table (CPT). The conditional probabilities can be estimated from data using standard statistical estimation techniques (like the maximum likelihood estimation). For continuous score S, it is associated with the Conditional Gaussian (CG) distributions instead of the CPT. Let $\mathbf{Q} = Pa(S) \backslash \{C\}$. For each value assignment $c, \mathbf{q}$ of parents of S, there is a CG distribution whose mean and variance are based on $c, \mathbf{q}$. Thus, the CG distribution of S is given by

$$P(s|c, \mathbf{q}) = \mathcal{N}(\mu_{c,\mathbf{q}}, \sigma_{c,\mathbf{q}}^2).$$

Finally, we fit each CG distribution $\mathcal{N}(\mu_{c,\mathbf{q}}, \sigma_{c,\mathbf{q}}^2)$ to the scores of all candidates with $C = c$ and $\mathbf{Q} = \mathbf{q}$ using standard statistical estimation techniques.

As an example, Figure 2 shows a causal graph of the toy example presented in the Introduction. Each of C, Z, E, I is associated with a CPT representing the conditional probability given the parents, and S is associated with a set of CG distribution where the mean and the variance are based on its parents, the other four variables.
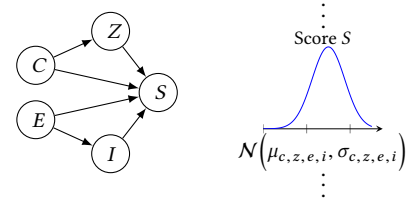

Figure 2: Causal graph of the toy example involving: *race* (C), *zip code* (Z), *education* (E), *interview result* (I), and *score* (S).

3.2 Quantitative Measurement

Now we show how direct and indirect discrimination in a ranked data can be quantitatively measured based on the causal graph we build. It is known that discrimination is a causal effect of the protected attribute on the decision. We first give the quantitative measure of the total causal effect of protected attribute C on score S as shown in Theorem 3.1.

THEOREM 3.1. *The total causal effect is given by*

$$TE(c^+, c^-) = \sum_{\mathbf{q} \in \mathfrak{X}_{\mathbf{Q}}} \left(\mu_{c^+, \mathbf{q}} P(\mathbf{q}|c^+) - \mu_{c^-, \mathbf{q}} P(\mathbf{q}|c^-) \right) \qquad (2)$$

PROOF. According to Definition 1, total causal effect is given by

$$TE(c^+, c^-) = \mathbb{E}\left[S|do(c^+)\right] - \mathbb{E}\left[S|do(c^-)\right]$$
$$= \int s \cdot P(s|do(c^+))ds - \int s \cdot P(s|do(c^-))ds.$$

According to Eq. (1), we have

$$P(s|do(c^+)) = \sum_{\mathbf{z} \in \mathfrak{X}_{\mathbf{Z}}} P(s, \mathbf{z}|do(c^+))$$

$$= \sum_{\mathbf{z} \in \mathfrak{X}_{\mathbf{Z}}} P(s|c^+, \mathbf{q}) \prod_{Z_j \in \mathbf{Z}} P(z_j|Pa(Z_j))\delta_{C=c^+}.$$

It can be shown that

$$\prod_{Z_j \in \mathbf{Z}} P(z_j|Pa(Z_j))\delta_{C=c^+} = P(\mathbf{z}|c^+). \tag{3}$$

In fact, if we sort all nodes in $\mathbf{Z}$ according to the topological ordering as $\{Z_1, \cdots, Z_j, \cdots\}$, we can see that all parents of each node Z_j are before it in the ordering. In addition, since C has no parent, it must be Z_j's non-descendant; since E has no child, it cannot be Z_j's parent. Thus, based on the local Markov condition, we have $P(z_j|Pa(Z_j)) = P(z_j|c^+, z_1, \cdots, z_{j-1})$. According to the chain rule we obtain $P(\mathbf{z}|c^+)$. Thus, it follows that

$$P(s|do(c^+)) = \sum_{\mathbf{z} \in \mathfrak{X}_{\mathbf{Z}}} P(s|c^+, \mathbf{q})P(\mathbf{z}|c^+)$$

$$= \sum_{\mathbf{q} \in \mathfrak{X}_{\mathbf{Q}}} P(s|c^+, \mathbf{q}) \sum_{\mathbf{Z}\setminus\mathbf{Q}} P(\mathbf{z}|c^+) = \sum_{\mathbf{q} \in \mathfrak{X}_{\mathbf{Q}}} P(s|c^+, \mathbf{q})P(\mathbf{q}|c^+).$$

As a result, we have

$$\int s \cdot P(s|do(c^+))ds = \int s \cdot \sum_{\mathbf{q} \in \mathfrak{X}_{\mathbf{Q}}} P(s|c^+, \mathbf{q})P(\mathbf{q}|c^+)ds$$

$$= \sum_{\mathbf{q} \in \mathfrak{X}_{\mathbf{Q}}} P(\mathbf{q}|c^+) \int sP(s|c^+, \mathbf{q})ds = \sum_{\mathbf{q} \in \mathfrak{X}_{\mathbf{Q}}} \mu_{c^+,\mathbf{q}} P(\mathbf{q}|c^+).$$

Hence, the theorem is proven. □

In [36], the authors show that in the single-type causal graph, total causal effect generally cannot correctly measure either direct discrimination or indirect discrimination, which should be modeled as the path-specific effects. By adopting similar strategy, we capture direct discrimination by the causal effect transmitted via the direct edge from C to S, and capture indirect discrimination by the causal effect transmitted via the paths that pass through redlining attributes. Formally, define π_d as the path set that contains only $C \rightarrow S$, and define π_i as the path set that contains all causal paths which are from C to S and pass through $\mathbf{R}$. Then, direct discrimination can be captured by the π_d-specific effect $SE_{\pi_d}(\cdot)$, and indirect discrimination can be captured by the π_i-specific effect $SE_{\pi_i}(\cdot)$. We extend the method in [36] for computing the path-specific effect from data to our mixed-variable causal graph for computing $SE_{\pi_d}(\cdot)$ and $SE_{\pi_i}(\cdot)$. The results are shown in Theorem 3.2.

