Achieving Non-Discrimination in Data Release

Lu Zhang
University of Arkansas
lz006@uark.edu

Yongkai Wu
University of Arkansas
yw009@uark.edu

Xintao Wu
University of Arkansas
xintaowu@uark.edu

ABSTRACT

Discrimination discovery and prevention/removal are increasingly important tasks in data mining. Discrimination discovery aims to unveil discriminatory practices on the protected attribute (e.g., gender) by analyzing the dataset of historical decision records, and discrimination prevention aims to remove discrimination by modifying the biased data before conducting predictive analysis. In this paper, we show that the key to discrimination discovery and prevention is to find the meaningful partitions that can be used to provide quantitative evidences for the judgment of discrimination. With the support of the causal graph, we present a graphical condition for identifying a meaningful partition. Based on that, we develop a simple criterion for the claim of non-discrimination, and propose discrimination removal algorithms which accurately remove discrimination while retaining good data utility. Experiments using real datasets show the effectiveness of our approaches.

CCS CONCEPTS

• **Information systems** → **Data mining**; • **Applied computing** → **Law, social and behavioral sciences**; • **Mathematics of computing** → *Causal networks*;

KEYWORDS

discrimination discovery and removal, causal graph

1 INTRODUCTION

Discrimination discovery and prevention/removal has been an active research area recently [9, 12, 16, 28, 29, 39]. Discrimination refers to unjustified distinctions of individuals based on their membership in a certain group. Federal Laws and regulations (e.g., Fair Credit Reporting Act or Equal Credit Opportunity Act) prohibit discrimination on several grounds, such as gender, age, marital status, sexual orientation, race, religion or belief, and disability or illness, which are referred to as the *protected attributes*. Different types of discrimination have been introduced, which can be generally categorized as direct and indirect discrimination [12, 28]. Direct discrimination occurs when individuals are treated less favorably in comparable situations explicitly due to their membership in a protected group; indirect discrimination refers to an apparently neutral practice which results in an unfair treatment of a protected group. In this paper, we focus on the problem of discrimination discovery and prevention on direct discrimination. In the following, we simply say discrimination for direct discrimination.

Table 1: Summary statistics of Example 1.

Test score	L				H			
Gender	Female		Male		Female		Male	
Major	CS	EE	CS	EE	CS	EE	CS	EE
No. applicants	450	150	150	450	300	100	100	300
Admission rate	20%	40%	20%	40%	50%	70%	50%	70%
	25%		35%		55%		65%	

Table 2: Summary statistics of Example 2.

Major	CS				EE			
Gender	Female		Male		Female		Male	
Test score	L	H	L	H	L	H	L	H
No. applicants	450	300	150	100	600	300	200	100
Admission rate	30%	50%	36%	40%	40%	60%	45%	50%
	38%		38%		47%		47%	

For a quantitative measurement of discrimination, a general legal principle is to measure the difference in the proportion of positive decisions between the protected group and non-protected group [28]. Ensuring non-discrimination in a released data is not trivial. Simply considering the difference measured at the whole dataset level fails to take into account the part of differences that are explainable by other attributes, and removing all the differences will result in reverse discrimination [32]. Thus, directly removing the discrimination at the whole dataset level or conditioning on an arbitrary attribute and then removing the difference within each produced subpopulation cannot achieve true discrimination-free. It is imperative to determine whether a partition is meaningful for measuring and removing discrimination and then to remove discrimination from all meaningful partitions.

Typically, given a dataset from an organization, a partition is determined by a subset of attributes and a subpopulation is specified by a value assignment to the attributes. We demonstrate that only meaningful partitions can be used to provide quantitative evidences for the judgment of discrimination. Using inappropriate partitions will result in misleading conclusions. Consider a toy model for a university admission system that contains four attributes: `gender`, `major`, `test_score`, and `admission`, where `gender` is the protected attribute, and `admission` is the decision. We assume there is no correlation between `gender` and `test_score`. The summary statistics of the admission rate is shown in Table 1. It can be observed that the average admission rate is 37% for females and 46% for males. It is already known that the judgment of discrimination cannot be made simply based on the average admission rates in the whole population and further partitioning is needed. If we partition the

KDD'17, , August 13–17, 2017, Halifax, NS, Canada.

ACM ISBN 978-1-4503-4887-4/17/08…$15.00
https://doi.org/10.1145/3097983.3098167

data conditioning on test_score as shown in the table, there exist significant differences (from either 35%−25% for L or from 65%−55% for H) between the admission rates of females and males in the two subpopulations. However, intuitively test_score should not be used for partitioning the data alone as it is uncorrelated with the protected attribute. In fact, this result is indeed misleading, since if we carefully examine the admission rates for each major or each combination {major, test_score}, it shows no bias in any of the subpopulations. Therefore, it would be groundless if a plaintiff tries to file a lawsuit of discrimination against the university by demonstrating admission rate difference either in the whole population or based on the partitioning on test_score.

Moreover, as there are often multiple meaningful partitions, examining one partition showing no bias does not guarantee no bias based on other partitions. Consider a different example on the same toy model shown in Table 2. The average admission rate now becomes 43% equally for both females and males. Further conditioning on major, which appears to be a reasonable partition, still shows that females and males have the same chance to be admitted in the two subpopulations. However, when partitioning the data based on the combination {major, test_score}, significant differences ($\geq$ 5%) between the admission rates of females and males present. The difference among applicants applied to either a major with test scores of L is clear evidence of discrimination against females. The difference among applicants applied to either a major with test scores of H can be treated as reverse discrimination against males, or tokenism where some strong male applicants are purposefully rejected to refute a claim of discrimination against females. In this case, the data publisher cannot make a non-discrimination claim.

The above examples show that, any quantitative evidence of discrimination must be measured under a meaningful partition. In addition, to ensure non-discrimination, we must show no bias for all meaningful partitions. In this paper, we make use of the causal graphs to identify meaningful partitions. A causal graph is a probabilistic graph model widely used for causation representation and inference [30]. As stated in [10], discrimination claims usually require plaintiffs to demonstrate a causal connection between the decision and the protected attribute. We then develop discrimination discovery and removal algorithms. Our results include: 1) a graphical condition for identifying meaningful partitions, which are defined by subsets of attributes called the *block sets*; 2) an efficient discrimination detection algorithm that only needs to examine one single block set but ensures non-discrimination across all meaningful partitions; and 3) discrimination removal algorithms which achieve non-discrimination while maximizing the data utility. Our approaches can be used to find quantitative evidences of discrimination for plaintiffs, or to achieve a non-discrimination guarantee for data owners. The experiments using real datasets show that our proposed approaches are effective in discovering and removing discrimination whereas previous works on discrimination removal cannot achieve non-discrimination in all meaningful partitions.

2 MODELING DISCRIMINATION USING CAUSAL GRAPH

Consider a dataset $\mathcal{D}$ which may contain discrimination against a certain protected group. Each individual in $\mathcal{D}$ is specified by a set of attributes $\mathbf{V}$, which includes the protected attribute (e.g., gender), the decision attribute (e.g., admission), and the non-protected attributes (e.g., major). Throughout this paper, we use an uppercase alphabet, e.g., X, to represent an attribute; a bold uppercase alphabet, e.g., $\mathbf{X}$, to represent a subset of attributes, e.g., {gender, major}. We use a lowercase alphabet, e.g., x, to represent a domain value of attribute X; a bold lowercase alphabet, e.g., $\mathbf{x}$, to represent a value assignment to $\mathbf{X}$. We denote the decision attribute by E, associated with domain values of positive decision e^+ and negative decision e^-; denote the protected attribute by C, associated with two domain values c^- (e.g., female) and c^+ (e.g., male).

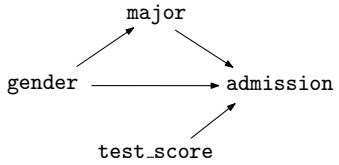


Figure 1: Causal graph of an example university admission system.

For the quantitative measurement of discrimination, we use *risk difference* [28] to measure the difference in the the proportion of positive decisions between the protected group and non-protected group. Formally, by assuming c^- is the protected group, risk difference is defined as $\Delta P|_{\mathbf{s}} = \Pr(e^+|c^+, \mathbf{s}) - \Pr(e^+|c^-, \mathbf{s})$, where $\mathbf{s}$ denotes a specified subpopulation produced by a partition $\mathbf{S}$. We say that the protected group is treated less favorably within subpopulation $\mathbf{s}$ if $\Delta P|_{\mathbf{s}} \geq \tau$, where $\tau > 0$ is a threshold for discrimination depending on law. For instance, the 1975 British legislation for sex discrimination sets $\tau = 0.05$, namely a 5% difference. To avoid reverse discrimination, we do not specify which group is the protected group. Thus, we use $|\Delta P|_{\mathbf{s}}|$ to deal with both scenarios where either c^- or c^+ is designated as the protected group.

