Causal Modeling-Based Discrimination Discovery and Removal: Criteria, Bounds, and Algorithms

Lu Zhang, Yongkai Wu, and Xintao Wu

Abstract—Anti-discrimination is an increasingly important task in data science. 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 are 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 graph 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 graph. For certain situations where indirect discrimination cannot be exactly measured due to the unidentifiability of some path-specific effects, we develop an upper bound and a lower bound to the effect of indirect discrimination. Based on the theoretical results, we propose effective algorithms for discovering direct and indirect discrimination, as well as algorithms for precisely removing both types of discrimination while retaining good data utility. Experiments using the real dataset show the effectiveness of our approaches.

Index Terms—Discrimination discovery and removal; Direct and indirect discrimination; Causal modeling; Path-specific effect

1 Introduction

Discrimination refers to unjustified distinctions in decisions against individuals based on their membership in a certain group. Laws and regulations (e.g., the Equal Credit Opportunity Act of 1974) 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*. Nowadays 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 contain 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 go into the predictive models and the decisions made are not subject to discrimination.

In the legal field, discrimination is divided 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 to 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 have received increasing attention over the past few years in data science [1], [2], [3], [4], [5]. 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. [4] treat all the difference in decisions as discrimination. [6] realized the explainable effect but failed to correctly measure it. They also failed to distinguish the effects of direct and indirect discrimination. For discrimination removal, a general requirement is to preserve the data utility, i.e., how the distorted data is close to the original one, 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 [5]) can eliminate discrimination but may suffer significant utility loss. To maximize the data utility, it is necessary to first accurately measure the discriminatory effects.

The causal modeling-based discrimination detection has been proposed most recently [7], [8], [9] 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 model. A causal model [10] is a structural equation-based mathematical object that describes the causal mechanisms of a system. Each causal model is associated with a causal graph for friendly causal inference, where causal effects are carried by the *causal paths* that trace arrows pointing from the cause to the effect. Using the causal 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

• *L. Zhang, Y. Wu and X. Wu are with Computer Science and Computer Engineering Dept., University of Arkansas, Fayetteville, AR 72701. E-mail: {lz006,yw009,xintaowu}@uark.edu*

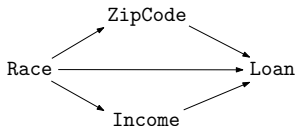


Fig. 1: The toy model.

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.

As shown above, measuring discrimination based on the causal graph requires to measure the causal effect transmitted along certain causal paths. To this end, we employ the technique of the *path-specific effect* [11], [12]. We define direct/indirect discrimination as different path-specific effects, and attempt to compute them using the observational data. In theory, the path-specific effect is not always able to be computed from the observational data. This situation is referred to as the unidentifiability of the path-specific effect. We show that direct discrimination is always identifiable, but indirect discrimination is not identifiable in some cases. For the unidentifiable situation, we provide an upper bound and a lower bound to the effect of indirect discrimination, which is achieved by representing the unidentifiable effect as the expression of counterfactual statements and then scaling up and down specific components of the expression. Based on the theoretical results, we propose effective algorithms that can deal with both identifiable and unidentifiable situations, including algorithms for discovering direct/indirect discrimination, as well as algorithms for precisely removing both types of discrimination while retaining good data utility. The experiments using real datasets show that our approaches are effective in discovering and removing discrimination, ensuring that all types of discrimination are removed while only small utility loss is incurred.

The rest of the paper is organized as follows. Section 2 summarizes the related work. Section 3 proposes the criteria and algorithms for discovering and removing both direct and indirect discrimination based on the path-specific effect. Section 4 deals with the situation where the indirect discrimination cannot be exactly measured from the observational data according to the unidentifiability of the path-specific effect. Section 5 discusses the assumption relaxation and extensions of the proposed methods to several realistic scenarios. The experimental setup and results are discussed in Section 6. Finally, Section 7 concludes the paper.