THEOREM 3.2. *The π_d-specific effect $SE_{\pi_d}(c^+, c^-)$ is given by*

$$SE_{\pi_d}(c^+, c^-) = \sum_{\mathbf{q} \in \mathfrak{X}_{\mathbf{Q}}} \left(\mu_{c^+,\mathbf{q}} - \mu_{c^-,\mathbf{q}}\right) P(\mathbf{q}|c^-), \tag{4}$$

The π_i-specific effect $SE_{\pi_i}(c^+, c^-)$ is given by

$$SE_{\pi_i}(c^+, c^-) = \sum_{\mathbf{z} \in \mathfrak{X}_{\mathbf{Z}}} \Big(\mu_{c^-,\mathbf{q}} \prod_{G \in \mathbf{V}_{\pi_i}} P(g|c^+, Pa(G)\setminus\{C\})$$

$$\prod_{H \in \bar{\mathbf{V}}_{\pi_i}} P(g|c^-, Pa(G)\setminus\{C\}) \prod_{O \in \mathbf{Z}\setminus Ch(C)} P(o|Pa(O))\Big) - \sum_{\mathbf{q} \in \mathfrak{X}_{\mathbf{Q}}} \left(\mu_{c^-,\mathbf{q}}P(\mathbf{q}|c^-)\right),$$

$$\tag{5}$$

where $\mathbf{V}_{\pi_i}$ and $\bar{\mathbf{V}}_{\pi_i}$ is obtained by dividing C's children except S based on the above method. Eq. (5) can be simplified to

$$SE_{\pi_i}(c^+, c^-) = \sum_{\mathbf{q} \in \mathfrak{X}_{\mathbf{Q}}} \mu_{c^-,\mathbf{q}} \left(P(\mathbf{q}|c^+) - P(\mathbf{q}|c^-)\right) \tag{6}$$

if π_i contains all causal paths from C to S except the direct edge $C \rightarrow S$.

PROOF. For the π_d-specific effect, according to Definition 2, we have

$$SE_{\pi_d} = \mathbb{E}\left[S|do(\mathbf{c}^+|_{\pi_d})\right] - \mathbb{E}\left[S|do(\mathbf{c}^-)\right]$$

$$= \int s \cdot P(s|do(c^+|_{\pi_d}))ds - \int s \cdot P(s|do(c^-))ds.$$

In the above equation, $P(s|do(c^-))$ can be computed according to the truncated factorization formula (1). To compute $P(s|do(c^+|_{\pi_d}))$, we follow the steps in [27]. First, express $P(s|do(c^+|_{\pi_d}))$ as the truncated factorization formula. Then, divide the children of C into two disjoint sets $\mathbf{V}_{\pi_d}$ and $\bar{\mathbf{V}}_{\pi_d}$. Let $\mathbf{V}_{\pi_d}$ contains C's each child V where edge $C \rightarrow V$ is a segment of a path in π_d; let $\bar{\mathbf{V}}_{\pi_d}$ contains C's each child V where either V is not included in any path from C to S, or edge $C \rightarrow V$ is a segment of a path not in π_d. Finally, replace values of C with c^+ for the terms corresponding to nodes in $\mathbf{V}_{\pi}$, and replace values of C with c^- for the terms corresponding to nodes in $\bar{\mathbf{V}}_{\pi_d}$.

Following the above procedure, we obtain

$$P(s|do(c^+|_{\pi_d})) = \sum_{\mathbf{z} \in \mathfrak{X}_{\mathbf{Z}}} P(s|c^+, \mathbf{q}) \prod_{Z_i \in \mathbf{Z}} P(z_i|Pa(Z_i))\delta_{C=c^-}.$$

By using Eq. (3), it follows that

$$P(s|do(c^+|_{\pi_d})) = \sum_{\mathbf{q} \in \mathfrak{X}_{\mathbf{Q}}} P(s|c^+, \mathbf{q})P(\mathbf{q}|c^-),$$

which leads to Eq. (4) in the theorem.

For the π_i-specific effect, following the above procedure similarly we can obtain

$$P(s|do(c^+|_{\pi_i})) = \sum_{\mathbf{z} \in \mathfrak{X}_{\mathbf{Z}}} \Big(P(s|c^-, \mathbf{q}) \prod_{G \in \mathbf{V}_{\pi_i}} P(g|c^+, Qa(G))$$

$$\prod_{H \in \bar{\mathbf{V}}_{\pi_i}} P(g|c^-, Pa(G)\setminus\{C\}) \prod_{O \in \mathbf{Z}\setminus Ch(C)} P(o|Pa(O))\Big),$$

which leads to Eq. (5). If π_i contains all causal paths from C to S except the direct edge, it means that $\mathbf{V}_{\pi_i} = Ch(C)\setminus\{S\}$, and $\bar{\mathbf{V}}_{\pi_i} = \emptyset$. Thus, it follows that

$$P(s|do(c^+|_{\pi_i})) = \sum_{\mathbf{z} \in \mathfrak{X}_{\mathbf{Z}}} P(s|c^-, \mathbf{q}) \prod_{Z \in \mathbf{Z}} P(z|Pa(Z)\setminus\{C\})\delta_{C=c^+}$$

$$= \sum_{\mathbf{q} \in \mathfrak{X}_{\mathbf{Q}}} P(s|c^-, \mathbf{q})P(\mathbf{q}|c^+),$$

which leads to Eq. (6). Hence, the theorem is proven. □

Theorems 3.1 and 3.2 present the quantitative measurement of the total causal effect as well as the π_d and π_i-specific effects. The following proposition reveals the relationship among $TE(\cdot)$, $SE_{\pi_d}(\cdot)$ and $SE_{\pi_i}(\cdot)$. It shows that the indirect (discriminatory) effect is equal to the total causal effect plus the "reversed" direct (discriminatory) effect.