A DAG $\mathcal{G}$ is represented by a set of nodes and a set of arcs. Each node represents an attribute in $\mathcal{D}$. Each arc, denoted by an arrow $\rightarrow$ in the graph, connects a pair of nodes where the node emanating the arrow is called the parent of the other node. The DAG is assumed to satisfy the *local Markov condition*, i.e., each node X is independent of all its non-descendants conditioning on its parents $\mathrm{Par}(X)$. Each node is associated with a conditional probability table (CPT) specified by $\Pr(X|\mathrm{Par}(X))$. The joint probability distribution can be computed using the factorization formula [19]:

$$\Pr(\mathbf{v}) = \prod_{X \in \mathbf{V}} \Pr(x|\mathrm{Par}(X)). \quad (1)$$

Spirtes et al. [30] have shown that, when causal interpretations are given to the DAG, i.e., each node's parents are this node's direct causes, the DAG represents a correct causal structure of the underlying data. In particular, the causation among the attributes are encoded in the missing arcs in the DAG: if there is no arc between two nodes in $\mathcal{G}$, then it represents no direct causal effect between the two attributes in $\mathcal{D}$. The DAG with the causal interpretation is called the *causal graph*. For example, the causal graph of the illustrative examples in the introduction is shown in Figure 1.

In this paper, we assume that we have a causal graph $\mathcal{G}$ that correctly captures the causal structure of the dataset. We also assume that the arc $C \rightarrow E$ is present in $\mathcal{G}$, since the absence of the arc represents a zero direct effect of C on E. A causal DAG can be learned from data and domain knowledge. In the past decades, many algorithms have been proposed to learn causal DAGs from data [6, 15, 24, 30]. In our implementation, we use the original PC algorithm [30] to build the causal graph.

2.1 Identifying Meaningful Partition

Discrimination occurs due to different decisions made explicitly based on the membership in the protected group. As stated, all discrimination claims require plaintiffs to demonstrate the existence of a causal connection between the decision and the protected attribute. Given a partition $\mathbf{S}$, for $\Delta P|_{\mathbf{s}}$ to capture the discriminatory effect and be considered as a quantitative evidence of discrimination, one needs to prove that this difference is indeed caused by the protected attribute. In other words, the partition $\mathbf{S}$ must guarantee that the influence from the protected attribute C to the decision E is only transmitted along $C \rightarrow E$ and all the influences from C to E through all other paths are shielded/blocked given the partition value $\mathbf{s}$. We first introduce the well-known concept of d-separation in causal modeling and then present our result of identifying meaningful partitions.

Definition 2.1 (d-Separation [30]). Consider a dataset $\mathcal{D}$ and its represented causal graph $\mathcal{G}$. $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are disjoint sets of attributes. $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{Z}$, denoted by $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}}$, if and only if $\mathbf{Z}$ blocks all paths from every node in $\mathbf{X}$ to every node in $\mathbf{Y}$. A path p is said to be blocked by $\mathbf{Z}$ if and only if

(1) p contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ such that the middle node m is in $\mathbf{Z}$, or
(2) p contains an collider $i \rightarrow m \leftarrow j$ such that the middle node m is not in $\mathbf{Z}$ and no descendant of m is in $\mathbf{Z}$.

$\mathbf{X}$ and $\mathbf{Y}$ are said to be conditionally independent given $\mathbf{Z}$, denoted by $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{D}}$, if $\Pr(\mathbf{x}|\mathbf{y},\mathbf{z}) = \Pr(\mathbf{x}|\mathbf{z})$ holds for all values $\mathbf{x}, \mathbf{y}, \mathbf{z}$. With the local Markov condition, the d-separation criterion in $\mathcal{G}$ and the conditional independence relations in $\mathcal{D}$ are connected such that, if we have $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}}$, then we must have $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{D}}$. This connection is referred to as the *global Markov condition.*

Being facilitated with the d-separation criterion, the following theorem shows under what conditions a node set $\mathbf{B}$ forms a meaningful partition for indeed measuring discrimination of C on E.

Theorem 2.2. *Given a causal graph $\mathcal{G}$, a node set $\mathbf{B}$ forms a meaningful partition for measuring discrimination with $\Delta P|_{\mathbf{b}}$ if and only if $\mathbf{B}$ satisfies: (1) $(C \perp\!\!\!\perp E \mid \mathbf{B})_{\mathcal{G}'}$ holds, (2) $\mathbf{B}$ contains none of E's decedents, where $\mathcal{G}'$ is constructed by deleting arc $C \rightarrow E$ from $\mathcal{G}$.*

Proof. Construct a new DAG $\mathcal{G}'$ by deleting the arc $C \rightarrow E$ from $\mathcal{G}$ and keeping everything else unchanged. Consider a node set $\mathbf{B}$ such that $(C \perp\!\!\!\perp E \mid \mathbf{B})_{\mathcal{G}'}$, and use $\mathbf{B}$ to examine the conditional independence relations in $\mathcal{D}$. The possible difference between the causal relationships represented by $\mathcal{G}'$ and $\mathcal{G}$ lies merely in the presence of the direct causal effect of C on E. Therefore, if there is no direct causal effect of C on E in $\mathcal{G}$, $\mathcal{G}$ can be considered as equivalent to $\mathcal{G}'$, which means that we should also obtain $(C \perp\!\!\!\perp E \mid \mathbf{B})_{\mathcal{G}}$. According to the global Markov condition, this entails $(C \perp\!\!\!\perp E \mid \mathbf{B})_{\mathcal{D}}$, i.e., $\Pr(e^+|c^+,\mathbf{b}) = \Pr(e^+|c^-,\mathbf{b}) = \Pr(e^+|\mathbf{b})$ for each value assignment $\mathbf{b}$ of $\mathbf{B}$. Thus, if we observe $\Pr(e^+|c^+,\mathbf{b}) \neq \Pr(e^+|c^-,\mathbf{b})$, the difference must be due to a non-zero direct causal effect of C on E. This implies that $\Delta P|_{\mathbf{b}} = \Pr(e^+|c^+,\mathbf{b}) - \Pr(e^+|c^-,\mathbf{b})$ can be used to measure the direct causal effect of C on E.

On the other hand, if a node set $\mathbf{B}$ does not satisfy $(C \perp\!\!\!\perp E \mid \mathbf{B})_{\mathcal{G}'}$, then this conditional dependence between C and E given $\mathbf{B}$ that is not caused by the direct causal effect will also exist in $\Delta P|_{\mathbf{b}}$. As a result, $\Delta P|_{\mathbf{b}}$ cannot correctly measure the direct causal effect.

In addition, for $\Delta P|_{\mathbf{b}}$ to correctly measure the direct causal effect, set $\mathbf{B}$ cannot contain any descendant of E even when it satisfies the requirement $(C \perp\!\!\!\perp E \mid \mathbf{B})_{\mathcal{G}'}$. This is because when conditioning on E's descendants, part of the knowledge of E is already given since the consequences caused by E is known.

Hence, the theorem is proven. □

We call the node set that satisfies the requirement in the above theorem as the *block set*. Theorem 2.2 shows that each block set $\mathbf{B}$ forms a meaningful partition on the dataset where the direct causal effect of C on E within each subpopulation $\mathbf{b}$ can be correctly measured by $\Delta P|_{\mathbf{b}}$. On the other hand, for any partition that is not defined by a block set, the measured differences, which either contain spurious influences or have been explained by the consequences of the decisions, cannot correctly measure the direct causal effect and hence cannot be used to prove discrimination or non-discrimination. In practice, if $|\Delta P|_{\mathbf{b}}| \geq \tau$, it is a quantitative evidence of discrimination against either c^- or c^+ for subpopulation $\mathbf{b}$. Thus, we have the following corollary.

Corollary 2.3. *Discriminatory effect is considered to present for subpopulation $\mathbf{b}$ if $\mathbf{B}$ is a block set and $|\Delta P|_{\mathbf{b}}| \geq \tau$, where $\Delta P|_{\mathbf{b}} = \Pr(e^+|c^+,\mathbf{b}) - \Pr(e^+|c^-,\mathbf{b})$.*

3 DISCRIMINATION DISCOVERY AND PREVENTION

3.1 Non-Discrimination Criterion

Now we develop the criterion that ensures non-discrimination for a dataset. Based on Theorem 2.2, if each block set $\mathbf{B}$ is examined and ensures that $|\Delta P|_{\mathbf{b}}| < \tau$ holds for each subpopulation $\mathbf{b}$, we can say that the dataset is not liable for any claim of direct discrimination. Otherwise, there exists a subpopulation $\mathbf{b}$ of a block set such that $|\Delta P|_{\mathbf{b}}| \geq \tau$, which implies that subpopulation $\mathbf{b}$ suffers the risk of being accused of discrimination. Therefore, we give the following theorem.

Theorem 3.1. *Non-discrimination is claimed for $\mathcal{D}$ if and only if inequality $|\Delta P|_{\mathbf{b}}| < \tau$ holds for each value assignment $\mathbf{b}$ of each block set $\mathbf{B}$.*

We use the example shown in Figure 1 to illustrate how the criterion works. There are two block sets in this graph: {`major`}, and {`major`,`test_score`}. Note that `test_score` alone is not a block set. That is why conditioning on it will produce misleading results. For the example shown in Table 1, examining both block sets shows no discriminatory effect. Thus, non-discrimination can be claimed. For the example shown in Table 2, although examining {`major`} shows

no discriminatory effect, when examining {major,test_score} we observe $|\Delta P|_{\{\text{math,B}\}}| = 0.06$, $|\Delta P|_{\{\text{math,A}\}}| = 0.10$, $|\Delta P|_{\{\text{biology,B}\}}| = 0.05$, and $|\Delta P|_{\{\text{biology,A}\}}| = 0.10$. Thus, the evidences of discrimination for four subpopulations are identified.