2 Related Work

How to discover discrimination from data has been widely studied and many techniques have been proposed in the literature. Among them a widely adopted concept is called the statistical parity, which means that the demographics of the set of individuals receiving positive (or negative) decisions are identical to the demographics of the population as a whole. Based on statistical parity, the classic statistical metrics of discrimination consider the difference between the proportion of having positive decision for the non-protected group (p_1), that for the protected group p_2, and that for the whole population (p). According to how the difference is measured, these metrics can be distinguished into $p_1 - p_2$ (a.k.a. risk difference), $\frac{p_1}{p_2}$ (a.k.a. risk ratio), $\frac{1-p_1}{1-p_2}$ (a.k.a. relative chance), $\frac{p_1(1-p_2)}{p_2(1-p_1)}$ (a.k.a. odds ratio), $p_1 - p$ (a.k.a. extended risk difference), $\frac{p_1}{p}$ (a.k.a. extended risk ratio), $\frac{1-p_1}{1-p}$ (a.k.a. extended change), etc. Data mining techniques have also been proposed. Pedreschi et al. proposed to extract from the dataset classification rules which represent certain discrimination patterns [3], [13], [14]. 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, [15] further proposed to use Bayesian networks to compute the confidence of the classification rules for detecting discrimination. The authors in [16] exploited the idea of situation testing to discover individual discrimination. For each member of the protected group with a negative decision outcome, testers with similar characteristics are searched from a historical dataset. When there are significantly different decision outcomes between the testers of the protected group and the testers of the non-protected group, the negative decision can be considered as discrimination. Conditional discrimination, i.e., part of discrimination may be explained by other legally grounded attributes, was studied in [6]. The task was to evaluate to which extent the discrimination apparent for a group is explainable on a legal ground. The metric is still based on the difference of the positive decision proportions for the protected and non-protected groups.

Proposed methods for discrimination removal are either based on data preprocessing or algorithm tweaking. Data preprocessing methods [2], [5], [6], [17] modify the historical data to remove discriminatory patterns. For example, [2], [6] proposed several methods for modifying data, including *Massaging*, which changes the labels of some individuals in the dataset to remove discrimination, *Reweighting*, which assigns weights to individuals to balance the dataset, and *Sampling*, which changes the sample sizes of different subgroups to make the dataset discrimination-free. Another work [5] studied how to remove indirect discrimination from data. The authors 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, which is determined by the non-protected attributes, cannot be used to predict C. In [17], the authors proposed the use of loglinear modeling to capture and measure discrimination, and developed a method for discrimination prevention by modifying significant coefficients of the fitted loglinear model and generate unbiased datasets. On the other hand, algorithm tweaking methods remove discrimination by modifying the model including the decision tree [18], naive Bayes classifier [19], and logistic regression [20]. For example, in [18], the authors developed a strategy for relabeling the leaf nodes of a decision tree to make it discrimination-free. In [21], the authors added the measure of fairness into the classification learning formulation as the constraints so that the classifier learned satisfies the fairness requirement. In [22], 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.

All of the above works are mainly based on correlation

or association. Recently, several studies have been devoted to analyzing discrimination from the causal perspective. In [23], 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. Studies in [7], [8], [9] are built on causal modeling and the associated causal graph, but cannot deal with indirect discrimination. The causal model [10] is a mathematical object that describes the causal mechanisms of a system as a set of structural equations. With well-established conceptual and algorithmic tools, the causal model provides a general, formal, yet friendly calculus of causal effects. In this paper, we adopt the causal model for the quantitative measuring of both direct/indirect discrimination. Specifically, we focus on the technique of path-specific effect [11] that measures the causal effect that is transmitted along certain paths in the causal graph. A recent work [24] proposes similar discrimination criteria that also consider indirect discrimination. However, they are more simplified in order to avoid the complexity in measuring path-specific effects. In addition, [24] suffers inherent limitations: (1) its proposed discrimination criteria can only qualitatively determine the existence of the discrimination, but cannot quantitatively measure the amount of discriminatory effects as we do; (2) its proposed algorithms for avoiding discrimination proposed only work under the linearity assumptions about the underlying causal model while our methods make no assumption.