PROPOSITION 3.3. *If π_i contains all causal paths from C to S except the direct edge $C \to S$, we have*

$$SE_{\pi_i}(c^+, c^-) = TE(c^+, c^-) + SE_{\pi_d}(c^-, c^+).$$

PROOF. The proof can be directly obtained from Eq. (2) and (6). □

3.3 Relationship between Ranking and Binary Decision

In the earlier work [36], we have derived the π_d and π_i-specific effects of the protected attribute C on a binary decision attribute E with positive decision e^+ and negative decision e^- (denoted by $SE_{\pi_d}^E(\cdot)$ and $SE_{\pi_i}^E(\cdot)$ for distinguishing with the path-specific effects derived for ranked data in this paper). Assume that the decision is made based on a cut-off point θ of the score. Then an interesting question is to ask, given a discrimination-free rank, whether a binary decision made based on the cut-off point θ is also discrimination free. Answering this question needs to derive a relationship between $SE_\pi(\cdot)$ and $SE_\pi^E(\cdot)$. In this subsection, we derive such relationships under the condition that $\forall \mathbf{q}, \theta \geq \mu_{c^+,\mathbf{q}} \geq \mu_{c^-,\mathbf{q}}$ and $\sigma_{c^+,\mathbf{q}} = \sigma_{c^-,\mathbf{q}} = \sigma$. We first obtain the formulas of $SE_{\pi_d}^E(\cdot)$ and $SE_{\pi_i}^E(\cdot)$ using the cut-off point θ.

LEMMA 3.4. *Given the causal graph based on score S, and a cut-off point θ for determining a binary decision E, we have*

$$SE_{\pi_d}^E(c^+, c^-) = \sum_{\mathbf{q} \in \mathfrak{X}_Q} \frac{1}{2} \left(\mathrm{erf}(\frac{\theta - \mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma}) - \mathrm{erf}(\frac{\theta - \mu_{c^+,\mathbf{q}}}{\sqrt{2}\sigma}) \right) P(\mathbf{q}|c^-),$$
(7)

$$SE_{\pi_i}^E(c^+, c^-) = \sum_{\mathbf{q} \in \mathfrak{X}_Q} \frac{1 - \mathrm{erf}(\frac{\theta - \mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma})}{2} \Delta_{\mathbf{q}}.$$
(8)

PROOF. Since θ is a cut-off point, we have $P(e^+|c^+, \mathbf{q}) = P(s \geq \theta|c^+, \mathbf{q})$ and $P(e^+|c^-, \mathbf{q}) = P(s \geq \theta|c^-, \mathbf{q})$. According to the CDF of the Gaussian distribution, we have

$$P(e^+|c^+, \mathbf{q}) = \frac{1 - \mathrm{erf}(\frac{\theta - \mu_{c^+,\mathbf{q}}}{\sqrt{2}\sigma})}{2}, \quad P(e^+|c^-, \mathbf{q}) = \frac{1 - \mathrm{erf}(\frac{\theta - \mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma})}{2}.$$

The lemma is proven by substituting $P(e^+|c^+, \mathbf{q})$ and $P(e^+|c^-, \mathbf{q})$ in the formulas of $SE_{\pi_d}^E$ and $SE_{\pi_i}^E$ in [36] with the above expressions. □

Then we present two lemmas to show the properties of $\mathrm{erf}(\cdot)$.

LEMMA 3.5. *For any $x_1 \geq x_2 \geq 0$, we have*

$$\frac{1}{2} \left(\mathrm{erf}(x_1) - \mathrm{erf}(x_2) \right) \leq \mathrm{erf}(\frac{x_1 - x_2}{2}).$$

PROOF. Since $\mathrm{erf}(x)$ $(x \geq 0)$ is concave and $\mathrm{erf}(0) = 0$, we have

$$\frac{\mathrm{erf}(x_2)}{x_2} \geq \frac{\mathrm{erf}(x_1)}{x_1} \implies \frac{x_2}{2x_1} \mathrm{erf}(x_1) \leq \frac{1}{2} \mathrm{erf}(x_2)$$

which follows that

$$\left(\frac{1}{2} - \frac{x_2}{2x_1} \right) \mathrm{erf}(x_1) \geq \frac{1}{2} \mathrm{erf}(x_1) - \frac{1}{2} \mathrm{erf}(x_2).$$

Again, since $\mathrm{erf}(x)$ $(x \geq 0)$ is concave and $\mathrm{erf}(0) = 0$, we have

$$\left(\frac{1}{2} - \frac{x_2}{2x_1} \right) \mathrm{erf}(x_1) \leq \mathrm{erf}(\frac{x_1}{2} - \frac{x_2}{2}).$$

Combining the above two inequalities, the lemma is proven. □

LEMMA 3.6. *For any $t \geq 0$, when $0 \leq x \leq t$, we have*

$$\alpha_t x \leq \mathrm{erf}(x) \leq \alpha_t x + \beta_t,$$

where

$$\alpha_t = \frac{\mathrm{erf}(t)}{t}, \quad \beta_t = \mathrm{erf}(\sqrt{\ln \frac{2t}{\sqrt{\pi}\,\mathrm{erf}(t)}}) - \frac{\mathrm{erf}(t)}{t} \sqrt{\ln \frac{2t}{\sqrt{\pi}\,\mathrm{erf}(t)}}.$$

PROOF. It is obvious that $\mathrm{erf}(x) \geq \alpha_t x$ $(0 \leq x \leq t)$. Then, β_t is obtained by calculating the tangent line with the slope α_t of $\mathrm{erf}(x)$. □

Based on the above results, the following two theorems characterize the relationship between SE_π and SE_π^E.