Although Theorem 3.1 provides a clear criterion for the claim of non-discrimination, it requires examining each subpopulation of each block set. A brute force algorithm may have an exponential complexity. Instead of examining all block sets, the following theorem shows that we only need to examine one node set $\mathbf{Q}$, which is the set of all E's parents except C, i.e., $\mathbf{Q} = \text{Par}(E)\backslash\{C\}$.

Theorem 3.2. *Non-discrimination is claimed if and only if inequality $|\Delta P|_{\mathbf{q}}| < \tau$ holds for each value assignment $\mathbf{q}$ of set $\mathbf{Q}$ where $\mathbf{Q} = \text{Par}(E)\backslash\{C\}$.*

Proof. We first give two lemmas and their proofs are given in Appendix.

Lemma 3.3. *Given a value assignment $\mathbf{b}$ of a block set $\mathbf{B}$, $\mathbf{Q} = \text{Par}(E)\backslash\{C\}$, we have*

$$\Delta P|_{\mathbf{b}} = \sum_{\mathbf{Q}'} \Pr(\mathbf{q}'|\mathbf{b}) \cdot \Delta P|_{\mathbf{q}},$$

where $\mathbf{q}'$ goes through all the possible combinations of the values of non-overlapping attributes $\mathbf{Q}' = \mathbf{Q}\backslash\mathbf{B}$. For overlapping attributes $\mathbf{Q} \cap \mathbf{B}$, $\mathbf{q}$ and $\mathbf{b}$ have the same values.

Lemma 3.4. *Node set $\mathbf{Q}$ of all E's parents except C, i.e., $\mathbf{Q} = \text{Par}(E)\backslash\{C\}$, must be a block set.*

Lemma 3.3 indicates that, for each value assignment $\mathbf{b}$ of each block set, $\Delta P|_{\mathbf{b}}$ can be expressed by a weighted average of $\Delta P|_{\mathbf{q}}$. If $|\Delta P|_{\mathbf{q}}| < \tau$ for each subpopulation of $\mathbf{Q}$, then it is guaranteed that $|\Delta P|_{\mathbf{b}}| < \tau$ holds for each subpopulation of each block set. According to Theorem 3.1, non-discrimination is claimed. Otherwise, Lemma 3.4 means that $|\Delta P|_{\mathbf{b}}| \geq \tau$ for at least one block set $\mathbf{Q}$, which provides the evidence of discrimination. □

3.2 Discrimination Discovery

We present the non-discrimination certifying algorithm. We emphasize that our algorithm only needs to examine one set $\mathbf{Q}$ (instead of all block sets) based on Theorem 3.2. The procedure of the algorithm is shown in Algorithm 1. It first finds set $\mathbf{Q}$ in the graph. Then, the algorithm computes $|\Delta P|_{\mathbf{q}}| = |\Pr(e^+|c^+, \mathbf{q}) - \Pr(e^+|c^-, \mathbf{q})|$ for each subpopulation $\mathbf{q}$, and makes the judgment of non-discrimination based on the criterion.

The complexity from Line 2 to 8 is $O(|\mathbf{Q}|)$, where $|\mathbf{Q}|$ is the number of value assignments of $\mathbf{Q}$. The function *findParent*(E) (Line 1) finds the parents of E in a causal graph. A straightforward way is to first build a causal graph from the dataset using a structure learning algorithm (e.g., the classic PC algorithm), then find the parents of E in the graph. The complexity of the PC algorithm is bounded by the largest degree in the undirected graph. In the worst case, the number of conditional independence tests required by the algorithm is bounded by $\frac{n^2(n-1)^{k-1}}{(k-1)!}$ where k is the maximal degree of any vertex and n is the number of vertices. However, in our algorithm we only need to identify the parents of E without the need of building the complete graph. Thus, we can use local causal discovery algorithms such as the Markov blanket [31] to determine the local structure for the decision attribute E. We leave this part as our future work.

3.3 Discrimination Removal

When non-discrimination is not claimed, the discriminatory effects need to be removed by modifying the data before it is used for predictive analysis (e.g., building a discrimination-free classifier). Since the modification makes the data distorted, it may cause losses in data utility when compared with the original data. Thus, a general requirement in discrimination removal is to maximize the utility of the modified data while achieving non-discrimination. A naive approach such as used in [9] would be totally removing the protected attribute from the dataset to eliminates discrimination. However, as we shall show in the experiments, in this way the data utility would be greatly suffered. In this section, we propose two strategies that exactly remove discrimination while retaining good data utility.

3.3.1 Discrimination Removal by Modifying Causal Graph. The first strategy modifies the constructed causal graph and uses it to generate a new dataset. Specifically, it modifies the CPT of E, i.e., $\Pr(e|c, \mathbf{q})$, to obtain a new CPT $\Pr'(e|c, \mathbf{q})$, to meet the non-discrimination criterion given by Theorem 3.2, i.e., $|\Pr'(e^+|c^+, \mathbf{q}) - \Pr'(e^+|c^-, \mathbf{q})| < \tau$ for all subpopulations $\mathbf{q}$. The CPTs of all the other nodes are kept unchanged. The joint distribution of the causal graph after the modification can be calculated using the factorization formula (1). After that, the algorithm generates a new dataset based on the modified joint distribution. Since the structure of the causal graph is not changed after the modification, $\mathbf{Q}$ is still the parent set of E excluding C. Thus, according to Theorem 3.2, the newly generated dataset satisfies the non-discrimination criterion.

To achieve a good data utility, we minimize the difference between the original distribution (denoted by P) and the modified distribution (denoted by P'). We use the Euclidean distance, i.e., $d(P', P) = \sqrt{\sum_{\mathbf{V}}(\Pr'(\mathbf{v}) - \Pr(\mathbf{v}))^2}$, to measure the distance between the two distributions. We sort the nodes according to the topological ordering of the graph, and represent the sorted nodes as $\{C, \mathbf{X}, E, \mathbf{Y}\}$. Note that we must have $\mathbf{Q} \subseteq \mathbf{X}$. Then, using the factorization formula (1), $d(P', P)$ can be formulated as

$$d(P', P) = \sqrt{\sum_{C,\mathbf{Q},E} \beta_{\mathbf{q}}^{c,e} \cdot \left(\Pr'(e|c, \mathbf{q}) - \Pr(e|c, \mathbf{q})\right)^2},$$

where $\beta_{\mathbf{q}}^{c,e} = \sum_{\mathbf{x}',\mathbf{y}} \left(\Pr(c)\Pr(\mathbf{x}|c)\Pr(\mathbf{y}|c, \mathbf{x}, e)\right)^2$ and $\mathbf{X}' = \mathbf{X}\backslash\mathbf{Q}$. Thus, the optimal solution (denoted by $\Pr^*(e|c, \mathbf{q})$) that minimizes $d(P', P)$ can be obtained by solving the following quadratic programming problem.

$$\begin{aligned}
\text{minimize} \quad & \sum_{C,\mathbf{Q},E} \beta_{\mathbf{q}}^{c,e} \cdot \left(\Pr'(e|c, \mathbf{q}) - \Pr(e|c, \mathbf{q})\right)^2 \\
\text{subject to} \quad & \forall \mathbf{q}, \quad |\Pr'(e^+|c^+, \mathbf{q}) - \Pr'(e^+|c^-, \mathbf{q})| < \tau, \\
& \forall c, \mathbf{q}, \quad \Pr'(e^-|c, \mathbf{q}) + \Pr'(e^+|c, \mathbf{q}) = 1, \\
& \forall c, \mathbf{q}, e, \quad \Pr'(e|c, \mathbf{q}) > 0.
\end{aligned}$$

The procedure of the algorithm is shown in Algorithm 2.