For the unidentifiability of the path-specific effect, a recent work [25] proposes three principled approaches: (1) obtaining the data on exogenous variables $\mathbf{U}$; (2) considering a identifiable path-specific effect that includes the paths of interest and some other paths; and (3) deriving bounds for unidentifiable path-specific effects, which is claimed to be an open problem in general. In this paper, we deal with this issue by adopting the third approach.

3 Discrimination Discovery and Removal

3.1 Preliminaries

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., $\mathbf{x}$.

A causal model is formally defined as follows [10].

Definition 1 (Causal Model). A causal model is a triple $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F} \rangle$ where

1) $\mathbf{U}$ is a set of arbitrarily distributed *unobserved* random variables (called exogenous) that are determined by factors outside the model. A joint probability distribution $P(\mathbf{u})$ is defined over the variables in $\mathbf{U}$.
2) $\mathbf{V}$ is a set of *observed* random variables (called endogenous) $\{X_1, \cdots, X_i, \cdots\}$ that are determined by variables in the model, namely, variables in $\mathbf{U} \cup \mathbf{V}$.
3) $\mathbf{F}$ is a set of deterministic functions $\{f_1, \cdots, f_i, \cdots\}$ where each f_i is a mapping from $\mathbf{U} \times (\mathbf{V} \backslash X_i)$ to X_i. Symbolically, the set of equations $\mathbf{F}$ can be represented by writing

$$x_i = f_i(\mathbf{pa}_{X_i}, \mathbf{u}_i)$$

where $\mathbf{pa}_{X_i}$ is any realization of the unique minimal set of variables $\mathbf{Pa}_{X_i}$ in $\mathbf{V} \backslash X_i$ that renders f_i nontrivial. Here variables in $\mathbf{Pa}_{X_i}$ are referred to as the parents of X_i. Similarly, $\mathbf{U}_i \subset \mathbf{U}$ stands for the unique minimal set of variables in $\mathbf{U}$ that renders f_i nontrivial.

Each causal model $\mathcal{M}$ is associated with a causal graph $\mathcal{G} = (\mathbf{V}, \mathbf{A})$, where $\mathbf{V}$ is a set of nodes and $\mathbf{A}$ is a set of edges. Each node in $\mathcal{G}$ corresponds to a variable X in $\mathbf{V}$. In this paper, terms node and variable are used interchangeably. Each edge, denoted by an arrow $\rightarrow$, points from each member of $\mathbf{Pa}_X$ toward X to represent the direct causal relationship. Standard terminology in the graph theory is used in the causal graph. For a node X, we also use symbol $\mathbf{Pa}_X$ to denote its parents, and use $\mathbf{Ch}_X$ to denote its children. The path that traces arrows directed from one node X to another node Y is called the *causal path* from X to Y.

It is generally assumed that the causal model is a *Markovian model*, which means that all exogenous variables in $\mathbf{U}$ are mutually independent. An equivalent graphical expression of the Markovian model is the *local Markov condition*, which means that each node is independent of its non-descendants conditional on all its parents. Under this assumption, the joint distribution over all attributes $P(\mathbf{v})$ can be computed using the factorization formula [26]

$$P(\mathbf{v}) = \prod_{V \in \mathbf{V}} P(v \mid \mathbf{pa}_V), \tag{1}$$

where $P(v \mid \mathbf{pa}_V)$ is the conditional probability table (CPT) associated with V.

In the causal model, measuring causal effects is facilitated with the *do*-operator [10], which simulates the physical interventions that force some variables $\mathbf{X}$ to take certain values $\mathbf{x}$. The post-intervention distributions, which represent the effect of the intervention, can be computed from the observational data. 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 variables $\mathbf{Y} = \mathbf{V} \backslash \mathbf{X}$, i.e., $P(\mathbf{Y} = \mathbf{y} \mid do(\mathbf{X} = \mathbf{x}))$ or simply $P(\mathbf{y} \mid do(\mathbf{x}))$, can be expressed as a truncated factorization formula [10]

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

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

$$P(y \mid do(x)) = \sum_{\mathbf{v}'} \prod_{V \in \mathbf{V} \backslash \{X\}} P(v \mid \mathbf{pa}_V) \delta_{X=x}, \tag{3}$$

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

By using the *do*-operator, the total causal effect of X on Y is defined in Definition 2 [10]. 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 2 (Total causal effect). The total causal effect $TE(x_2, x_1)$ 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

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

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 [11]. 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

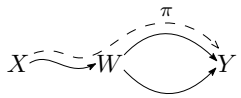


Fig. 2: The "kite pattern".

