Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)

A Causal Framework for Discovering and Removing Direct and Indirect Discrimination

Lu Zhang, Yongkai Wu, and Xintao Wu
University of Arkansas
{lz006,yw009,xintaowu}@uark.edu

Abstract

In this paper, we investigate the problem of discovering both direct and indirect discrimination from the historical data, and removing the discriminatory effects before the data is used for predictive analysis (e.g., building classifiers). The main drawback of existing methods is that they cannot distinguish the part of influence that is really caused by discrimination from all correlated influences. In our approach, we make use of the causal network to capture the causal structure of the data. Then we model direct and indirect discrimination as the *path-specific effects*, which accurately identify the two types of discrimination as the causal effects transmitted along different paths in the network. Based on that, we propose an effective algorithm for discovering direct and indirect discrimination, as well as an algorithm for precisely removing both types of discrimination while retaining good data utility. Experiments using the real dataset show the effectiveness of our approaches.

1 Introduction

Discrimination refers to unjustified distinctions in decisions against individuals based on their membership in a certain group. Laws and regulations have been established to prohibit discrimination on several grounds, such as gender, age, sexual orientation, race, religion, and disability, which are referred to as the *protected attributes*. Various predictive models have been built around the collection and use of historical data to make important decisions like employment, credit and insurance. If the historical data contains discrimination, the predictive models are likely to learn the discriminatory relationship present in the historical data and apply it when making new decisions. Therefore, it is imperative to ensure that the data goes into the predictive models and the decisions made with its assistance are not subject to discrimination.

In the legal field, discrimination falls into direct and indirect discrimination. Direct discrimination occurs when individuals receive less favorable treatment explicitly based on the protected attributes. An example would be rejecting a qualified female applicant in applying a university just because of her gender. Indirect discrimination refers to the situation where the treatment is based on apparently neutral non-protected attributes but still results in unjustified distinctions against individuals from the protected group. A well-known example of indirect discrimination is redlining, where the residential Zip Code of the individual is used for making decisions such as granting a loan. Although Zip Code is apparently a neutral attribute, it correlates with race due to the racial composition of residential areas. Thus, the use of Zip Code may indirectly lead to racial discrimination.

Discrimination discovery and removal from historical data has received an increasing attention over the past few years in data science [Hajian and Domingo-Ferrer, 2013; Kamiran and Calders, 2012; Ruggieri *et al.*, 2010; Romei and Ruggieri, 2014; Feldman *et al.*, 2015]. Many approaches have been proposed to deal with both direct and indirect discrimination but significant issues exist. For discrimination discovery, the difference in decisions across the protected and non-protected groups is a combined (not necessarily linear) effect of direct discrimination, indirect discrimination, and explainable effect that should not be considered as discrimination (e.g., the difference in average income of females and males caused by their different working hours per week). However, existing methods cannot explicitly and correctly identify the three different effects when measuring discrimination. For example, the classic metrics *risk difference*, *risk ratio*, *relative chance*, *odds ratio*, etc. [Romei and Ruggieri, 2014] treat all the difference in decisions as discrimination. [Žliobaitė *et al.*, 2011] realized the explainable effect but failed to distinguish the effects of direct and indirect discrimination. For discrimination removal, a general requirement is to preserve the data utility while achieving non-discrimination. As we shall show in the experiments, a crude method that totally removes all connections between the protected attribute and decision (e.g., in [Feldman *et al.*, 2015]) can eliminate discrimination but may suffer significant utility loss. To maximize the preserved data utility, it is necessary to first accurately measure the discriminatory effects and then precisely remove them.

The causal modeling based discrimination detection has been proposed most recently [Zhang *et al.*, 2016c; 2016b] for improving the correlation based approaches. However, these work also do not tackle indirect discrimination. In this paper, we develop a framework for discovering and removing both direct and indirect discrimination based on the causal network. A causal network is a directed acyclic graph

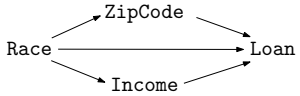

Figure 1: The toy model.

(DAG) widely used for causal representation, reasoning and inference [Pearl, 2009], where causal effects are carried by the *causal paths* that trace arrows pointing from the cause to the effect. Using this model, direct and indirect discrimination can be respectively captured by the causal effects of the protected attribute on the decision transmitted along different causal paths. To be specific, direct discrimination is modeled as the causal effect transmitted along the direct path from the protected attribute to the decision. Indirect discrimination, on the other hand, is modeled as the causal effect transmitted along other causal paths that contain any unjustified attribute. For example, consider a toy model of a loan application system shown in Figure 1. Assume that we treat `Race` as the protected attribute, `Loan` as the decision, and `ZipCode` as the unjustified attribute that triggers redlining. Direct discrimination is then transmitted along path `Race` → `Loan`, and indirect discrimination is transmitted along path `Race` → `ZipCode` → `Loan`. Assume that the use of `Income` can be objectively justified as it is reasonable to deny a loan if the applicant has low income. In this case, path `Race` → `Income` → `Loan` is explainable, which means that part of the difference in loan issuance across different race groups can be explained by the fact that some race groups in the dataset tend to be under-paid.