THEOREM 3.7. *Given the causal graph based on score S and an arbitrary cut-off point θ, if for the ranking derived from the score we have*

$$SE_{\pi_d}(c^+, c^-) \leq \frac{2\sqrt{2}(\tau - \beta_t)\sigma}{\alpha_t},$$

for the binary decision derived from the score we must have $SE_{\pi_i}^E(c^+, c^-) \leq \tau$, where

$$t = \max_{\mathbf{q}} \left\{ \frac{\mu_{c^+,\mathbf{q}} - \mu_{c^-,\mathbf{q}}}{2\sqrt{2}\sigma} \right\}.$$

PROOF. Let $x_1 = \frac{\theta - \mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma}$, $x_2 = \frac{\theta - \mu_{c^+,\mathbf{q}}}{\sqrt{2}\sigma}$, according to Lemma 3.5 we have

$$\frac{1}{2} \left(\mathrm{erf}(x_1) - \mathrm{erf}(x_2) \right) \leq \mathrm{erf}(\frac{x_1 - x_2}{2}) = \mathrm{erf}(\frac{\mu_{c^+,\mathbf{q}} - \mu_{c^-,\mathbf{q}}}{2\sqrt{2}\sigma}).$$

According to Lemma 3.6 it follows that

$$\mathrm{erf}(\frac{\mu_{c^+,\mathbf{q}} - \mu_{c^-,\mathbf{q}}}{2\sqrt{2}\sigma}) \leq \alpha_t \frac{\mu_{c^+,\mathbf{q}} - \mu_{c^-,\mathbf{q}}}{2\sqrt{2}\sigma} + \beta_t.$$

Combining the above inequality with Eq. (7), we have

$$SE_{\pi_d}^E \leq \sum_{\mathbf{q} \in \mathfrak{X}_Q} \left(\alpha_t \frac{\mu_{c^+,\mathbf{q}} - \mu_{c^-,\mathbf{q}}}{2\sqrt{2}\sigma} + \beta_t \right) P(\mathbf{q}|c^-)$$

$$= \frac{\alpha_t}{2\sqrt{2}\sigma} SE_{\pi_d} + \beta_t \leq \tau.$$

□

THEOREM 3.8. *Given the causal graph based on score S and an arbitrary cut-off point θ, if for the ranking derived from the score we have*

$$SE_{\pi_i}(c^+, c^-) \leq \frac{2\sqrt{2}(\tau - c)\sigma}{\alpha_t},$$

for the binary decision derived from the score we must have $SE_{\pi_i}^E(c^+, c^-) \leq \tau$, where

$$t = \max_{\mathbf{q}} \left\{ \frac{\max\{s\} - \mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma} \right\},$$

$$c = \frac{1}{2} - \sum_{\mathbf{q}:\Delta_{\mathbf{q}} \geq 0} \left(\frac{\alpha_t \max_{\mathbf{q}}\{\mu_{c^+,\mathbf{q}}\}}{\sqrt{2}} \right) - \sum_{\mathbf{q}:\Delta_{\mathbf{q}} < 0} \left(\frac{\alpha_t}{2\sqrt{2}} + \beta_t \right).$$

PROOF. According to Lemma 3.6 we have

$$\text{erf}(\frac{\theta - \mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma}) \geq \text{erf}(\frac{\max_{\mathbf{q}}\{\mu_{c^+,\mathbf{q}}\} - \mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma}) \geq \alpha_t \frac{\max_{\mathbf{q}}\{\mu_{c^+,\mathbf{q}}\} - \mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma},$$

$$\text{erf}(\frac{\theta - \mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma}) \leq \text{erf}(\frac{\max\{s\} - \mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma}) \leq \alpha_t \frac{\max\{s\} - \mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma} + \beta_t.$$

Combining the above inequalities with Eq. (8), we have

$$SE_{\pi_i}^E = \sum_{\mathbf{q}:\Delta_{\mathbf{q}} \geq 0} \frac{1 - \text{erf}(\frac{\theta - \mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma})}{2} \Delta_{\mathbf{q}} + \sum_{\mathbf{q}:\Delta_{\mathbf{q}} < 0} \frac{1 - \text{erf}(\frac{\theta - \mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma})}{2} \Delta_{\mathbf{q}}$$

$$\leq \frac{1}{2} + \sum_{\mathbf{q}:\Delta_{\mathbf{q}} \geq 0} \alpha_t \frac{\mu_{c^-,\mathbf{q}} - \max_{\mathbf{q}}\{\mu_{c^+,\mathbf{q}}\}}{2\sqrt{2}\sigma} \Delta_{\mathbf{q}} + \sum_{\mathbf{q}:\Delta_{\mathbf{q}} < 0} (\alpha_t \frac{\mu_{c^-,\mathbf{q}} - \max\{s\}}{2\sqrt{2}\sigma} - \beta_t) \Delta_{\mathbf{q}}$$

$$= \frac{\alpha_t}{2\sqrt{2}\sigma} SE_{\pi_i} + c \leq \tau.$$

$\square$

4 DISCOVERY AND REMOVAL ALGORITHMS

We develop the discrimination discovery and removal algorithms based on the derived π_d and π_i-specific effects. Since the values of $SE_{\pi_d}(c^+, c^-)$ and $SE_{\pi_i}(c^+, c^-)$ can be arbitrarily large, we give the criterion of direct and indirect discrimination in terms of relative difference. We require that the ratio of $SE_{\pi_d}(c^+, c^-)$ and $SE_{\pi_i}(c^+, c^-)$ over the expected score of the non-protected group, i.e., $\mathbb{E}[S|c^+]$, is smaller than a given threshold τ. For example, the Equality and Human Rights Commission (EHRC) consider 0.05 as a significant threshold for the gender pay gap. By defining the discrimination measures

$$DE_d(c^+, c^-) = \frac{SE_{\pi_d}(c^+, c^-)}{\mathbb{E}[S|c^+]}$$

and

$$DE_i(c^+, c^-) = \frac{SE_{\pi_i}(c^+, c^-)}{\mathbb{E}[S|c^+]},$$

the criterion of discrimination is shown below. To avoid reverse discrimination, we also similarly define $DE_d(c^+, c^-)$ and $DE_i(c^+, c^-)$. Then, we give the criterion of discrimination as follows.