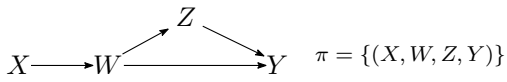


Fig. 3: The recanting witness criterion satisfied.

transmitted along π, while the effect of X on Y without the intervention is transmitted along paths not in π, i.e., $\bar{\pi}$. We denote by $P(y \mid do(x_2|_\pi, x_1|_{\bar{\pi}}))$ 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 3 (Path-specific effect). Given a path set π, the π-specific effect $SE_\pi(x_2, x_1)$ measures the effect of the change of X from x_1 to x_2 on $Y = y$ transmitted along π. It is given by

$$SE_\pi(x_2, x_1) = P\left(y \mid do(x_2|_\pi, x_1|_{\bar{\pi}})\right) - P\left(y \mid do(x_1)\right).$$

The identifiability of path-specific effect $SE_\pi(x_2, x_1)$, i.e., whether it can be computed from the observational data, depends on the identifiability of $P(y \mid do(x_2|_\pi, x_1|_{\bar{\pi}}))$. The authors in [11] have given the necessary and sufficient condition for $P(y \mid do(x_2|_\pi, x_1|_{\bar{\pi}}))$ to be identifiable, known as the recanting witness criterion.

Definition 4 (Recanting witness criterion). Given a path set π pointing from X to Y, let W be a node in $\mathcal{G}$ such that: 1) there exists a path from X to W which is a segment of a path in π; 2) there exists a path from W to Y which is a segment of a path in π; 3) there exists another path from W 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 W as a witness.

The graphical pattern of the recanting witness criterion is known as the "kite pattern", as shown in Figure 2. Figure 3 shows an example where $\pi = \{(X, W, Z, Y)\}$. It is easy to see that the recanting witness criterion is satisfied with W as the witness.

Theorem 1 (Identifiability). For path-specific effect $SE_\pi(x_2, x_1)$, $P(y \mid do(x_2|_\pi, x_1|_{\bar{\pi}}))$ can be computed from the observational data if and only if the recanting witness criterion for the π-specific effect is not satisfied.

If the recanting witness criterion is not satisfied, $P(y \mid do(x_2|_\pi, x_1|_{\bar{\pi}}))$ can be computed as shown in Theorem 2 [12].

Theorem 2. For path-specific effect $SE_\pi(x_2, x_1)$, if the recanting witness criterion is not satisfied, then $P(y \mid do(x_2|_\pi, x_1|_{\bar{\pi}}))$ can be computed in following steps. Firstly, express $P(y \mid do(x_1))$ as the truncated factorization formula according to Eq. (3). Secondly, divide the children of X other than Y into two sets $\mathbf{S}_\pi$ and $\bar{\mathbf{S}}_\pi$, i.e., $\mathbf{Ch}_X \backslash \{Y\} = \mathbf{S}_\pi \cup \bar{\mathbf{S}}_\pi$. Let $\mathbf{S}_\pi$ contain X's each child S where edge $X \to S$ is a segment of a path in π; let $\bar{\mathbf{S}}_\pi$ contain X's each child S where either S is not included in any path from C to E, or edge $X \to S$ 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: $\mathbf{S}_\pi \cap \bar{\mathbf{S}}_\pi = \emptyset$ if and only if the recanting witness criterion for the π-specific effect is not satisfied.

3.2 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). For simplifying representation, we also make two reasonable assumptions: (1) C has no parent in $\mathcal{G}$; (2) E has no child in $\mathcal{G}$. The first one is due to the fact that the protected attribute is usually an inherent nature of an individual, and second one is because that E is usually the output of a decision making system. We will discuss the relaxation of above assumptions in Section 5. We assume that a causal graph $\mathcal{G}$ can be built to correctly represent the causal structure of dataset $\mathcal{D}$. Many algorithms have been proposed to learn the causal graph from data [27], [28], [29], [30].