Our analysis shows that measuring discrimination based on the causal network requires to measure the causal effect transmitted along certain causal paths. To this end, we employ the technique of the *path-specific effect* [Avin *et al.*, 2005; Shpitser, 2013]. We define direct/indirect discrimination as different path-specific effects, and show how to measure them using the observational data. Based on that, we propose an effective algorithm for discovering direct/indirect discrimination, as well as an algorithm for precisely removing both types of discrimination while retaining good data utility. The experiments using the real dataset show that our approaches are effective in discovering and removing discrimination.

2 Preliminary Concepts

A causal network is a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{A})$ where $\mathbf{V}$ is a set of nodes and $\mathbf{A}$ is a set of arcs. Each node in the network represents an attribute. Each arc, denoted by an arrow → pointing from the cause to the effect represents the direct causal relationship. Throughout the paper, we denote an attribute by an uppercase alphabet, e.g., X; denote a subset of attributes by a bold uppercase alphabet, e.g., $\mathbf{X}$. We denote a domain value of attribute X by a lowercase alphabet, e.g., x; denote a value assignment of attributes $\mathbf{X}$ by a bold lowercase alphabet, e.g., $\mathbf{x}$. For a node X, its parents are denoted by $Pa(X)$, and its children are denoted by $Ch(X)$. Each node is associated with a conditional probability table (CPT), i.e., $P(x|Pa(X))$. The joint distribution over all attributes $P(\mathbf{v})$ can be computed using the factorization formula [Koller and Friedman, 2009]

$$P(\mathbf{v}) = \prod_{V \in \mathbf{V}} P(v|Pa(V)), \tag{1}$$

where $P(v|Pa(V))$ is the observational distribution.

In the causal network, measuring causal effects is facilitated with the *do*-calculus [Pearl, 2009], which simulates the physical interventions that force some attributes $\mathbf{X}$ to take certain values $\mathbf{x}$. The post-intervention distributions represent the effect of the intervention. Formally, the intervention that sets the value of $\mathbf{X}$ to $\mathbf{x}$ is denoted by $do(\mathbf{X} = \mathbf{x})$. The post-intervention distribution of all other attributes $\mathbf{Y} = \mathbf{V}\backslash\mathbf{X}$, i.e., $P(\mathbf{Y} = \mathbf{y}|do(\mathbf{X} = \mathbf{x}))$ or simply $P(\mathbf{y}|do(\mathbf{x}))$, can be computed by the truncated factorization formula [Pearl, 2009]

$$P(\mathbf{y}|do(\mathbf{x})) = \prod_{Y \in \mathbf{Y}} P(y|Pa(Y))\delta_{\mathbf{X}=\mathbf{x}}, \tag{2}$$

where $\delta_{\mathbf{X}=\mathbf{x}}$ means assigning attributes in $\mathbf{X}$ involved in the term ahead with the corresponding values in $\mathbf{x}$. Specifically, the post-intervention distribution of a single attribute Y given an intervention on a single attribute X is given by

$$P(y|do(x)) = \sum_{\mathbf{V}\backslash\{X,Y\}, Y=y} \prod_{V \in \mathbf{V}\backslash\{X\}} P(v|Pa(V))\delta_{X=x}, \tag{3}$$

where the summation is a marginalization that traverses all value combinations of $\mathbf{V}\backslash\{X, Y\}$.

By using the *do*-calculus, the total causal effect of X on Y is defined in Definition 1 [Pearl, 2009]. Note that in this definition, the effect of the intervention is transmitted along all causal paths from the cause X to the effect Y.

Definition 1 (Total causal effect) *The total causal effect measures the effect of the change of X from x_1 to x_2 on $Y = y$ transmitted along all causal paths from X to Y. It is given by*

$$TE(x_2, x_1) = P(y|do(x_2)) - P(y|do(x_1)).$$

The path-specific effect is an extension to the total causal effect in the sense that the effect of the intervention is transmitted only along a subset of causal paths from X to Y [Avin *et al.*, 2005]. Denote a subset of causal paths by π. The π-specific effect considers a counterfactual situation where the effect of X on Y with the intervention is transmitted along π, while the effect of X on Y without the intervention is transmitted along paths not in π. We denote by $P(y \mid do(x_2|_\pi))$ the distribution of Y after an intervention of changing X from x_1 to x_2 with the effect transmitted along π. Then, the π-specific effect of X on Y is described as follows.