CRITERION 1. *Given a user-defined threshold τ, direct discrimination exists if either $DE_d(c^+, c^-) > \tau$ or $DE_d(c^-, c^+) > \tau$ holds, and indirect discrimination exists if either $DE_i(c^+, c^-) > \tau$ or $DE_i(c^-, c^+) > \tau$ holds.*

Based on the above analysis, we develop the algorithm for discovering discrimination in a rank, referred to as *FDetect*, as shown in Algorithm 1. Once direct or indirect discrimination is detected, the discriminatory effects need to be eliminated before the ranked data is used for training or sharing. We propose a path-specific-effect-based Fair Ranking (*FRank*) algorithm to remove both discrimination from the ranked data and reconstruct a fair ranking. We first modify the score distributions so that the causal graph contains no discrimination, and then reconstruct a fair ranking based on the modified causal graph. As shown in Theorem 3.2, the discriminatory effect only depends on the means of the score distributions. Hence we only need to modify the means of the score.

To maximize the utility during the modification process, we minimize the distance between the original score distributions and the

---

Algorithm 1: *FDetect*

Input : Ranked dataset $\mathcal{D}$, protected attribute C, user-defined parameter τ.
Output: Direct/indirect discrimination $judge_d$, $judge_i$.

1 $judge_d = judge_i = false$;
2 Derive the score S using the Bradley-Terry model;
3 Build the causal graph for S and attributes in $\mathcal{D}$;
4 Compute $DE_d(\cdot)$ according to Theorem 3.2;
5 **if** $DE_d(c^+, c^-) > \tau \parallel DE_d(c^-, c^+) > \tau$ **then**
6 $\quad$ $judge_d = true$;
7 Divide C's children except S into $\mathbf{V}_{\pi_i}$ and $\bar{\mathbf{V}}_{\pi_i}$;
8 Compute $DE_i(\cdot)$ according to Theorem 3.2;
9 **if** $DE_i(c^+, c^-) > \tau \parallel DE_i(c^-, c^+) > \tau$ **then**
10 $\quad$ $judge_i = true$;
11 **return** $[judge_d, judge_i]$;

---

modified score distributions, as measured by the Bhattacharyya distance [4]. Specifically, for each score distribution $\mathcal{N}(\mu_{c,\mathbf{q}}, \sigma_{c,\mathbf{q}}^2)$, denote the modified distribution by $\mathcal{N}(\mu'_{c,\mathbf{q}}, \sigma_{c,\mathbf{q}}^2)$. The Bhattacharyya distance between the two distributions is given by

$$D_B = -\ln \int \sqrt{\mathcal{N}(\mu_{c,\mathbf{q}}, \sigma_{c,\mathbf{q}}^2)\mathcal{N}(\mu'_{c,\mathbf{q}}, \sigma_{c,\mathbf{q}}^2)} ds = \frac{\left(\mu_{c,\mathbf{q}} - \mu'_{c,\mathbf{q}}\right)^2}{8\sigma_{c,\mathbf{q}}^2}.$$

We define the objective function as the sum of the Bhattacharyya distances for all score distributions. As a result, we obtain the following quadratic programming problem with $\mu_{c,\mathbf{q}}$ as the variables.

$$\begin{aligned}
\text{minimize} \quad & \sum_{c \in \mathfrak{X}_C, \mathbf{q} \in \mathfrak{X}_Q} \frac{\left(\mu_{c,\mathbf{q}} - \mu'_{c,\mathbf{q}}\right)^2}{\sigma_{c,\mathbf{q}}^2} \\
\text{subject to} \quad & DE_d(c^+, c^-) \leq \tau, \quad DE_d(c^-, c^+) \leq \tau, \\
& DE_i(c^+, c^-) \leq \tau, \quad DE_i(c^-, c^+) \leq \tau.
\end{aligned}$$

After obtaining the modified score distribution by solving the quadratic programming problem, we reconstruct a fair ranking as follows. Consider the individuals with the same profile $c, \mathbf{q}$, i.e., $\forall i, c_i = c, \mathbf{q}_i = \mathbf{q}$. For each individual i, the new score s'_i is regenerated from the new CG distribution $\mathcal{N}(\mu'_{c,\mathbf{q}}, \sigma_{c,\mathbf{q}}^2)$ at the same percentile as the score s_i in the original distribution. Specifically, since $s_i = \mu_{c,\mathbf{q}} + \rho\sigma_{c,\mathbf{q}}$ and $s'_i = \mu'_{c,\mathbf{q}} + \rho\sigma_{c,\mathbf{q}}$ where ρ is the value from the standard normal distribution for the percentile, we have $s'_i = s_i + (\mu'_{c,\mathbf{q}} - \mu_{c,\mathbf{q}})$. Finally, we re-rank all individuals according to the descending order of their new scores. Since the new scores contain no discrimination, so does the new rank. The procedure is shown in Algorithm 2, referred to as *FRank*.

The computational complexity of our discovery and removal algorithms depends on how efficiently to derive the score S using Bradley-Terry model. [29] proved that the likelihood function is convex and the optimal solution can be efficiently obtained using gradient descent. The complexity also depends on the complexities of building the causal graph and computing the path-specific effect. Many researches have been devoted to improving the performance of network construction [1, 14, 28] and probabilistic inference in

---

Algorithm 2: *FRank*

Input : Ranked dataset $\mathcal{D}$, protected attribute C, user-defined
parameter τ.

Output: Modified dataset $\mathcal{D}^*$.