The complexity of Algorithm 2 depends on the complexity of building the causal graph and solving the quadratic programming. The complexity of building a causal graph has been discussed in Section 3.2. Note that since deriving the objective function needs

Algorithm 1 Certifying of Non-Discrimination (Certify)

Require: dataset $\mathcal{D}$, protected attribute C, decision E, user-defined parameter τ
Ensure: judgment of non-discrimination *judge*, parents of E except C $\mathbf{Q}$
1: $\mathbf{Q} = findParent(E)\backslash\{C\}$
2: **for all** value assignment $\mathbf{q}$ of $\mathbf{Q}$ **do**
3: $|\Delta P|_{\mathbf{q}}| = |\Pr(e^+|c^+, \mathbf{q}) - \Pr(e^+|c^-, \mathbf{q})|$
4: **if** $|\Delta P|_{\mathbf{q}}| \geq \tau$ **then**
5: **return** [*false*, $\mathbf{Q}$]
6: **end if**
7: **end for**
8: **return** [*true*, $\mathbf{Q}$]

Algorithm 2 Removal by Modifying Graph (MGraph)

Require: dataset $\mathcal{D}$, protected attribute C, decision E, user-defined parameter τ
Ensure: modified dataset $\mathcal{D}^*$
1: [*judge*, $\mathbf{Q}$] = *Certify*($\mathcal{D}$, C, E, τ)
2: **if** *judge* == *true* **then**
3: $\mathcal{D}^* = \mathcal{D}$
4: **else**
5: Calculate the modified CPT of E: $\Pr^*(e|c, \mathbf{q})$
6: **for all** $X \in \mathbf{V}\backslash\{E\}$ **do**
7: $\Pr^*(x|\text{Par}(X)) = \Pr(x|\text{Par}(X))$
8: **end for**
9: Calculate P^* using Equation (1)
10: Generate $\mathcal{D}^*$ based on P^*
11: **end if**
12: **return** $\mathcal{D}^*$

information of the whole graph, local causal discovery cannot be used to improve the algorithm. For the quadratic programming, it can be easily shown that, the coefficients of the quadratic terms in the objective function form a positive definite matrix. According to [20], the quadratic programming can be solved in polynomial time.

3.3.2 Discrimination Removal by Modifying Dataset. The second strategy directly modifies the decisions of selected tuples from the dataset to meet the non-discrimination criterion. For each value assignment $\mathbf{q}$, if $\Delta P|_{\mathbf{q}} \geq \tau$, we randomly select[1] a number of tuples with $C = c^-$ and $E = e^-$, and change their E values from e^- to e^+. If $\Delta P|_{\mathbf{q}} \leq -\tau$, we select tuples similarly and change their E values from e^+ to e^-. As result, we ensure $|\Delta P|_{\mathbf{q}}| \leq \tau$ for each $\mathbf{q}$.

One issue here is that, in order to claim non-discrimination for the modified dataset according to Theorem 3.2, $\mathbf{Q}$ should still be the parent set of E excluding C after the modification. For any E's non-decedent X, according to the local Markov condition, X is independent of E in each subpopulation specified by E's parents, i.e., C and $\mathbf{Q}$. Since the modified tuples are randomly selected in the subpopulation specified by C and $\mathbf{Q}$, X would still be independent of E after the modification. Thus, all E's non-decedents would be conditionally independent of E given C and $\mathbf{Q}$, implying that $\mathbf{Q}$ is still the parent set of E excluding C.

To calculate the number of tuples to be modified within each subpopulation $\mathbf{q}$, we express $\Delta P|_{\mathbf{q}}$ as $n_{\mathbf{q}}^{c^+e^+}/n_{\mathbf{q}}^{c^+} - n_{\mathbf{q}}^{c^-e^+}/n_{\mathbf{q}}^{c^-}$. Refer

[1]For ease of representation, in the proposed algorithm we only select the tuples with $C = c^-$. In practice, the selected tuples can come from the group with either $C = c^-$ or $C = c^+$ or both groups.

Table 3: Contingency table within subpopulation q.

	positive decision (e^+)	negative decision (e^-)	total
protected group (c^-)	$n_{\mathbf{q}}^{c^-e^+}$	$n_{\mathbf{q}}^{c^-e^-}$	$n_{\mathbf{q}}^{c^-}$
non-protected group (c^+)	$n_{\mathbf{q}}^{c^+e^+}$	$n_{\mathbf{q}}^{c^+e^-}$	$n_{\mathbf{q}}^{c^+}$
total	$n_{\mathbf{q}}^{e^+}$	$n_{\mathbf{q}}^{e^-}$	$n_{\mathbf{q}}$

to Table 3 for the meaning of the notations. For subpopulations with $\Delta P|_{\mathbf{q}} \geq \tau$, by selecting $\lceil n_{\mathbf{q}}^{c^-} \cdot (|\Delta P|_{\mathbf{q}}| - \tau)\rceil$ tuples with $C = c^-$ and $E = e^-$, and changing their E values from e^- to e^+, the value of $\Delta P|_{\mathbf{q}}$ would decrease by $\lceil n_{\mathbf{q}}^{c^-} \cdot (|\Delta P|_{\mathbf{q}}| - \tau)\rceil / n_{\mathbf{q}}^{c^-} \geq \Delta P|_{\mathbf{q}} - \tau$. Therefore, we have $\Delta P|_{\mathbf{q}} < \tau$ after the modification. The result is similar when $\Delta P|_{\mathbf{q}} \leq -\tau$. The pseudo-code of the algorithm is shown in Algorithm 3.

Algorithm 3 Removal by Modifying Data (MData)

Require: dataset $\mathcal{D}$, protected attribute C, decision E, user-defined parameter τ
Ensure: modified dataset $\mathcal{D}^*$
1: [*judge*, $\mathbf{Q}$] = *Certify*($\mathcal{D}$, C, E, τ)
2: **if** *judge* == *true* **then**
3: $\mathcal{D}^* = \mathcal{D}$
4: **else**
5: **for all** value assignment $\mathbf{q}$ of $\mathbf{Q}$ **do**
6: **if** $\Delta P|_{\mathbf{q}} > \tau$ **then**
7: randomly select a set $\mathbf{T}$ of $\lceil n_{\mathbf{q}}^{c^-} \cdot (|\Delta P|_{\mathbf{q}}| - \tau)\rceil$ tuples with $C = c^-$ and $E = e^-$ in subpopulation $\mathbf{q}$, and change the values of Es from e^- to e^+ to get the set $\mathbf{T}^*$ of the modified tuples
8: **else if** $\Delta P|_{\mathbf{q}} < -\tau$ **then**
9: randomly select a set $\mathbf{T}$ of $\lceil n_{\mathbf{q}}^{c^-} \cdot (|\Delta P|_{\mathbf{q}}| - \tau)\rceil$ tuples with $C = c^-$ and $E = e^+$ in subpopulation $\mathbf{q}$, and change the values of Es from e^+ to e^- to get the set $\mathbf{T}^*$ of the modified tuples
10: **end if**
11: $\mathcal{D}^* = \mathcal{D}^*\backslash\mathbf{T} \cup \mathbf{T}^*$
12: **end for**
13: **end if**
14: **return** $\mathcal{D}^*$

The complexity of Algorithm 3 includes the complexity of finding $\mathbf{Q}$. Similar to Algorithm 1, we can identify E's parents without building the whole graph. Therefore, local discovery algorithms can be employed to improve the efficiency of algorithm. The complexity from Line 5 to 14 is bounded by the size of the original dataset, i.e., $O(|\mathcal{D}|)$.

4 RELAXED NON-DISCRIMINATION CRITERION

So far, we treat dataset $\mathcal{D}$ as the whole population. In real situations, $\mathcal{D}$ may be a sample of the whole population, and $\Delta P|_{\mathbf{b}}$s under a block set $\mathbf{B}$ may vary from one subpopulation to another due to randomness in sampling, especially when the sample size is small. The $|\Delta P|_{\mathbf{b}}|$ values of a few $\mathbf{b}$ could be larger than τ due to the small sample size although the majority of $|\Delta P|_{\mathbf{b}}|$ values are smaller than τ. In this situation, the dataset is claimed as containing discrimination based on the above criterion where all $|\Delta P|_{\mathbf{b}}|$s should be smaller than τ no matter of the majority of $\Delta P|_{\mathbf{b}}$ values.

We propose a relaxed α-non-discrimination criterion which may perform better under the context of randomness and small samples

by finding statistical evidences. Formally, for a given block set $\mathbf{B}$, we treat $\Delta P|_{\mathbf{B}}$ as a variable and treat the values of $\Delta P|_{\mathbf{b}}$s observed across all subpopulations as samples. We introduce a user-defined parameter, α ($0 < \alpha < 1$), to indicate a threshold for the probability of $|\Delta P|_{\mathbf{B}}| < \tau$. If $|\Pr(\Delta P|_{\mathbf{B}}| < \tau) \geq \alpha$, then we say no significant bias is observed under partition $\mathbf{B}$. If $\Pr(|\Delta P|_{\mathbf{B}}| < \tau) \geq \alpha$ holds for each block set, then α-non-discrimination can be claimed for $\mathcal{D}$.

Definition 4.1. Given α, α-non-discrimination is claimed if inequality $\Pr(|\Delta P|_{\mathbf{B}}| < \tau) \geq \alpha$ holds for each block set $\mathbf{B}$.

One challenge here is that we do not know the exact distribution of $\Delta P|_{\mathbf{B}}$ for estimating $\Pr(|\Delta P|_{\mathbf{B}}| < \tau)$ accurately. We propose to employ the Chebyshev's inequality [2], which provides a lower bound of the probability for the value of a random variable lying within a given region, using its mean and variance. Note that the Chebyshev's inequality holds for any random variable irrespective of its distribution. The general form of the Chebyshev's inequality is given as follows.