Definition 2 (Path-specific effect) *Given a path set π, the π-specific effect measures the effect of the change of X from x_1 to x_2 on $Y = y$ transmitted along π. It is given by*

$$SE_\pi(x_2, x_1) = P(y \mid do(x_2|_\pi)) - P(y \mid do(x_1)).$$

The authors in [Avin *et al.*, 2005] have given the condition under which the path-specific effect can be estimated from the observational data, known as the recanting witness criterion.

Definition 3 (Recanting witness criterion) *Given a path set π, let Z be a node in $\mathcal{G}$ such that: 1) there exists a path*

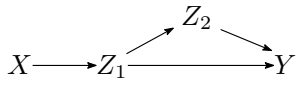

Figure 2: The recanting witness criterion satisfied.

from X to Z which is a segment of a path in π; 2) there exists a path from Z to Y which is a segment of a path in π; 3) there exists another path from Z to Y which is not a segment of any path in π. Then, the recanting witness criterion for the π-specific effect is satisfied with Z as a witness.

Figure 2 shows an example where $\pi = \{X \to Z_1 \to Z_2 \to Y\}$. It is easy to see that the recanting witness criterion is satisfied with Z_1 as the witness.

Theorem 1 (Identifiability) *The π-specific effect can be estimated from the observational data if and only if the recanting witness criterion for the π-specific effect is not satisfied.*

If and only if the recanting witness criterion is not satisfied, the π-specific effect $SE_\pi(x_2, x_1)$ can be computed from the observational data, as shown in Theorem 2 [Shpitser, 2013].

Theorem 2 *When the recanting witness criterion is not satisfied, the π-specific effect $SE_\pi(x_2, x_1)$ can be computed in following steps. First, express $P(y|do(x_1))$ as the truncated factorization formula according to Equation* (3)*. Second, to compute $P(y \mid do(x_2|_\pi))$, divide the children of X into two sets $\mathbf{S}_\pi$ and $\bar{\mathbf{S}}_\pi$, i.e., $Ch(X) = \mathbf{S}_\pi \cup \bar{\mathbf{S}}_\pi$. Let $\mathbf{S}_\pi$ contains X's each child S where arc $X \to S$ is a segment of a path in π; let $\bar{\mathbf{S}}_\pi$ contains X's each child S where either S is not included in any path from C to E, or arc $X \to S$ is a segment of a path not in π. Finally, replace values x_1 with x_2 for the terms corresponding to nodes in $\mathbf{S}_\pi$, and keep values x_1 unchanged for the terms corresponding to nodes in $\bar{\mathbf{S}}_\pi$.*

Note that the above computation requires $\mathbf{S}_\pi \cap \bar{\mathbf{S}}_\pi = \emptyset$. Theorem 1 is reflected here in the sense that: $\mathbf{S}_\pi \cap \bar{\mathbf{S}}_\pi \neq \emptyset$ if and only if the recanting witness criterion for the π-specific effect is satisfied.

3 Modeling Direct and Indirect Discrimination as Path-Specific Effects

Consider a historical dataset $\mathcal{D}$ that contains a group of tuples, each of which describes the profile of an individual. Each tuple is specified by a set of attributes $\mathbf{V}$, including the protected attributes, the decision, and the non-protected attributes. Among the non-protected attributes, assume there is a set of attributes that cannot be objectively justified if used in the decision making process, which we refer to as the *redlining attributes* denoted by $\mathbf{R}$. For ease of presentation, we assume that there is only one protected/decision attribute with binary values. We denote the protected attribute by C associated with two domain values c^- (e.g., female) and c^+ (e.g., male); denote the decision by E associated with two domain values e^- (i.e., negative decision) and e^+ (i.e., positive decision). Our approach can extend to handling multiple domain values of C and even multiple Cs. We assume that a causal network $\mathcal{G}$ is built to correctly represent the causal structure of dataset $\mathcal{D}$. Many algorithms have been proposed to learn the causal network from data [Spirtes *et al.*, 2000; Neapolitan and others, 2004; Colombo and Maathuis, 2014; Kalisch and Bühlmann, 2007]. We also make a reasonable assumption that C has no parent in $\mathcal{G}$, as the protected attribute is always an inherent nature of an individual.

Discrimination is the causal effect of C on E. As discussed, the causal effect of C on E includes direct, indirect discriminatory effects and the explainable effects. To distinguish the different effects, we model them as the causal effects transmitted along different paths. For direct discrimination, we consider the causal effect transmitted along the direct path from C to E, i.e., $C \to E$. Define π_d as the path set that contains only $C \to E$. Then, the above causal effect that is caused by the change of C from c^- to c^+ is given by the π_d-specific effect $SE_{\pi_d}(c^+, c^-)$. For a better understanding, the physical meaning of $SE_{\pi_d}(c^+, c^-)$ can be explained as the expected change in decisions of individuals from protected group c^-, if the decision makers are told that these individuals were from the other group c^+. When applied to the example in Figure 1, it means the expected change in loan approval of applications is actually from the disadvantage group (e.g., black), when the bank is instructed to treat the applicants as from the advantage group (e.g., white). We can see that the π_d-specific effect matches the definition of direct discrimination in law and hence is an appropriate measure for direct discrimination.