1 **if** *PSE-DD*$(\mathcal{D}, C, \tau) == [false, false]$ **then**
2   **return**;
3 Obtain the modified distributions of S by solving the quadratic
programming problem;
4 **foreach** c, $\mathbf{q}$ **do**
5   **foreach** $i : c_i = c$, $\mathbf{q}_i = \mathbf{q}$ **do**
6        $s_i' = s_i + (\mu_{c,\mathbf{q}}' - \mu_{c,\mathbf{q}})$;
7 Compute the new rank of each individual according to the
descending order of S, and replace the rank in $\mathcal{D}$ with the new
one to obtain $\mathcal{D}^*$;
8 **return** $\mathcal{D}^*$;

---

causal graphs [12, 13]. The complexity analysis can be found in these related literature.

5 EXPERIMENTS

5.1 Experimental Setup

In the experiments, the causal graphs are then constructed using the open source software TETRAD [11] and parameterized as described in Section 3.1. The quadratic programming is solved using CVXOPT [2]. The discrimination threshold τ is set as 0.05 for both direct and indirect discrimination.

Dataset. We use a real world dataset, the German Credit [20], which is also used in previous works [30, 32]. The German Credit dataset consists of 1000 individuals with 20 attributes applying for loans. Due to the small sample size, we only select 8 attributes in our experiments including *age, dependent, duration, housing, job, property, purpose, residence*. We treat *age* as the protected attribute, *housing* as the redlining attribute.

Based on the German Credit dataset, we generate three ranked datasets for experiments. We employ the weighted-sum ranking strategy proposed in [30, 32] to generate two ranked datasets, denoted by **D1** and **D2**. The weighted sum is computed through a weighted linear summation of certain attributes, and then all candidates are ranked according to the weighted sum. In **D1**, all attributes are summed up with equal weight, while in **D2**, the summation is for all attributes except *age*. We also use another ranked data **D** where the ranking is directly based on an original attribute *credit amount*. After that, we derive the continuous qualification scores from each ranked dataset using the Bradley-Terry model and build the causal graph. As an example, the constructed causal graph for **D** is shown in Figure 3.

Baseline. We involve the statistical parity-based discrimination discovery and removal algorithms proposed by Yang et al. [30] and Zehlike et al. [32]. For discrimination discovery, Yang et al. [30] proposed three set-based discrimination measures called **rRD**, **rND**, and **rKL** to compute the difference between the protected group and the whole dataset in terms of *risk difference*, *risk ratio*, and *Kullback-Leibler distance*. They compute the values of difference at several discrete points (e.g., top-10, top-20, $\cdots$) and sum up all values with the logarithmic discounts. All measures are normalized to 0-1 range
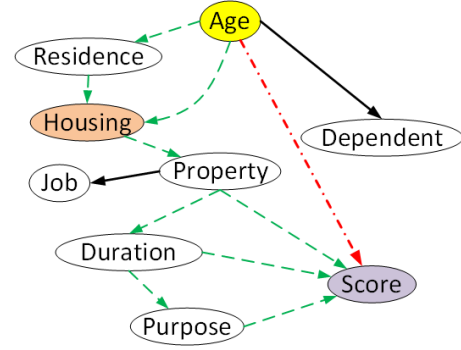
Figure 3: The causal graph of D. The yellow node Age is the protected attribute, the orange node Housing is the redlining attribute, and the purple node Score is the decision attribute. The red dash-dot line captures the direct discrimination from Age to Score, and the green dashed line captures the indirect discrimination through Housing.

(0 is the most fair value and 1 is the least fair value). Since they don't provide any criterion for discrimination discovery, we simply use 0.05 as the threshold for all three measures. Zehlike et al. [32] proposed an adjusted fairness condition (*FairCon*) that requires the minimum number of protected candidates in every prefix of the ranking list. For discrimination removal, Yang et al. [30] proposed a fair data generator (*FairGen*) that manipulates the permutation according to the user-defined preference f. For example, if $f = 0.05$, all the candidates are well mixed in equal proportion at every prefix; if $f = 1$, the candidates from the protected group are ranked at the bottom. Zehlike et al. [32] proposed discrimination removal methods, *FA*IR*, to select the most qualified candidate from the corresponding group at every prefix in order to satisfy the adjusted fairness condition.

To evaluate the data utility of all removal approaches, we adopt two widely used metrics, the Spearman's footrule distance (*SFD*) and the Kendall's tau distance (*KTD*) [19]. The Spearman's footrule distance (*SFD*) measures the total element-wise displacement between the modified permutation and the original one. The Kendall's tau distance (*KTD*) measures the total number of pairwise inversions between the two permutations. For both of the distance metrics, the larger values indicate more data utility loss.

5.2 Discrimination Discovery

We quantify the strength of direct and indirect discrimination using our method *FDetect* for all three ranked datasets. The results are shown in Table 1. For dataset **D1**, all attributes including the protected attribute are used for ranking directly. Thus, the ground-truth is that both direct and indirect discrimination occurs in this dataset. Our method obtains $DE_d(c^+, c^-) = 0.231$ and $DE_i(c^+, c^-) = 0.055$, showing that both direct and indirect discrimination are correctly identified. For dataset **D2**, since we use all the other attributes except the protected attribute in the ranking process, the ground-truth is that the indirect discrimination occurs but the direct discrimination doesn't. Our method shows $DE_d(c^+, c^-) = 0.026$ and $DE_i(c^+, c^-) = 0.061$, which is also consistent with the ground-truth.

For dataset **D**, we don't have the ground-truth. Our method obtains that $DE_d(c^+, c^-) = 0.005$ and $DE_i(c^+, c^-) = 0.013$. The results imply that neither direct discrimination nor indirect discrimination exists in this dataset.