Theorem 4.2 (Chebyshev's inequality). *Let X be a random variable with finite expected value μ and finite non-zero variance σ^2. Then for any real numbers $a < b$,*

$$\Pr(a < X < b) \geq 1 - \frac{\sigma^2 + (\mu - \frac{b+a}{2})^2}{(\frac{b-a}{2})^2}.$$

The following theorem shows a sufficient condition to satisfy Definition 4.1 using the Chebyshev's inequality.

Theorem 4.3. *Given α, α-non-discrimination is claimed if the following inequality holds for each block set $\mathbf{B}$:*

$$1 - \frac{\sigma^2_{\mathbf{B}} + \mu^2_{\mathbf{B}}}{\tau^2} \geq \alpha,$$

where $\mu_{\mathbf{B}}$ and $\sigma^2_{\mathbf{B}}$ are mean and variance of $\Delta P|_{\mathbf{B}}$.

The proof is straightforward by replacing X with $\Delta P|_{\mathbf{B}}$, a with $-\tau$, and b with τ in the Chebyshev's inequality. We show Theorem 4.3 can be achieved by examining $\mathbf{Q}$ only.

Theorem 4.4. *Given α, α-non-discrimination is claimed if the following inequalities holds for set $\mathbf{Q}$:*

$$1 - \frac{\hat{\sigma}^2_{\mathbf{Q}} + \hat{\mu}^2_{\mathbf{Q}}}{\tau^2} \geq \alpha,$$

where $\hat{\mu}_{\mathbf{B}} = \sum_{\mathbf{B}} \Pr(\mathbf{b}) \cdot \Delta P|_{\mathbf{b}}$ and $\hat{\sigma}^2_{\mathbf{B}} = \sum_{\mathbf{B}} \Pr(\mathbf{b})(\Delta P|_{\mathbf{b}} - \hat{\mu}_{\mathbf{B}})^2$.

Proof. The proof is straightforward by giving two lemmas:

Lemma 4.5. *For each block set $\mathbf{B}$, $\hat{\mu}_{\mathbf{B}} = \hat{\mu}_{\mathbf{Q}}$, where $\mathbf{Q} = \mathrm{Par}(E)\backslash\{C\}$.*

Lemma 4.6. *For each block set $\mathbf{B}$, $\hat{\sigma}^2_{\mathbf{B}} \leq \hat{\sigma}^2_{\mathbf{Q}}$, where $\mathbf{Q} = \mathrm{Par}(E)\backslash\{C\}$.*

Refer to the appendix for proof details. □

5 EXPERIMENTS

We conduct experiments for discrimination discovery and removal algorithms by using two real data sets: the Adult dataset [21] and the Dutch Census of 2001 [25], and compare our algorithms with previous methods proposed in [32] and [9].

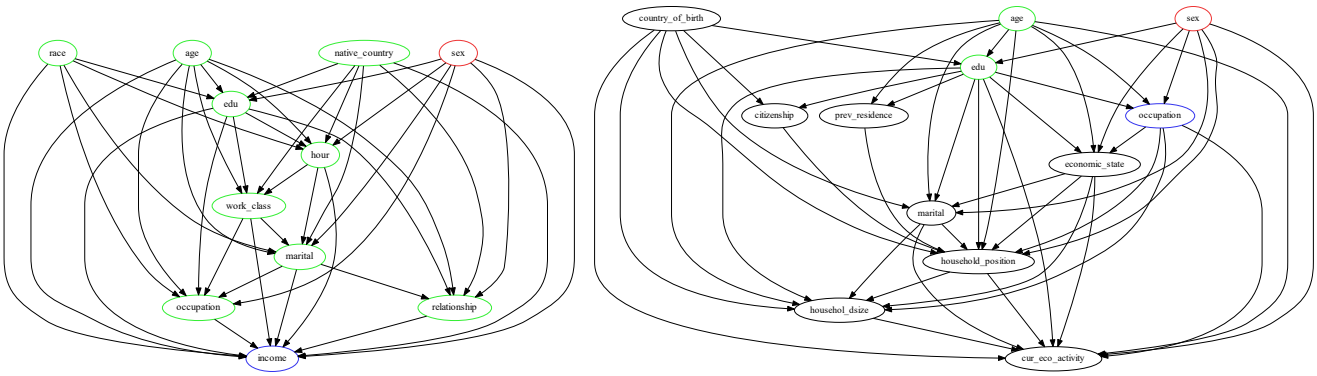
The causal graphs are constructed by utilizing an open-source software TETRAD [11], which is a widely used platform for causal modeling. We employ the original PC algorithm and set the significance threshold 0.01 used for conditional independence testing in causal graph construction. The quadratic programming is solved using CVXOPT [3]. All experiments were conducted with a PC workstation with 16GB RAM and Intel Core i7-4770 CPU. All data and source codes used in the paper are available at https://goo.gl/CjKssf.

5.1 Discrimination Discovery

The Adult dataset consists of 65123 tuples with 11 attributes such as `age`, `edu`, `sex`, `occupation`, `income`, etc. We binarize each attribute's domain values into two classes to reduce the domain sizes because of the sparsity of the dataset. We use three tiers in the partial order for temporal priority: `sex`, `age`, `native_country`, `race` are defined in the first tier, `edu` is defined in the second tier, and all other attributes are defined in the third tier. The constructed causal graph is shown in Figure 2(a). We treat `sex` (female and male) as the protected attribute and `income` (low_income and high_income) as the decision. An arc pointing from `sex` to `income` is observed. We first find set $\mathbf{Q}$ of `income`, which contains all the non-protected attributes. There are 512 subpopulations specified by $\mathbf{Q}$, and 376 subpopulations with non-zero number of tuples due to the sparseness. Then, we compute $\Delta P|_{\mathbf{q}}$ for the 376 subpopulations. On ignoring the subpopulations with the number of females or males smaller than 10, the value of $\Delta P|_{\mathbf{q}}$ ranges from -0.40 to 0.35. Among them, 63 subpopulations have $|\Delta P|_{\mathbf{q}}| > 0.05$, indicating the existence of discrimination in the Adult. Moreover, the mean and standard variance of $\Delta P|_{\mathbf{q}}$ across all subpopulations are 0.011 and 0.142, which has small $\Pr(|\Delta P|_{\mathbf{q}}| < \tau)$ based on the Chebyshev's inequality, e.g., $\Pr(|\Delta P|_{\mathbf{q}}| < 0.15) \geq 9.84\%$. It indicates that the non-discrimination cannot be claimed for the Adult dataset even under the relaxed α-non-discrimination model with large τ and small α.

Another dataset Dutch census consists of 60421 tuples with 12 attributes. Similarly, we binarize the domain values of attribute `age` due to its large domain size. Three tiers are used in the partial order for temporal priority: `sex`, `age`, `country_birth` are in the first tire, `edu` is in the second tire, and all other attributes are in the third tire. The constructed causal graph is shown in Figure 2(b). We treat `sex` (female and male) as the protected attribute and `occupation` (occupation_w_low_income, occupation_w_high_income) as the decision. An arc from `sex` to `occupation` is observed in the causal graph. Set $\mathbf{Q}$ of `occupation` is $\mathbf{Q} = \{\mathtt{edu}, \mathtt{age}\}$. The value of $\Delta P|_{\mathbf{q}}$ ranges from 0.062 to 0.435 across all the 12 subpopulations specified by $\mathbf{Q}$. Thus, discrimination against females is detected in the Dutch census based on the non-discrimination criterion. Moreover, the mean and standard variance of $\Delta P|_{\mathbf{q}}$ are 0.222 and 0.125, which has small $\Pr(|\Delta P|_{\mathbf{q}}| < \tau)$ based on the Chebyshev's inequality, e.g., $\Pr(|\Delta P|_{\mathbf{q}}| < 0.30) \geq 27.94\%$. Hence, the Dutch census still contains discrimination based on the relaxed α-non-discrimination criterion.

Our current implementation uses the PC algorithm to construct the complete causal graph. In our experiment, the PC algorithm with the default significance threshold 0.01 takes 51.59 seconds to build the graph for the binarized Adult dataset and 139.96 seconds for the binarized Dutch census dataset. We also run the PC algorithm on the original Adult dataset, which incurs 4492.36 seconds. In our future work, we will explore the use of the local causal discovery algorithms to improve the efficiency.



(a) Adult (b) Dutch Census

Figure 2: Causal graphs: the red node represents the protected attribute, the blue node represents the decision, the green nodes represent set Q, and the black nodes represent the others.

5.2 Discrimination Removal

Table 4: Comparison of MGraph, MData, Naive, two conditional discrimination removal algorithms (LM and LPS), and disparity impact removal algorithm (DI) on Adult and Dutch Census.