Discrimination is the causal effect of C on E. As we have 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 edge 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 the disadvantage group (e.g., black), if the bank was instructed to treat these applicants as from the advantage group (e.g., white). We can see that the π_d-specific effect perfectly follows 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. As can be seen, the π_i-specific effect also follows the definition of indirect discrimination and is appropriate for measuring indirect discrimination.

Therefore, we have the following claim.

Claim 1. The effect of direct discrimination is captured by the π_d-specific effect $SE_{\pi_d}(c^+, c^-)$, and the effect of indirect discrimination is 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 identifying 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 3. Given the protected attribute C, decision E, and redlining attributes $\mathbf{R}$, direct discrimination exists if either $SE_{\pi_d}(c^+, c^-) > \tau$ or $SE_{\pi_d}(c^-, c^+) > \tau$ holds, and indirect discrimination exists if either $SE_{\pi_i}(c^+, c^-) > \tau$ 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 by using Theorem 2.

Theorem 4. The π_d-specific effect $SE_{\pi_d}(c^+, c^-)$ is given by

$$SE_{\pi_d}(c^+, c^-) = \sum_{\mathbf{q}} \left(P(e^+|c^+, \mathbf{q})P(\mathbf{q}|c^-)\right) - P(e^+|c^-), \quad (4)$$

where $\mathbf{Q}$ is the parents of E except C, i.e., $\mathbf{Q} = \mathbf{Pa}_E\backslash\{C\}$. For the π_i-specific effect $SE_{\pi_i}(c^+, c^-)$, divide C's children other than E into $\mathbf{S}_{\pi_i}$ and $\bar{\mathbf{S}}_{\pi_i}$ whose definitions are the same as those in Theorem 2. If $\mathbf{S}_{\pi_i} \cap \bar{\mathbf{S}}_{\pi_i} = \emptyset$, then $SE_{\pi_i}(c^+, c^-)$ is given by

$$\begin{aligned} SE_{\pi_i}(c^+, c^-) = \sum_{\mathbf{v}'} \Bigg(& P(e^+|c^-, \mathbf{q}) \prod_{G \in \mathbf{S}_{\pi_i}} P(g|c^+, \mathbf{pa}_G\backslash\{C\}) \\ & \prod_{H \in \bar{\mathbf{S}}_{\pi_i}\backslash\{E\}} P(h|c^-, \mathbf{pa}_H\backslash\{C\}) \prod_{O \in \mathbf{V}\backslash\mathbf{Ch}_C} P(o|\mathbf{pa}_O) \Bigg) - P(e^+|c^-), \end{aligned} \quad (5)$$

where $\mathbf{V}' = \mathbf{V}\backslash\{C, E\}$. It can be simplified to

$$SE_{\pi_i}(c^+, c^-) = \sum_{\mathbf{q}} \left(P(e^+|c^-, \mathbf{q})P(\mathbf{q}|c^+)\right) - P(e^+|c^-), \quad (6)$$

if π_i contains all causal paths from C to E except direct edge $C \to E$.

Proof: According to the definition of $SE_{\pi_d}(c^+, c^-)$, we have

$$SE_{\pi_d}(c^+, c^-) = P(e^+ \mid do(c^+|_{\pi_d}, c^-|_{\bar{\pi}_d})) - P(e^+ \mid do(c^-)).$$

Since C has no parent, it is straitforward that $P(e^+ \mid do(c^-)) = P(e^+|c^-)$. For $P(e^+ \mid do(c^+|_{\pi_d}, c^-|_{\bar{\pi}_d}))$, following Theorem 2, we express $P(e^+|c^-)$ as the truncated factorization formula, given by

$$P(e^+|c^-) = \sum_{\mathbf{v}'} \left(P(e^+|c^-, \mathbf{q}) \prod_{V \in \mathbf{V}'} P(v \mid \mathbf{pa}_V) \right), \quad (7)$$