The statistical parity-based methods **rRD**, **rND**, **rKL** and *FairCon* cannot distinguish direct and indirect discrimination. We directly report the results produced by these methods as shown in Table 1. For **D1**, the method proposed by Yang et al. shows that **rRD** = 0.590, **rND** = 0.440, and **rKL** = 0.204, while Zehlike's *FairCon* shows that the third position doesn't satisfy the minimum fair requirement. For **D2**, Yang's method shows that **rRD** = 0.160, **rND** = 0.102, and **rKL** = 0.022, while *FairCon* reports that the 5-th position doesn't satisfy the fair requirement. Most methods conclude discrimination for both dataset, which kind of match our conclusions. However, for **D**, Yang's method shows that **rRD** = 0.109, **rND** = 0.070, and **rKL** = 0.008, where three values make the contradictory conclusions: there is no discrimination according to **rRD** but **rND** and **rKL** report significant discrimination. *FairCon* shows that the ranking cannot satisfy the fair requirement at the 20-th position. All methods cannot obtain the results that are consistent with ours, implying that they may produce incorrect or misleading conclusions.

Table 1: Comparison of discrimination discovery methods on the direct discrimination. The second column represents the ground-truth for direct and indirect discrimination.

	Ground-Truth	DE_d	DE_i	*FairCon*	**rRD**	**rND**	**rKL**
D1	Y/Y	0.231	0.055	3rd	0.590	0.440	0.204
D2	N/Y	0.026	0.061	5th	0.160	0.102	0.022
D	-	0.005	0.013	20th	0.109	0.070	0.008

5.3 Discrimination Removal

We perform *FRank* to remove discrimination and reconstruct fairly ranked datasets with neither direct nor indirect discrimination. Our theoretical results guarantee that there is no discrimination after modification. For comparison, we also execute *FairGen* [30] and *FA*IR* [32]. After removing discrimination, we further apply *FDetect* to evaluate whether the newly-generated data achieves truly discrimination-free. The results of three removal methods are shown in Table 2. As can be seen, our method *FRank* removes both direct and indirect discrimination precisely. However, *FairGen* and *FA*IR* cannot achieve discrimination-free. *FairGen* removes neither direct nor indirect discrimination. It even introduces more discrimination to **D**. *FA*IR* can mitigate part of direct discrimination, but fails to remove indirect discrimination.

We adopt the Spearman's footrule distance (*SFD*) and the Kendall's tau distance (*KTD*) to evaluate the data utility loss when mitigating the discrimination. As can be seen from the last two columns of Table 2, our method *FRank* incurs relatively small data utility loss, but *FairGen* suffers large data utility loss while not achieving discrimination-free. Although *FA*IR* introduces quite a small data utility loss, it fails to mitigate indirect discrimination. It is worth pointing out that there is no direct or indirect discrimination in **D** so our *FRank* doesn't result in any distortion. On the contrary, *FairGen* leads to too much utility loss.

Table 2: Discrimination and data utility measured on the new ranked data produced different methods. Values violating the discrimination criterion are marked in bold.

Data	Methods	DE_d	DE_i	*KTD*	*SFD*
D1	*FRank*	0.050	0.050	24602	72938
	FairGen	**0.234**	**0.064**	11150	44600
	FA*IR	**0.077**	**0.066**	13882	55528
D2	*FRank*	0.029	0.050	5851	18090
	FairGen	**0.246**	**0.060**	19483	77934
	FA*IR	0.022	**0.061**	231	924
D	*FRank*	0.005	0.013	0	0
	FairGen	**0.250**	0.012	20806	83226
	FA*IR	0.003	0.013	143	572

We also examine how the data utility varies with different values of the discrimination threshold τ. We perform *FRank* on **D1** and vary the threshold τ for *FRank* from 0.00 to 0.25 for evaluating how much data utility loss is incurred. In Table 3, we can see that both the Spearman's footrule distance (*SFD*) and the Kendall's tau distance (*KTD*) decrease with the increase of τ, which means that less utility loss is incurred with a larger threshold. This observation is consistent with our analysis since the larger τ, the more relaxed the constraints in *FRank*.

Table 3: Comparison of *FRank* with varied τ.

τ	0.00	0.05	0.10	0.15	0.20	0.25
$DE_d(c^+, c^-)$	0.000	0.050	0.100	0.150	0.200	0.231
$DE_i(c^+, c^-)$	0.000	0.050	0.055	0.055	0.055	0.055
SFD	43490	24602	14370	9041	3444	0
KTD	123626	72938	45636	28652	11054	0

6 RELATED WORK

Fairness-aware learning is an active research area in machine learning and data mining. Many methods have been proposed for constructing discrimination-free machine learning models, which either based on data preprocessing or model tweaking. Data preprocessing methods [8, 10, 15, 24, 36] modify the historical training data to remove discrimination before it is used for learning a model. Model tweaking methods [7, 16, 17, 31] require some tweak or adjustment of the constructed machine learning models. Fair ranking is an emerging topic in fairness-aware learning. Current works in fair ranking are mainly based on the statistical parity. In [32], it is required that a preset proportion of protected individuals that must be maintained in each prefix of the ranking for the rank to be fair. However, many works (e.g., [36]) have shown that statistical parity alone is insufficient as a general notion of fairness.

Recently, several studies have been devoted to analyzing discrimination from the causal perspective [5, 24, 33–38]. In [5], the authors proposed a framework based on the Suppes-Bayes causal network and developed several random-walk-based methods to detect different types of discrimination. However, it is unclear how the number of random walks is related to practical discrimination metrics. In addition, the construction of the Suppes-Bayes causal

network is impractical with the large number of attribute-value pairs. In this work we adopt the causal graph used in [24, 33–38]. A causal graph is a probabilistic graph model widely used for causation representation, reasoning and inference [25]. The limitation of these works is that they focus on the classification problems and cannot be applied directly to the fair ranking problem. This is because in their models, the decision of each individual is treated as an independent random variable, but the ranking positions of different individuals are correlated. In this paper we address above limitations and develop the causal-based fair ranking algorithms.