Adult	MGraph	MData	Naive	LM	LPS	DI
$d(\times 10^{-3})$	1.18	10.27	39.35	38.65	35.60	60.65
n_T	1114	4122	29944	16048	16366	44582
χ^2	153	8470	18428	26900	10819	99770
Dutch	MGraph	MData	Naive	LM	LPS	DI
$d(\times 10^{-3})$	5.68	6.75	13.91	18.00	15.48	14.10
n_T	10422	8838	32516	29288	24648	35728
χ^2	2832	4825	14014	10555	5039	19684

After executing our two proposed discrimination removal algorithms, MGraph and MData, the value of $|\Delta P|_{\mathbf{q}}|$ of all subpopulations are within 0.05 for both datasets. The performance of the algorithms in terms of the utility of the modified data is shown in Table 4. We also report the results from the Naive method used in [9] in which we completely reshuffle the gender information. We measure the utility by three metrics: the Euclidean distance (d), the number of modified tuples (n_T), and the utility loss (χ^2). We can observe from Table 4 that the MGraph algorithm retains the highest utility. Both MGraph and MData algorithms significantly outperform the Naive method. We also examine how utility in terms of three metrics vary with different τ values for our MGraph and MData algorithms. We can see from Table 5 that both discrimination removal algorithms incur less utility loss with larger τ values. This observation validates our analysis of non-discrimination model.

We measure the execution times of our removal algorithms. As expected, MGraph takes longer time than MData since the former requires quadratic programming and data generation based on the whole modified graph while the latter only requires the information of **Q**. For the Adult dataset with $\tau = 0.05$, MGraph takes 3.42s while MData takes 0.82s. For the Dutch dataset the difference is even larger, i.e., 36.3s for MGraph and 0.07s for MData, since the size of **Q** of Dutch census is much smaller.

Table 5: Comparison of utility with varied τ values for MGraph and MData.

Adult	MGraph				MData			
τ	0.025	0.050	0.075	0.100	0.025	0.050	0.075	0.100
$d(\times 10^{-3})$	1.56	1.18	0.91	0.69	10.95	10.27	5.19	4.83
n_T	1390	1114	932	792	4900	4122	2672	2312
χ^2	228	153	107	74	9255	8470	4340	3103
Dutch	MGraph				MData			
τ	0.025	0.050	0.075	0.100	0.025	0.050	0.075	0.100
$d(\times 10^{-3})$	6.17	5.68	5.20	4.62	7.39	6.75	6.12	5.54
n_T	12024	10422	9076	8016	10114	8838	7702	6800
χ^2	3310	2832	2334	1905	6158	4825	3869	3082

5.3 Comparison with Previous Work

In [32], the authors measured the "bad" discrimination i.e., the effect that can be explained by conditioning on one attribute. They developed two methods, local massaging (LM) and local preferential sampling (LPS), to remove the unexplainable (bad) discrimination when one of the attributes is considered to be explanatory for the discrimination. However, their methods do not distinguish whether a partition is meaningful or not. Therefore, they cannot decrease and even increase the number of subpopulations with discrimination. Our experiments show that, their methods cannot remove discrimination conditioning on any single attribute. The results are skipped due to space limitation. In addition, even if we remove "bad" discrimination using their methods by conditioning on each attribute one by one, a significant amount of discriminatory effects still exist. After running the local massaging (LM) method, there are 104 subpopulations with discrimination for Adult and 9 subpopulations with discrimination for Dutch census. The local preferential sampling (LPS) method performs also poor — there are 77 subpopultions with discrimination for Adult and 4 subpopulations with discrimination for Dutch census. This is because for both datasets,

any single attribute is not a block set and hence does not form a meaningful partition. Even assuming each attribute forms a meaningful partition, removing discrimination for each partition one by one does not guarantee to remove discrimination since the modification under one partition may change the distributions under other partitions.

In [9], the authors studied how to remove disparate impact (measured by risk ratio) from the data. They proposed to test disparate impact based on how well the protected attribute C can be predicted with the non-protected attributes, and remove disparate impact by modifying the non-protected attributes while keeping the decision values E unchanged. Although the approach can completely remove the discrimination measured by disparate impact at the whole dataset level, it cannot remove discrimination from each meaningful partition. After running their repair method (DI), there are 127 subpopulations with discrimination for Adult and 12 subpopulations with discrimination for Dutch Census.

Differently from previous work, our approaches remove discrimination based on block set $\mathbf{Q}$ and ensure that there is no discrimination in all meaningful partitions. Moreover, our approaches keep the causal structure unchanged after the modification and preserve data utility. All previous methods, local massaging (LM), local preferential sampling (LPS), and disparity impact removal method (DI) incur much larger utility loss (as shown in the last three columns of Table 4) than our algorithms.

6 RELATED WORK

A number of data mining techniques have been proposed to discover discrimination in the literature. Classification rule-based methods [26, 27, 29] were proposed to represent certain discrimination patterns. If the presence of the protective attribute increases the confidence of a classification rule, it indicates possible discrimination in the data set. Based on that, the authors in [23] further proposed to use the Bayesian network to compute the confidence of the classification rules for detecting discrimination. In [22], the authors dealt with the individual discrimination by finding a group of similar individuals. When there are significantly different decision outcomes between the individuals from the protected group and the individuals from the non-protected group, the difference implies discrimination. Conditional discrimination, i.e., part of discrimination may be explained by other legally grounded attributes, was studied in [32]. The task was to evaluate to which extent the discrimination apparent for a group is explainable on a legal ground. However, their method does not distinguish whether a partition is meaningful for measuring discrimination, and it also does not consider the situation where there exist multiple meaningful partitions.

Proposed methods for discrimination prevention are either based on data preprocessing or algorithm tweaking. Data preprocessing methods [1, 9, 12, 16, 23, 32, 34] modify the historic data to remove discriminatory effect before conducting predictive analysis. For example, in [16] several methods for modifying data were proposed, including *Massaging*, which corrects the labels of some individuals in the data, *Reweighting*, which assigns weights to individuals to balance the data, and *Sampling*, which changes the sample sizes of different subgroups to remove the bias in the data. In [9], the authors studied how to remove disparate impact, i.e., indirect discrimination. The authors first remove direct discrimination by completely deleting the protected attribute from the data. Then, they test disparate impact based on how well the protected attribute can be predicted with the non-protected attributes, and remove disparate impact by modifying the distribution of the non-protected attributes such that the protected attribute cannot be estimated from the non-protected attributes. As shown by our experiments, removing the protected attribute from the released data would significantly damage the data utility. Proposed methods for discrimination prevention using algorithm tweaking require some tweak of predictive models, including the decision tree [17], the naive bayes classifier [5], the logistic regression [18], and the log-linear model [33]. In [8], the authors addressed the problem of constructing a predictive model that achieves both statistical parity and individual fairness, i.e., similar individuals should be treated similarly. In [13], the authors proposed to sanitize discriminatory patterns by incorporating the privacy preserving methods. In [14], the authors proposed a framework for optimally adjusting any predictive model so as to remove discrimination. All the above works are based on correlations rather than the causation.

Most recently, several studies have been devoted to analyzing discrimination from the causal perspective. The authors in [4] proposed a framework based on the Suppes-Bayes causal network and developed several random-walk-based methods to detect different types of discrimination. However, the construction of the Suppes-Bayes causal network is impractical with the large number of attribute-value pairs. In addition, it is unclear how the number of random walks is related to practical discrimination metrics, e.g., the difference in positive decision rates τ. Studies in [36–39] are built on the causal graph and causal modeling, and [35] provides a review of these works. They focused on detecting and removing either discrimination at the individual level or discrimination at the whole dataset level. Differently, this paper studies how to remove discrimination from all meaningful partitions of a dataset.

7 CONCLUSIONS

In this paper, we have investigated how to detect and remove discrimination from all meaningful partitions. With the support of the causal graph, we have shown that the discriminatory effect can only be identified under the partition defined by the block set. We have provided the condition for the block set. Based on that, we have developed a simple non-discrimination criterion and two strategies for removing discrimination. We also proposed a relaxed non-discrimination criterion to deal with sampling randomness in the data. The experiment results using real datasets show that our proposed approaches are effective in discovering and completely removing discrimination.

Appendices

Proof of Lemma 3.3

Proof. We first sort the nodes in the causal graph according to the topological ordering of the DAG, so that for each sorted pair of nodes X and Y that X is ahead of Y, X must be Y's non-descendent

and Y must be X's non-ancestor. The topological ordering is guaranteed to be found in a DAG [7]. We represent the sorted nodes by an ordered list $\{\cdots, C, \cdots, E, \cdots\}$. According to the local Markov condition, we have

$$\Pr(V|\text{Prior}(V)) = \Pr(V|\text{Par}(V)), \tag{2}$$

where $\text{Prior}(V)$ represents all the nodes prior to V in the ordering. Now we consider a topological ordering such that, (i) node E and all nodes in $\mathbf{Q}$ are consecutive in ordering, (ii) all nodes posterior to E are E's descendents. It is easy to prove that such a topological ordering can always be constructed.[2] Denote by $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ the set of nodes that are prior to C, between C and $\mathbf{Q}$, and posterior to E respectively. The topological ordering can be represented as the list $\{\mathbf{X}, C, \mathbf{Y}, \mathbf{Q}, E, \mathbf{Z}\}$. According to the definition of the block set, $\mathbf{B}$ contains no node in $\mathbf{Z}$. Thus, $\mathbf{B} \subseteq \mathbf{X} \cup \mathbf{Y} \cup \mathbf{Q}$. We define $\mathbf{X}' = \mathbf{X} \backslash \mathbf{B}$, $\mathbf{Y}' = \mathbf{Y} \backslash \mathbf{B}$, $\mathbf{Q}' = \mathbf{Q} \backslash \mathbf{B}$. Since sets $\mathbf{X}, \mathbf{Y}, \mathbf{Q}$ are mutually exclusive, we have $\mathbf{B} = (\mathbf{X} \backslash \mathbf{X}') \cup (\mathbf{Y} \backslash \mathbf{Y}') \cup (\mathbf{Q} \backslash \mathbf{Q}')$, which entails that

$$(\mathbf{X} \backslash \mathbf{X}') \cup (\mathbf{Y} \backslash \mathbf{Y}') = \mathbf{B} \cup \mathbf{Q}'. \tag{3}$$