where $\mathbf{V}' = \mathbf{V}\backslash\{C, E\}$. It can be shown that $\prod_{V \in \mathbf{V}'} P(v \mid \mathbf{pa}_V) = P(\mathbf{v}'|c^-)$. In fact, if we sort all nodes in $\mathbf{V}'$ according to the topological ordering as $\{V_1, \cdots, V_j, \cdots\}$, we can see that all parents of each node V_j are before it in the ordering. In addition, since C has no parent, it must be V_j's non-descendant; since E has no child, it cannot be V_j's parent. Thus, based on the local Markov condition, we have $P(v_j \mid \mathbf{pa}_{V_j}) = P(v_j \mid c^-, v_1, \cdots, v_{j-1})$. According to the chain rule we obtain $P(\mathbf{v}'|c^-)$. Therefore, it follows that

$$P(e^+|c^-) = \sum_{\mathbf{q}} \left(P(e^+|c^-, \mathbf{q})P(\mathbf{q}|c^-)\right).$$

Then, we divide the children of C into $\mathbf{S}_{\pi_d}$ and $\bar{\mathbf{S}}_{\pi_d}$, and replace c^- with c^+ for the terms corresponding to nodes in $\mathbf{S}_{\pi_d}$. Note that $\mathbf{S}_{\pi_d}$ contains only one node E. As a result, we have

$$P(e^+ \mid do(c^+|_{\pi_d}, c^-|_{\bar{\pi}_i})) = \sum_{\mathbf{q}} \left(P(e^+|c^+, \mathbf{q})P(\mathbf{q}|c^-)\right),$$

which leads to Eq. (4).

For the indirect discrimination, by definition we have

$$SE_{\pi_i}(c^+, c^-) = P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i})) - P(e^+|do(c^-)).$$

To compute the first term, we also express $P(e^+|c^-)$ as Eq. (7), and divide the children of C into $\mathbf{S}_{\pi_i}$ and $\bar{\mathbf{S}}_{\pi_i}$. Then, node set $\mathbf{V}'$ can be divided into three disjoint subsets: $\mathbf{S}_{\pi_i}$, $\bar{\mathbf{S}}_{\pi_i}$ and $\mathbf{V}'\backslash\mathbf{Ch}_C$. We replace c^- with c^+ only for the terms corresponding to nodes in $\mathbf{S}_{\pi_i}$. As a result, we obtain Eq. (5).

If π_i contains all causal paths from C to E except $C \to E$, it means that $\mathbf{S}_{\pi_i} = \mathbf{Ch}_C\backslash\{E\}$ and $\bar{\mathbf{S}}_{\pi_i} = \emptyset$. Note that

$$\prod_{G \in \mathbf{Ch}_C\backslash\{E\}} P(g \mid c^+, \mathbf{pa}_G\backslash\{C\}) \prod_{O \in \mathbf{V}'\backslash\mathbf{Ch}_C} P(o \mid \mathbf{pa}_O) = \prod_{V \in \mathbf{V}'} P(v \mid \mathbf{pa}_V),$$

which can be similarly shown to equal to $P(\mathbf{v}'|c^+)$. As a result we obtain Eq. (6). □

Theorem 4 shows that $SE_{\pi_d}(c^+, c^-)$ can always be computed from the observational data but $SE_{\pi_i}(c^+, c^-)$ may not[1]. 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 the unidentifiable 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 3, Definition 2 and Eq. (3). $P(e^+|c^+) - P(e^+|c^-)$ is known as the *risk difference* [4] 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 3 and Theorem 2. In fact, as shown in [31], the above equality holds if all functions in $\mathbf{F}$ of the causal model are linear, and π_i contains all causal paths from C to E other than $C \to E$. Thus, Proposition 2 implies that if the causal relationship is not linear, then a linear connection between direct and indirect discrimination also does not exist.

3.3 Discovery Algorithm

We propose a Path-Specific based Discrimination Discovery (*PSE-DD*) algorithm based on Theorem 3. It first builds the causal graph from the historical dataset, and then computes $SE_{\pi_d}(\cdot)$ and $SE_{\pi_i}(\cdot)$ according to Eq. (4) and (5). The procedure of the algorithm is shown in Algorithm 1.