Similarly, for indirect discrimination, we consider the causal effect transmitted along all the indirect paths from C to E that contain the redlining attributes. Given the set of redlining attributes $\mathbf{R}$, we define π_i as the path set that contains all the causal paths from C to E which pass through $\mathbf{R}$, i.e., each of the paths includes at least one node in $\mathbf{R}$. Thus, the above causal effect is given by the π_i-specific effect $SE_{\pi_i}(c^+, c^-)$. The physical meaning of $SE_{\pi_i}(c^+, c^-)$ is the expected change in decisions of individuals from protected group c^-, if the values of the redlining attributes in the profiles of these individuals were changed as if they were from the other group c^+. When applied to the example in Figure 1, it means the expected change in loan approval of the disadvantage group if they had the same racial makeups shown in the Zip Code as the advantage group. The π_i-specific effect also matches the definition of indirect discrimination and is appropriate for measuring indirect discrimination.

Therefore, we have the following theorem.

Theorem 3 *The effect of direct discrimination can be captured by the π_d-specific effect $SE_{\pi_d}(c^+, c^-)$, and the effect of indirect discrimination can be captured by the π_i-specific effect $SE_{\pi_i}(c^+, c^-)$.*

Based on the above path-specific effect metrics, we propose the criterion for direct and indirect discrimination. We define that direct discrimination against protected group c^- exists if $SE_{\pi_d}(c^+, c^-) > \tau$, where $\tau > 0$ is a use-defined threshold for discrimination depending on the law. For instance, the 1975 British legislation for sex discrimination sets $\tau = 0.05$, namely a 5% difference. Similarly, given the redlining attributes $\mathbf{R}$, we define that indirect discrimination against protected group c^- exists if $SE_{\pi_i}(c^+, c^-) > \tau$. To avoid reverse discrimination, we do not specify which group is the protected group. As a result, we give the following criterion.

Theorem 4 *Direct discrimination exists if $SE_{\pi_d}(c^+, c^-) > \tau$ or $SE_{\pi_d}(c^-, c^+) > \tau$ holds, and indirect discrimination exists if $SE_{\pi_i}(c^+, c^-) > \tau$ or $SE_{\pi_i}(c^-, c^+) > \tau$ holds.*

The following theorem shows how to compute $SE_{\pi_d}(c^+, c^-)$ and $SE_{\pi_i}(c^+, c^-)$ from the observational data.

Theorem 5 *The π_d-specific effect $SE_{\pi_d}(c^+, c^-)$ is given by*

$$SE_{\pi_d}(c^+, c^-) = \sum_{\mathbf{V}\setminus\{C,E\}} \Big(P(e^+|c^+, Pa(E)\setminus\{C\}) \prod_{V\in\mathbf{V}\setminus\{C,E\}} P(v|Pa(V))\delta_{C=c^-} \Big) - P(e^+|c^-). \tag{4}$$

For the π_i-specific effect $SE_{\pi_i}(c^+, c^-)$, divide C's children into $\mathbf{S}_{\pi_i}$ and $\bar{\mathbf{S}}_{\pi_i}$. If $\mathbf{S}_{\pi_i} \cap \bar{\mathbf{S}}_{\pi_i} = \emptyset$, then $SE_{\pi_i}(c^+, c^-)$ is given by

$$SE_{\pi_i}(c^+, c^-) = \sum_{\mathbf{V}\setminus\{C\}} \Big(\prod_{G\in\mathbf{S}_{\pi_i}} P(g|c^+, Pa(G)\setminus\{C\}) \prod_{H\in\bar{\mathbf{S}}_{\pi_i}} P(h|c^-, Pa(H)\setminus\{C\}) \prod_{O\in\mathbf{V}\setminus(\{C\}\cup Ch(C))} P(o|Pa(O))\delta_{C=c^-} \Big) - P(e^+|c^-). \tag{5}$$

The proof directly follows Theorem 2 and the truncated factorization Equation (2). Theorem 5 shows that $SE_{\pi_d}(c^+, c^-)$ can always be computed from the observational data but $SE_{\pi_i}(c^+, c^-)$ may not. This is because the recanting witness criterion for the π_d-specific effect is guaranteed to be not satisfied, but the recanting witness criterion for the π_i-specific effect might be satisfied. The situation where $SE_{\pi_i}(c^+, c^-)$ cannot be computed is referred to as the unidentifiable situation. How to deal with this situation will be discussed later in the next section.

The following two propositions further show two properties of the path-specific effect metrics.