From probability theories, we have

$$\Pr(e^+|c^+, \mathbf{b}) = \frac{\Pr(e^+, c^+, \mathbf{b})}{\Pr(c^+, \mathbf{b})} = \frac{1}{\Pr(c^+, \mathbf{b})} \sum_{\mathbf{X}', \mathbf{Y}', \mathbf{Q}', \mathbf{Z}} \Pr(\mathbf{x}, c^+, \mathbf{y}, \mathbf{q}, e^+, \mathbf{z}).$$

According to the chain rule of probability calculus, we have

$$\begin{aligned}
&\Pr(e^+|c^+, \mathbf{b}) \\
&= \frac{1}{\Pr(c^+, \mathbf{b})} \sum_{\mathbf{X}', \mathbf{Y}', \mathbf{Q}', \mathbf{Z}} \Pr(\mathbf{x}, c^+, \mathbf{y}, \mathbf{q}) \cdot \Pr(e^+|\text{Prior}(E)) \cdot \Pr(\mathbf{z}|\text{Prior}(\mathbf{Z})) \\
&= \frac{1}{\Pr(c^+, \mathbf{b})} \sum_{\mathbf{X}', \mathbf{Y}', \mathbf{Q}'} \Pr(\mathbf{x}, c^+, \mathbf{y}, \mathbf{q}) \cdot \Pr(e^+|\text{Prior}(E)).
\end{aligned}$$

From Equation (2), it follows that

$$\begin{aligned}
\Pr(e^+|c^+, \mathbf{b}) &= \frac{1}{\Pr(c^+, \mathbf{b})} \sum_{\mathbf{X}', \mathbf{Y}', \mathbf{Q}'} \Pr(\mathbf{x}, c^+, \mathbf{y}, \mathbf{q}) \cdot \Pr(e^+|\text{Par}(E)) \\
&= \frac{1}{\Pr(c^+, \mathbf{b})} \sum_{\mathbf{X}', \mathbf{Y}', \mathbf{Q}'} \Pr(\mathbf{x}, c^+, \mathbf{y}, \mathbf{q}) \cdot \Pr(e^+|c^+, \mathbf{q}) \\
&= \frac{1}{\Pr(c^+, \mathbf{b})} \sum_{\mathbf{Q}'} \Big\{ \Pr(e^+|c^+, \mathbf{q}) \cdot \sum_{\mathbf{X}', \mathbf{Y}'} \Pr(\mathbf{x}, c^+, \mathbf{y}, \mathbf{q}) \Big\}.
\end{aligned}$$

From Equation (3), we have

$$\begin{aligned}
\Pr(e^+|c^+, \mathbf{b}) &= \frac{1}{\Pr(c^+, \mathbf{b})} \sum_{\mathbf{Q}'} \Pr(e^+|c^+, \mathbf{q}) \cdot \Pr(c^+, \mathbf{b}, \mathbf{q}) \\
&= \sum_{\mathbf{Q}'} \Pr(\mathbf{q}|c^+, \mathbf{b}) \cdot \Pr(e^+|c^+, \mathbf{q}) = \sum_{\mathbf{Q}'} \Pr(\mathbf{q}'|c^+, \mathbf{b}) \cdot \Pr(e^+|c^+, \mathbf{q}).
\end{aligned}$$

If $(C \perp\!\!\!\perp \mathbf{Q}' \mid \mathbf{B})_{\mathcal{G}'}$, then we can find a path from C to E through $\mathbf{Q}'$ that is not blocked, which means that $(C \perp\!\!\!\perp E \mid \mathbf{B})_{\mathcal{G}'}$. This contradicts $\mathbf{B}$ being a block set. Therefore, we must have $(C \perp\!\!\!\perp \mathbf{Q}' \mid \mathbf{B})_{\mathcal{G}'}$, which entails $(C \perp\!\!\!\perp \mathbf{Q}' \mid \mathbf{B})_{\mathcal{G}}$ according to the d-separation criterion. Thus, it follows that

$$\Pr(e^+|c^+, \mathbf{b}) = \sum_{\mathbf{Q}'} \Pr(\mathbf{q}'|\mathbf{b}) \cdot \Pr(e^+|c^+, \mathbf{q}).$$

[2]For (i), if any node lies between E and some of its parents, we can move the node to the front of all E's parents and the resultant list is still a topological ordering. Similarly we can prove (ii).

We can obtain similar result for $\Pr(e^+|c^-, \mathbf{b})$. Therefore, we have

$$\Delta P|_{\mathbf{b}} = \sum_{\mathbf{Q}'} \Pr(\mathbf{q}'|\mathbf{b}) \cdot \Delta P|_{\mathbf{q}}. \tag{4}$$

Hence, the lemma is proven. □

Proof of Lemma 3.4

Proof. We classify the paths from C to E other than arc $C \to E$ into two cases based on the last node X ahead of E on the path. For the first case, X is a parent of E. Thus, X is a noncollider and belongs to $\mathbf{Q}$. Based on the definition, each path in the first case is blocked by $\mathbf{Q}$. For the second case, X is a child of E. Then, there must be at least one collider Y on each path in the second case. Otherwise, the path is mono-directional with all the arcs pointing from E to C, forming a circle with the arc $C \to E$. This contradicts to that a CBN is a directed acyclic graph. Let Y be the last collider ahead of E on a path. Then, neither Y nor its descendant Z can be E's parent. Otherwise, mono-directional path $E \to \cdots \to Y \to E$ or $E \to \cdots \to Y \to \cdots \to Z \to E$ forms a circle, which again contradicts to that a CBN is a directed acyclic graph. Thus, according to the definition, each path in the second case is blocked by $\mathbf{Q}$. Finally, $\mathbf{Q}$ contains none of E's descendents. Therefore, $\mathbf{Q}$ is a block set. Hence, the lemma is proven. □

Proof of Lemma 4.5

Proof. By definition, we have

$$\hat{\mu}_{\mathbf{B}} = \sum_{\mathbf{B}} \Pr(\mathbf{b}) \cdot \Delta P|_{\mathbf{b}}.$$

According to Equation (4), we have

$$\hat{\mu}_{\mathbf{B}} = \sum_{\mathbf{B}} \Pr(\mathbf{b}) \cdot \sum_{\mathbf{Q}'} \Pr(\mathbf{q}'|\mathbf{b}) \cdot \Delta P|_{\mathbf{q}},$$

where $\mathbf{Q}' = \mathbf{Q} \backslash \mathbf{B}$. It follows that

$$\begin{aligned}
\hat{\mu}_{\mathbf{B}} &= \sum_{\mathbf{B}, \mathbf{Q}'} \Pr(\mathbf{b}) \cdot \Pr(\mathbf{q}'|\mathbf{b}) \cdot \Delta P|_{\mathbf{q}} = \sum_{\mathbf{B}, \mathbf{Q}'} \Pr(\mathbf{b}, \mathbf{q}') \cdot \Delta P|_{\mathbf{q}} \\
&= \sum_{\mathbf{X} = \mathbf{B} \cup \mathbf{Q}'} \Pr(\mathbf{x}) \cdot \Delta P|_{\mathbf{q}} = \sum_{\mathbf{B}', \mathbf{Q}} \Pr(\mathbf{b}', \mathbf{q}) \cdot \Delta P|_{\mathbf{q}},
\end{aligned}$$

where $\mathbf{B}' = \mathbf{B} \backslash \mathbf{Q}$. Then, it follows that

$$\hat{\mu}_{\mathbf{B}} = \sum_{\mathbf{Q}} \Delta P|_{\mathbf{q}} \cdot \sum_{\mathbf{B}'} \Pr(\mathbf{b}', \mathbf{q}) = \sum_{\mathbf{Q}} \Delta P|_{\mathbf{q}} \cdot \Pr(\mathbf{q}) = \mu_{\mathbf{Q}}.$$

Hence, the lemma is proven. □

Proof of Lemma 4.6

Proof. By definition, we have

$$\begin{aligned}
\hat{\sigma}^2_{\mathbf{B}} &= \sum_{\mathbf{B}} \Pr(\mathbf{b})(\Delta P|_{\mathbf{b}} - \hat{\mu}_{\mathbf{B}})^2 \\
&= \sum_{\mathbf{B}} \Pr(\mathbf{b}) \Big((\Delta P|_{\mathbf{b}})^2 - 2\hat{\mu}_{\mathbf{B}} \Delta P|_{\mathbf{b}} + \hat{\mu}^2_{\mathbf{B}} \Big) \\
&= \sum_{\mathbf{B}} \Pr(\mathbf{b})(\Delta P|_{\mathbf{b}})^2 - 2\hat{\mu}_{\mathbf{B}} \sum_{\mathbf{B}} \Pr(\mathbf{b}) \Delta P|_{\mathbf{b}} + \hat{\mu}^2_{\mathbf{B}} \sum_{\mathbf{B}} \Pr(\mathbf{b}).
\end{aligned}$$