1. Note that Eq. (6) can still be computed from the observational data since $\bar{\mathbf{S}}_{\pi_i} = \emptyset$ when π_i contains all causal paths from C to E except $C \to E$.

Algorithm 1: *PSE-DD*

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

1 $\mathcal{G} = buildCausalNetwork(\mathcal{D})$;
2 $judge_d = judge_i = false$;
3 Compute $SE_{\pi_d}(\cdot)$ according to Eq. (4);
4 **if** $SE_{\pi_d}(c^+, c^-) > \tau \parallel SE_{\pi_d}(c^-, c^+) > \tau$ **then**
5 $\quad judge_d = true$;
6 Call subroutine $[\mathbf{S}_{\pi_i}, \bar{\mathbf{S}}_{\pi_i}] = DivideChildren(\mathcal{G}, C, E, \mathbf{R})$;
7 **if** $\mathbf{S}_{\pi_i} \cap \bar{\mathbf{S}}_{\pi_i} \neq \emptyset$ **then**
8 $\quad judge_i = unknown$;
9 $\quad$ **return** $[judge_d, judge_i]$;
10 Compute $SE_{\pi_i}(\cdot)$ according to Eq. (5);
11 **if** $SE_{\pi_i}(c^+, c^-) > \tau \parallel SE_{\pi_i}(c^-, c^+) > \tau$ **then**
12 $\quad judge_i = true$;
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 descendant and E is R's descendant. Since the descendants 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*

Input : Causal graph $\mathcal{G}$, protected attribute C, decision attribute E, redlining attributes $\mathbf{R}$.
Output: $\mathbf{S}_{\pi_i}$ and $\bar{\mathbf{S}}_{\pi_i}$.

1 $\mathbf{S}_{\pi_i} = \emptyset$, $\bar{\mathbf{S}}_{\pi_i} = \emptyset$;
2 **foreach** $S \in \mathbf{Ch}_C \backslash \{E\}$ **do**
3 $\quad$ **foreach** $R \in \mathbf{R}$ **do**
4 $\quad\quad$ **if** $R \in \mathbf{De}_S \cup \{S\}$ && $E \in \mathbf{De}_R$ **then**
5 $\quad\quad\quad \mathbf{S}_{\pi_i} = \mathbf{S}_{\pi_i} \cup \{S\}$;
6 $\quad\quad$ **else**
7 $\quad\quad\quad \bar{\mathbf{S}}_{\pi_i} = \bar{\mathbf{S}}_{\pi_i} \cup \{S\}$;
8 **return** $[\mathbf{S}_{\pi_i}, \bar{\mathbf{S}}_{\pi_i}]$;

The computational complexity of *PSE-DD* also depends on the complexities of building the causal graph and computing the path-specific effect according to Eq. (4) or (5). Many researches have been devoted to improving the performance of network construction [30], [32], [33] and probabilistic inference in causal graphs [34], [35]. The complexity analysis can be found in these related literature.

3.4 Removal Algorithm

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 incurs 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 graph and then use it to generate a new dataset. Specifically, we modify the CPT of E, i.e., $P(e \mid \mathbf{pa}_E)$, to obtain a new CPT $P'(e \mid \mathbf{pa}_E)$, 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 graph (denoted by $P(\mathbf{v})$) and the joint distribution of the modified causal graph (denoted by $P'(\mathbf{v})$). As a result, we obtain the following quadratic programming problem with $P'(e \mid \mathbf{pa}_E)$ as the variables.

$$
\begin{aligned}
\text{minimize} \quad & \sum_{\mathbf{v}} \left(P'(\mathbf{v}) - P(\mathbf{v})\right)^2 \\
\text{subject to} \quad & SE_{\pi_d}(c^+, c^-) \leq \tau, \quad SE_{\pi_d}(c^-, c^+) \leq \tau, \\
& SE_{\pi_i}(c^+, c^-) \leq \tau, \quad SE_{\pi_i}(c^-, c^+) \leq \tau, \\
& \forall \mathbf{Pa}_E, \quad P'(e^- \mid \mathbf{pa}_E) + P'(e^+ \mid \mathbf{pa}_E) = 1, \\
& \forall \mathbf{Pa}_E, e, \quad P'(e \mid \mathbf{pa}_E) \geq 0,
\end{aligned}
$$