Proposition 1 *If path set π contains all causal paths from C to E, then we have*

$$SE_{\pi}(c^+, c^-) = TE(c^+, c^-) = P(e^+|c^+) - P(e^+|c^-).$$

The proof can be directly obtained from Definition 2, Definition 1 and Equation (3). $P(e^+|c^+) - P(e^+|c^-)$ is known as the *risk difference* [Romei and Ruggieri, 2014] widely used for discrimination measurement in the anti-discrimination literature. Therefore, the path-specific effect metrics can be considered as a significant extension to the risk difference for explicitly distinguishing the discriminatory effects of direct and indirect discrimination from the total causal effect.

Proposition 2 *For any path sets π_d and π_i, we do not necessarily have $SE_{\pi_d}(c^+, c^-) + SE_{\pi_i}(c^+, c^-) = SE_{\pi_d\cup\pi_i}(c^+, c^-)$.*

The proof can be obtained from Definition 2 and Theorem 2. This implies that there might not be a linear connection between direct and indirect discrimination.

4 Discrimination Discovery and Removal

4.1 Discrimination Discovery

We propose a Path-Specific based Discrimination Discovery (*PSE-DD*) algorithm based on Theorem 4. It first builds the causal network from the historical dataset, and then computes $SE_{\pi_d}(\cdot)$ and $SE_{\pi_i}(\cdot)$ according to Equations (4) and (5). The procedure of the algorithm is shown in Algorithm 1.

Algorithm 1: *PSE-DD*

Input : Dataset $\mathcal{D}$, protected attribute C, decision attribute E, redlining attributes $\mathbf{R}$, user-defined threshold τ.
Output: Direct/indirect discrimination $judge_d$, $judge_i$.

```
1  G = buildCausalNetwork(D);
2  judge_d = judge_i = false;
3  Compute SE_{π_d}(·) according to Equation (4);
4  if SE_{π_d}(c+, c−) > τ || SE_{π_d}(c−, c+) > τ then
5      judge_d = true;
6  Call subroutine [S_{π_i}, S̄_{π_i}] = DivideChildren;
7  if S_{π_i} ∩ S̄_{π_i} ≠ ∅ then
8      judge_i = unknown;
9      return [judge_d, judge_i];
10 Compute SE_{π_i}(·) according to Equation (5);
11 if SE_{π_i}(c+, c−) > τ || SE_{π_i}(c−, c+) > τ then
12     judge_i = true;
13 return [judge_d, judge_i];
```

The complexity of line 6 depends on how to identify $\mathbf{S}_{\pi_i}$ and $\bar{\mathbf{S}}_{\pi_i}$. A straightforward method is to find all paths in π_i, and for C's each child S check whether $C \to S$ is contained in any path in π_i. However, finding all paths between two nodes in a DAG has an exponential complexity. In our algorithm, we examine the existence of a path from S to E passing through $\mathbf{R}$. It can be easily observed that, a node S belongs to $\mathbf{S}_{\pi_i}$ if and only if there exists a path from S to E passing through $\mathbf{R}$ (a path from S to E passing through $\mathbf{R}$ also includes the path where S itself belongs to $\mathbf{R}$). Similarly, S belongs to $\bar{\mathbf{S}}_{\pi_i}$ if and only if there does not exist a path from S to E passing through $\mathbf{R}$. The subroutine of finding $\mathbf{S}_{\pi_i}$ and $\bar{\mathbf{S}}_{\pi_i}$ is presented in Algorithm 2, which checks whether there exists a node $R \in \mathbf{R}$ so that R is S's decedent and E is R's decedent. Since the decedents of all the nodes involved in the algorithm can be obtained by traversing the network starting from C within the time of $O(|\mathbf{A}|)$, the computational complexity of the subroutine is given by $O(|\mathbf{V}|^2 + |\mathbf{A}|)$.

Algorithm 2: subroutine *DivideChildren*

```
1 S_{π_i} = ∅, S̄_{π_i} = ∅;
2 foreach S ∈ Ch(C)\{E} do
3     foreach R ∈ R do
4         if R ∈ De(S) ∪ {S} && E ∈ De(R) then
5             S_{π_i} = S_{π_i} ∪ {S};
6         else
7             S̄_{π_i} = S̄_{π_i} ∪ {S};
8 return [S_{π_i}, S̄_{π_i}];
```

The complexity of *PSE-DD* also depends on the complexities of building the causal network and computing the path-specific effect according to Equation (4) or (5). Many researches have been devoted to improving the performance of network construction [Kalisch and Bühlmann, 2007; Aliferis and others, 2010] and probabilistic inference in causal net-

works [Heckerman and Breese, 1994; 1996]. The complexity analysis can be found in these related literature.

4.2 Discrimination Removal

When direct or indirect discrimination is discovered for a dataset, the discriminatory effects need to be removed before the dataset is released for predictive analysis. A naive approach would be simply deleting the protected attribute from the dataset, which often incur significant utility loss. In addition, this approach can eliminate direct discrimination, but indirect discrimination still presents.