According to Equation (4), we have

$$\sum_{\mathbf{B}} \Pr(\mathbf{b}) \cdot (\Delta P|_{\mathbf{b}})^2 = \sum_{\mathbf{B}} \Pr(\mathbf{b}) \cdot \Big(\sum_{\mathbf{Q}'} \Pr(\mathbf{q}'|\mathbf{b}) \cdot \Delta P|_{\mathbf{q}} \Big)^2$$
$$= \sum_{\mathbf{B}} \Pr(\mathbf{b}) \cdot \Big(\sum_{\mathbf{Q}'} \sqrt{\Pr(\mathbf{q}'|\mathbf{b})} \cdot \sqrt{\Pr(\mathbf{q}'|\mathbf{b})} \Delta P|_{\mathbf{q}} \Big)^2.$$

According to Cauchy's Inequality, it follows that

$$\sum_{\mathbf{B}} \Pr(\mathbf{b}) \cdot (\Delta P|_{\mathbf{b}})^2$$
$$\leq \sum_{\mathbf{B}} \Pr(\mathbf{b}) \cdot \Big(\sum_{\mathbf{Q}'} \Pr(\mathbf{q}'|\mathbf{b}) \Big) \cdot \Big(\sum_{\mathbf{Q}'} \Pr(\mathbf{q}'|\mathbf{b}) \cdot (\Delta P|_{\mathbf{q}})^2 \Big)$$
$$= \sum_{\mathbf{B}} \Pr(\mathbf{b}) \cdot \Big(\sum_{\mathbf{Q}'} \Pr(\mathbf{q}'|\mathbf{b}) \cdot (\Delta P|_{\mathbf{q}})^2 \Big).$$

Similar to the proof of Lemma 4.5, it follows that

$$\sum_{\mathbf{B}} \Pr(\mathbf{b}) \cdot (\Delta P|_{\mathbf{b}})^2 \leq \sum_{\mathbf{B},\mathbf{Q}'} \Pr(\mathbf{b}, \mathbf{q}') \cdot (\Delta P|_{\mathbf{q}})^2$$
$$= \sum_{\mathbf{X}=\mathbf{B}\cup\mathbf{Q}'} \Pr(\mathbf{x}) \cdot (\Delta P|_{\mathbf{q}})^2 = \sum_{\mathbf{B}',\mathbf{Q}} \Pr(\mathbf{b}', \mathbf{q}) \cdot (\Delta P|_{\mathbf{q}})^2 = \sum_{\mathbf{Q}} \Pr(\mathbf{q}) \cdot (\Delta P|_{\mathbf{q}})^2.$$

Hence, we have

$$\sum_{\mathbf{B}} \Pr(\mathbf{b}) \cdot (\Delta P|_{\mathbf{b}})^2 \leq \sum_{\mathbf{Q}} \Pr(\mathbf{q}) \cdot (\Delta P|_{\mathbf{q}})^2.$$

According to Lemma 4.5, we have

$$\hat{\mu}_{\mathbf{B}} = \sum_{\mathbf{B}} \Pr(\mathbf{b}) \cdot \Delta P|_{\mathbf{b}} = \hat{\mu}_{\mathbf{Q}} = \sum_{\mathbf{Q}} \Pr(\mathbf{q}) \cdot \Delta P|_{\mathbf{q}}.$$

Besides, we have

$$\sum_{\mathbf{B}} \Pr(\mathbf{b}) = \sum_{\mathbf{Q}} \Pr(\mathbf{q}) = 1.$$

Thus, it follows that

$$\hat{\sigma}^2_{\mathbf{B}} \leq \sum_{\mathbf{Q}} \Pr(\mathbf{q}) (\Delta P|_{\mathbf{q}})^2 - 2\hat{\mu}_{\mathbf{Q}} \sum_{\mathbf{Q}} \Pr(\mathbf{q}) \Delta P|_{\mathbf{q}} + \hat{\mu}^2_{\mathbf{Q}} \sum_{\mathbf{Q}} \Pr(\mathbf{q}) = \hat{\sigma}^2_{\mathbf{Q}}.$$

Hence, the lemma is proven. □

ACKNOWLEDGMENTS

This work was supported in part by NSF 1646654.

REFERENCES

[1] Philip Adler, Casey Falk, Sorelle A Friedler, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2016. Auditing black-box models for indirect influence. In *ICDM'16*. IEEE, 1–10.

[2] George A Anastassiou. 2009. *Probabilistic inequalities*. World Scientific.

[3] Martin Andersen, Joachim Dahl, and Lieven Vandenberghe. 2004. CVXOPT. http://cvxopt.org/. (2004).

[4] Francesco Bonchi, Sara Hajian, Bud Mishra, and Daniele Ramazzotti. 2017. Exposing the probabilistic causal structure of discrimination. *International Journal of Data Science and Analytics* 3, 1 (2017), 1–21.

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

[6] Diego Colombo and Marloes H Maathuis. 2014. Order-independent constraint-based causal structure learning. *JMLR* 15, 1 (2014), 3741–3782.

[7] Thomas H Cormen. 2009. *Introduction to algorithms*. MIT press.

[8] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of ITCS*. ACM, 214–226.

[9] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *SIGKDD'15*. ACM, 259–268.

[10] Sheila R Foster. 2004. Causation in Antidiscrimination Law: Beyond Intent Versus Impact. *Hous. L. Rev.* 41 (2004), 1469.

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

[12] Sara Hajian and Josep Domingo-Ferrer. 2013. A methodology for direct and indirect discrimination prevention in data mining. *JKDE* 25, 7 (2013), 1445–1459.

[13] Sara Hajian, Josep Domingo-Ferrer, Anna Monreale, Dino Pedreschi, and Fosca Giannotti. 2015. Discrimination-and privacy-aware patterns. *Data Mining and Knowledge Discovery* 29, 6 (2015), 1733–1782.

[14] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *NIPS'16*. 3315–3323.

[15] Markus Kalisch and Peter Bühlmann. 2007. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research* 8 (2007), 613–636.

[16] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *KAIS* 33, 1 (2012), 1–33.

[17] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination aware decision tree learning. In *ICDM'10*. IEEE, 869–874.

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

[19] Daphne Koller and Nir Friedman. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

[20] Mikhail K Kozlov, Sergei P Tarasov, and Leonid G Khachiyan. 1980. The polynomial solvability of convex quadratic programming. *U. S. S. R. Comput. Math. and Math. Phys.* 20, 5 (1980), 223–228.

[21] M. Lichman. 2013. UCI Machine learning repository. http://archive.ics.uci.edu/ml. (2013).

[22] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *SIGKDD'11*. ACM, 502–510.

[23] Koray Mancuhan and Chris Clifton. 2014. Combating discrimination using Bayesian networks. *Artificial intelligence and law* 22, 2 (2014), 211–238.

[24] Richard E Neapolitan and others. 2004. *Learning bayesian networks*. Vol. 38. Prentice Hall Upper Saddle River.

[25] Statistics Netherlands. 2001. Volkstelling. https://sites.google.com/site/faisalkamiran/. (2001).

[26] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2009. Measuring Discrimination in Socially-Sensitive Decision Records. In *SIAM SDM*. Society for Industrial and Applied Mathematics, 581.

[27] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *SIGKDD'08*. ACM, 560–568.

[28] Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* 29, 05 (2014), 582–638.

[29] Salvatore Ruggieri, Dino Pedreschi, and Franco Turini. 2010. Data mining for discrimination discovery. *TKDD* 4, 2 (2010), 9.

[30] Peter Spirtes, Clark N Glymour, and Richard Scheines. 2000. *Causation, prediction, and search*. Vol. 81. MIT press.

[31] Alexander Statnikov, Jan Lemeir, and Constantin F Aliferis. 2013. Algorithms for discovery of multiple Markov boundaries. *JMLR* 14, 1 (2013), 499–566.

[32] Indre Žliobaitė, Faisal Kamiran, and Toon Calders. 2011. Handling conditional discrimination. In *ICDM'11*. IEEE, 992–1001.

[33] Yongkai Wu and Xintao Wu. 2016. Using loglinear model for discrimination discovery and prevention. In *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*. IEEE, 110–119.

[34] Richard S Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. *ICML* 28 (2013), 325–333.

[35] Lu Zhang and Xintao Wu. 2017. Anti-discrimination learning: a causal modeling-based framework. *International Journal of Data Science and Analytics* (2017), 1–16.

[36] Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. On discrimination discovery using causal networks. In *SBP-BRiMS 2016*.

[37] Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. Situation testing-based discrimination discovery: a causal inference approach. In *IJCAI'16*.

[38] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. Achieving non-discrimination in prediction. *arXiv preprint arXiv:1703.00060* (2017).

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