In data science, it is well-studied in data mining how to model a ranking using a continuous score space [23]. Several models, such as the Plackett-Luce model [21, 26], the Mallows model [22] and the Bradley-Terry model [6], are widely used in this field. In this work, we adopt the Bradley-Terry model to characterize the ranked data and obtain the continuous scores from the ranks.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we studied the problem of discovering discrimination in a rank and reconstructing a fair rank if discrimination is detected. We made use of the causal graph to capture the biases in the rank as the causal effect. To address the limitation of the existing single data-type causal graph, we modeled the ranking positions using a continuous score, and built the causal graph for the profile attributes as well as the score. Then, we extended the path-specific effect technique to the mixed-variable causal graph, which is used to quantitatively measure direct and indirect discrimination in the ranked data. We also theoretically analyzed the relationship between the path-specific effects for the ranked data and those for the binary decision. Based on that, we developed an algorithm for discovering both direct and indirect discrimination, as well as an algorithm to reconstruct a fair rank from the causal graph. The experiments using the German Credit dataset showed that our methods correctly measure the discrimination in the rank and reconstruct a rank that does not contain either direct or indirect discrimination, while the statistical parity-based method may obtain incorrect and misleading results.

In Theorem 3.2 we assume that the π_i-specific effect is identifiable from the data. In some cases, the π_i-specific effect is not able to be computed from the data due to the inherent unidentifiability of the path-specific effect [3]. In our another work [34], we have discussed how to deal with this situation, referred to as the unidentifiable situation, and developed a lower and upper bound to the unidentifiable path-specific effect. Similar ideas can be adopted to deal with unidentifiable situation for ranked data. We leave this part for future work.

Repeatability: Our software together with the datasets used in this paper are available at http://tiny.cc/fair-ranking.

ACKNOWLEDGMENTS

REFERENCES

[1] Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. 2010. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. *JMLR* 11 (2010), 171–234.

[2] M S Andersen, J Dahl, and L Vandenberghe. 2013. CVXOPT: A Python package for convex optimization. (2013). www.cvxopt.org

[3] Chen Avin, Ilya Shpitser, and Judea Pearl. 2005. Identifiability of path-specific effects. *IJCAI'05* (2005), 357–363.

[4] A. Bhattacharyya. 1946. On a measure of divergence between Two multinomial populations. *Sankhyā: The Indian Journal of Statistics* 7, 4 (1946), 401–406.

[5] Francesco Bonchi, Sara Hajian, Bud Mishra, and Daniele Ramazzotti. 2017. Exposing the probabilistic causal structure of discrimination. *JDSA* 3, 1 (2017).

[6] Ralph Allan Bradley. 1954. Rank analysis of incomplete block designs. *Biometrika* 41, 3-4 (1954), 502–537.

[7] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292.

[8] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. In *NIPS'17*. 3995–4004.

[9] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *ITCS '12*. 214–226.

[10] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *KDD*. ACM, 259–268.

[11] Clark Glymour et al. 2004. The TETRAD project. http://www.phil.cmu.edu/tetrad. (2004).

[12] David Heckerman and John S Breese. 1994. A new look at causal independence. In *UAI'94*. 286–292.

[13] David Heckerman and John S Breese. 1996. Causal independence for probability assessment and inference using Bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 26, 6 (1996), 826–831.

[14] Markus Kalisch and Peter Bühlmann. 2007. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *JMLR* 8 (2007), 613–636.

[15] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *IC4'09*. IEEE, 1–6.

[16] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *ECML PKDD*.

[17] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In *ICDMW*. IEEE, 643–650.

[18] Daphne Koller and Nir Friedman. 2009. *Probabilistic graphical models: Principles and techniques*. The MIT Press.

[19] Ravi Kumar and Sergei Vassilvitskii. 2010. Generalized distances between rankings. In *WWW '10*. 571.

[20] M Lichman. 2013. UCI Machine Learning Repository. (2013).

[21] Robert Duncan Luce. 1959. *Individual choice behavior: A theoretical analysis*. Wiley. 153 pages.

[22] C L Mallows. 1957. Non-null ranking models. I. *Biometrika* 44, 1/2 (1957).

[23] John I Marden. 1996. *Analyzing and modeling rank data*. CRC Press.

[24] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *AAAI'18*.

[25] Judea Pearl. 2009. *Causality: Models, reasoning and inference*. Cambridge University Press.

[26] R L Plackett. 1975. The analysis of permutations. *Appl. Stat.* 24, 2 (1975), 193–202.

[27] Ilya Shpitser. 2013. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cogn. Sci.* 37, 6 (2013).

[28] Ioannis Tsamardinos, Constantin F Aliferis, Alexander Statnikov, and Laura E Brown. 2003. Scaling-up Bayesian network learning to thousands of variables using local learning techniques. *Vanderbilt University DSL TR-03-02* (2003).

[29] Fangzhao Wu, Jun Xu, Hang Li, and Xin Jiang. 2014. Ranking optimization with constraints. *CIKM '14* (2014), 1049–1058.

[30] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *SSDBM '17*, Vol. format. 1–6.

[31] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*. 962–970.

[32] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A fair top-k ranking algorithm. In *CIKM'17*.

[33] Lu Zhang and Xintao Wu. 2017. Anti-discrimination learning: a causal modeling-based framework. *JDSA* (may 2017).

[34] Lu Zhang and Xintao Wu. 2018. Causal modeling-based discrimination discovery and removal: criteria, bounds, and algorithms. *Submitted to TKDE* (2018).

[35] Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. Situation testing-based discrimination discovery: A causal inference approach. In *IJCAI'16*, Vol. 2016-Janua. 2718–2724.

[36] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. A causal framework for discovering and removing direct and indirect discrimination. In *IJCAI'17*. 3929–3935.

[37] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. Achieving non-discrimination in data release. In *KDD '17*. 1335–1344.

[38] Lu Zhang, Yongkai Wu, and Xintao Wu. 2018. Achieving non-discrimination in prediction. In *IJCAI'18*.