We propose a Path-Specific Effect based Discrimination Removal (*PSE-DR*) algorithm to remove both direct and indirect discrimination. The general idea is to modify the causal network and then use it to generate a new dataset. Specifically, we modify the CPT of E, i.e., $P(e|Pa(E))$, to obtain a new CPT $P'(e|Pa(E))$, so that the direct and indirect discriminatory effects are below the threshold τ. To maximize the utility of the modified dataset, we minimize the Euclidean distance between the joint distribution of the original causal network (denoted by $P(\mathbf{v})$) and the joint distribution of the modified causal network (denoted by $P'(\mathbf{v})$). As a result, we obtain the following quadratic programming problem with $P'(e|Pa(E))$ as the variables.

$$\begin{aligned}
\text{minimize} \quad & \sum_{\mathbf{V}} \left(P'(\mathbf{v}) - P(\mathbf{v})\right)^2 \\
\text{subject to} \quad & SE_{\pi_d}(c^+, c^-) \le \tau, \quad SE_{\pi_d}(c^-, c^+) \le \tau, \\
& SE_{\pi_i}(c^+, c^-) \le \tau, \quad SE_{\pi_i}(c^-, c^+) \le \tau, \\
& \forall Pa(E), \quad P'(e^-|Pa(E)) + P'(e^+|Pa(E)) = 1, \\
& \forall Pa(E), e, \quad Pr'(e|Pa(E)) \ge 0,
\end{aligned}$$

where $P'(\mathbf{v})$ and $P(\mathbf{v})$ are computed according to Equation (1) using $P'(e|Pa(E))$ and $P(e|Pa(E))$ respectively, and $SE_{\pi_d}(\cdot)$ and $SE_{\pi_i}(\cdot)$ are computed according to Equations (4) and (5) respectively using $P'(e|Pa(E))$. The optimal solution is obtained by solving the quadratic programming problem. After that, the joint distribution of the modified causal network is computed using Equation (1), and the new dataset is generated based on the joint distribution. The procedure of *PSE-DR* is shown in Algorithm 3, where lines 1-2 deal with the unidentifiable situation discussed in the next subsection.

Algorithm 3: *PSE-DR*

```
1 if S_{π_i} ∩ S̄_{π_i} ≠ ∅ then
2     Call subroutine NetworkPreprocess;
3 Obtain the modified CPT of E by solving the quadratic
  programming problem;
4 Calculate P*(v) according to Equation (1) using the modified
  CPTs;
5 Generate D* based on P*(v);
6 return D*;
```

The complexity of *PSE-DR* depends on the complexity of solving the quadratic programming problem. It can be easily shown that, the coefficients of the quadratic terms in the objective function form a positive definite matrix. According to [Kozlov *et al.*, 1980], the quadratic programming can be solved in polynomial time. Finally, it is also worth noting that our approach can be easily extended to handle the situation where either direct or indirect discrimination needs to be removed.

Dealing with Unidentifiable Situation As stated in Theorem 1, when the recanting witness criterion is satisfied, the π_i-specific effect cannot be estimated from the observational data. However, the structure of the recanting witness criterion implies potential indirect discrimination as there exist causal paths from C to E passing through the redlining attributes. From a practical perspective, it is meaningful to ensure non-discrimination while preserving reasonable data utility even though the discriminatory effect cannot be accurately measured. In this case, we preprocess the network as follows. Recall that $\mathbf{S}_{\pi_i} \cap \bar{\mathbf{S}}_{\pi_i} \neq \emptyset$ if and only if the recanting witness criterion is satisfied. For each node $S \in \mathbf{S}_{\pi_i} \cap \bar{\mathbf{S}}_{\pi_i}$, we cut off all the causal paths from S to E that pass through $\mathbf{R}$, so that S would not belong to $\mathbf{S}_{\pi_i}$ any more. Then, we must have $\mathbf{S}_{\pi_i} \cap \bar{\mathbf{S}}_{\pi_i} = \emptyset$ after the modification. To cut off the paths, we focus on the arc from E's each parent Q, i.e., $Q \to E$. If these exists a path from S to Q passing through $\mathbf{R}$, then arc $Q \to E$ is removed from the network. The pseudo-code of this procedure is shown below, which can be added before line 1 in Algorithm 3 to deal with this situation.

Algorithm 4: subroutine *NetworkPreprocess*

```
1 foreach S ∈ S_{π_i} ∩ S̄_{π_i} do
2     foreach Q ∈ Pa(E) do
3         foreach R ∈ R do
4             if R ∈ De(S) && Q ∈ De(R) then
5                 Remove arc Q → E from G;
6                 Break;
```

5 Experiments

In this section, we conduct experiments using the Adult dataset [Lichman, 2013]. We compare our algorithms with the local massaging (*LMSG*) and local preferential sampling (*LPS*) algorithms proposed in [Žliobaitė *et al.*, 2011] and disparate impact removal algorithm (*DI*) proposed in [Feldman *et al.*, 2015; Adler *et al.*, 2016]. The causal networks are constructed and presented by utilizing an open-source software TETRAD [Glymour and others, 2004]. We employ the original PC algorithm [Spirtes *et al.*, 2000] and set the significance threshold 0.01 for conditional independence testing in causal network construction. The quadratic programming is solved using CVXOPT [Dahl and others, 2006].

5.1 Discrimination Discovery

The Adult dataset consists of 48,842 tuples with 11 attributes including `age`, `education`, `sex`, `occupation`, `income`, `marital_status` etc. Due to the sparse data issue, we binarize each attribute's domain values into two classes to reduce the domain sizes. For experiment details including the causal network please refer to [Zhang *et al.*, 2016a]. We treat `sex` as the protected attribute, `income` as the decision, and `marital_status` as the redlining attribute. We set the

Table 1: Discrimination in the modified data ($\tau = 0.05$), and comparison of utility with varied τ values.

	Remove Algorithm				τ			
	PSE-DR	*DI*	*LMSG*	*LPS*	0.025	0.05	0.075	0.1
Direct	0.013	0.001	-0.142	-0.142	0.008	0.012	0.019	0.024
Indirect	0.049	0.050	0.288	0.174	0.024	0.049	0.074	0.100
$\chi^2(\times 10^4)$	1.038	4.964	1.924	1.292	1.247	1.038	1.029	0.819

threshold τ as 0.05. By computing the path-specific effects, we obtain that $SE_{\pi_d}(c^+, c^-) = 0.025$ and $SE_{\pi_i}(c^+, c^-) = 0.175$, which indicate no direct discrimination but significant indirect discrimination against females according to our criterion. In [Žliobaitė *et al.*, 2011], it has been shown that each of the attributes `relationship`, `age` and `working_hours` can explain some of the discrimination. However, no conclusion regarding direct/indirect discrimination is drawn.

5.2 Discrimination Removal

We run the removal algorithm *PSE-DR* to remove discrimination from the dataset, and then run the discovery algorithm *PSE-DD* to further examine whether discrimination is truly removed in the modified dataset. For comparison, we include removal algorithms from previous works: *LMSG*, *LPS* and *DI*. The discriminatory effects of the modified dataset are shown in Table 1 (left). As can be seen, our method *PSE-DR* completely removes both direct and indirect discrimination from the data. In addition, *PSE-DR* produces relatively small data utility loss in term of χ^2. For *LMSG* and *LPS*, indirect discrimination is not removed, and direct discrimination seems to be over removed. The *DI* algorithm provides a parameter λ to indicate the amount of discrimination to be removed, where $\lambda = 0$ represents no modification and $\lambda = 1$ represents full discrimination removal. However, λ has no direct connection with the threshold τ. In our experiments, we execute *DI* multiple times with different λs and report the one that is closest to achieve $\tau = 0.05$. Although *DI* indeed removes direct and indirect discrimination, its data utility is far more worse than *PSE-DR*, implying that it removes many information unrelated to discrimination.

We then examine how the data utility in term of χ^2 varies with different thresholds τ for *PSE-DR*. We change the value of τ from 0.025 to 0.1. From Table 1 (right) we can see that less utility loss is incurred when larger τ value is used. This observation is consistent with our analysis since the larger the value of τ, the more relaxed the constraints in *PSE-DR*.

We also examine whether the predictive models built from the dataset modified by *PSE-DR* incur discrimination in decision making. We divide the original dataset into the training and testing datasets, and remove discrimination from the training dataset to obtain the modified training dataset. Then, we build the predictive models from the modified training dataset, and use them to make predictive decisions over the testing data. Four classifiers, logistic regression (*LR*), decision tree (*DT*), random forest (*RF*) and *SVM*, are used for prediction with five-fold cross-validation. Finally, we run *PSE-DR* to examine whether the predictions for the testing data contain discrimination. The prediction accuracy using both original and modified training dataset are reported as well.

Table 2: Discrimination in prediction. ($\tau = 0.05$)

		LR	*DT*	*RF*	*SVM*
Direct		0.045	0.023	0.022	0.023
Indirect		0.047	0.042	0.050	0.041
Accuracy(%)	Original	81.70	81.77	81.81	81.78
	Modified	81.30	80.55	80.56	80.54

The results are shown in Table 2. As can be seen, the predictions of all classifiers do not incur direct or indirect discrimination, with the accuracy slightly decreased.

6 Related Work

In the literature, classification rule-based methods such as *elift* [Pedreshi *et al.*, 2008] and *belift* [Mancuhan and Clifton, 2014] were proposed to represent certain discrimination patterns. [Luong *et al.*, 2011; Zhang *et al.*, 2016c] dealt with the individual discrimination by finding a group of similar individuals. [Žliobaitė *et al.*, 2011] proposed conditional discrimination which considers some part of the discrimination may be explainable by certain attributes. None of these work explicitly identifies direct discrimination, indirect discrimination, and explainable effects. In [Bonchi *et al.*, 2017], 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.

Proposed methods for discrimination removal are either based on data preprocessing [Kamiran and Calders, 2012; Žliobaitė *et al.*, 2011] or algorithm tweaking [Kamiran *et al.*, 2010; Calders and Verwer, 2010; Kamishima *et al.*, 2011]. In a recent work [Feldman *et al.*, 2015], the authors first ensure no direct discrimination by completely removing the protected attribute C from data, and then modify all the non-protected attributes to ensure that C cannot be predicted from the non-protected attributes. As a result, indirect discrimination is removed since the decision E has no connection with C via the non-protected attributes. However, as shown in our experiment results, this approach suffers significant utility loss as it removes all the connections between C and E.

7 Conclusions

In this paper, we studied the problem of discovering both direct/indirect discrimination from historical data, and removing them before performing predictive analysis. We made use of the causal network to capture the causal structure of the data, and modeled direct and indirect discrimination as different path-specific effects. Based on that, we proposed the discovery algorithm *PSE-DD* to discover both direct and indirect discrimination, and the removal algorithm *PSE-DR* to remove them. The experiments using the real dataset show that, our approach can ensure that the modified data dose not contain any type of discrimination while incurring small utility loss. As a result, the predictive models built from the modified data are not subject to discrimination.

References

[Adler *et al.*, 2016] Philip Adler, Casey Falk, Sorelle A Friedler, et al. Auditing black-box models for indirect influence. In *Proceedings of ICDM 2016*, 2016.

[Aliferis and others, 2010] Constantin F Aliferis et al. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11(Jan):171–234, 2010.

[Avin *et al.*, 2005] Chen Avin, Ilya Shpitser, and Judea Pearl. Identifiability of path-specific effects. In *IJCAI'05*, pages 357–363, 2005.

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

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

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

[Dahl and others, 2006] Joachin Dahl et al. Cvxopt: a python package for convex optimization. In *Proc. eur. conf. op. res*, 2006.

[Feldman *et al.*, 2015] Michael Feldman, Sorelle A Friedler, et al. Certifying and removing disparate impact. In *KDD*, pages 259–268. ACM, 2015.

[Glymour and others, 2004] Clark Glymour et al. The TETRAD project. `http://www.phil.cmu.edu/tetrad`, 2004.

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

[Heckerman and Breese, 1994] David Heckerman and John S Breese. A new look at causal independence. In *Proceedings of UAI 1994*, pages 286–292, 1994.

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

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

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

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

[Kamishima *et al.*, 2011] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *ICDMW 2011*, 2011.

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

[Kozlov *et al.*, 1980] Mikhail K Kozlov, Sergei P Tarasov, et al. The polynomial solvability of convex quadratic programming. *USSR Computational Mathematics and Mathematical Physics*, 20(5):223–228, 1980.

[Lichman, 2013] M. Lichman. UCI machine learning repository. `http://archive.ics.uci.edu/ml`, 2013.

[Luong *et al.*, 2011] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. k-nn as an implementation of situation testing for discrimination discovery and prevention. In *KDD*, pages 502–510. ACM, 2011.

[Mancuhan and Clifton, 2014] Koray Mancuhan and Chris Clifton. Combating discrimination using bayesian networks. *AI & L*, 22(2):211–238, 2014.

[Neapolitan and others, 2004] Richard E Neapolitan et al. *Learning bayesian networks*, volume 38. Prentice Hall Upper Saddle River, 2004.

[Pearl, 2009] Judea Pearl. *Causality: models, reasoning and inference*. Cambridge university press, 2009.

[Pedreshi *et al.*, 2008] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *KDD*, pages 560–568. ACM, 2008.

[Romei and Ruggieri, 2014] Andrea Romei and Salvatore Ruggieri. A multidisciplinary survey on discrimination analysis. *Knowl Eng Rev*, 29(05):582–638, 2014.

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

[Shpitser, 2013] Ilya Shpitser. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive science*, 37(6):1011–1035, 2013.

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

[Žliobaitė *et al.*, 2011] Indre Žliobaitė, Faisal Kamiran, and Toon Calders. Handling conditional discrimination. In *ICDM*, pages 992–1001. IEEE, 2011.

[Zhang *et al.*, 2016a] Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. *arXiv preprint arXiv:1611.07509*, 2016.

[Zhang *et al.*, 2016b] Lu Zhang, Yongkai Wu, and Xintao Wu. On discrimination discovery using causal networks. In *Proceedings of SBP-BRiMS 2016*, 2016.

[Zhang *et al.*, 2016c] Lu Zhang, Yongkai Wu, and Xintao Wu. Situation testing-based discrimination discovery: a causal inference approach. In *Proceedings of IJCAI*, 2016.