where $P'(\mathbf{v})$ and $P(\mathbf{v})$ are computed according to Eq. (1) using $P'(e \mid \mathbf{pa}_E)$ and $P(e \mid \mathbf{pa}_E)$ respectively, and $SE_{\pi_d}(\cdot)$ and $SE_{\pi_i}(\cdot)$ are computed according to Eq. (4) and (5) respectively using $P'(e \mid \mathbf{pa}_E)$. The optimal solution is obtained by solving the quadratic programming problem. After that, the joint distribution of the modified causal graph is computed using Eq. (1), and the new dataset is generated based on the joint distribution. The procedure of *PSE-DR* is shown in Algorithm 3

Algorithm 3: *PSE-DR*

Input : Historical dataset $\mathcal{D}$, protected attribute C, decision attribute E, redlining attributes $\mathbf{R}$, threshold τ.
Output: Modified dataset $\mathcal{D}^*$.

1 $[judge_d, judge_i] = PSE\text{-}DD(\mathcal{D}, C, E, \mathbf{R}, \tau)$;
2 **if** $[judge_d, judge_i] == [false, false]$ **then**
3 $\quad$ **return** $\mathcal{D}$;
4 $\mathcal{G} = buildCausalNetwork(\mathcal{D})$;
5 **if** $judge_i == unknown$ **then**
6 $\quad$ Call subroutine *GraphPreprocess*;
7 Obtain the modified CPT of E by solving the quadratic programming problem;
8 Calculate $P^*(\mathbf{v})$ according to Eq. (1) using the modified CPTs;
9 Generate $\mathcal{D}^*$ based on $P^*(\mathbf{v})$;
10 **return** $\mathcal{D}^*$;

As stated in Theorems 1 and 4, when the recanting witness criterion is satisfied, the π_i-specific effect cannot be estimated from the observational data. However, the "kite pattern" implies potential indirect discrimination as there exist causal paths from C to E passing through the redlining attributes. Although the indirect discriminatory effect cannot be accurately measured, from a practical perspective, it is still meaningful to ensure non-discrimination while preserving reasonable data utility. As a straightforward method, we can first modify the causal graph to remove the "kite pattern", and then obtain the modified CPT of E by solving the quadratic programming problem similar to the identifiable situation. To remove the "kite pattern", 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. When cutting off the paths, we focus on the edge from E's each parent Q, i.e., $Q \to E$. If there exists a path from S to Q passing through $\mathbf{R}$, then edge $Q \to E$ is removed from the network. The pseudo-code of this procedure called *GraphPreprocess* is shown below, which is added as a subroutine in line 5 of *PSE-DR*.

Algorithm 4: subroutine *GraphPreprocess*

Input: Causal graph $\mathcal{G}$, protected attribute C, decision attribute E, redlining attributes $\mathbf{R}$.

```
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 edge Q → E from G;
6                 Break;
```

The computational 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 [36], 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.

4 Dealing with Unidentifiable Situation

Under the unidentifiable situation where the recanting witness criterion is satisfied, *PSE-DD* and *PSE-DR* provide workable but crude solutions to the discrimination discovery and removal. In this section, we develop the refined discrimination discovery and removal algorithms by deriving upper and lower bounds for the unidentifiable indirect discrimination. Compared to the presence of the "kite pattern", the bounds can be used as better indicators for discovering indirect discrimination, i.e., the upper bound smaller than τ indicates no indirect discrimination, while the lower bound larger than τ indicates its existence. We also prove that the refined removal algorithm is at least as good as *PSE-DR* in term of preserving the data utility. We start by giving several necessary preliminaries in addition to those presented in Section 3.1.

4.1 Preliminaries

In Section 3.1, we have shown that variables $\mathbf{Y}$ under an intervention $do(\mathbf{x})$ is still a set of random variables, whose distribution $P(\mathbf{y} \mid do(\mathbf{x}))$ is different from the observational distribution of $\mathbf{Y}$. We denote $\mathbf{Y}$ under intervention $do(\mathbf{x})$ by $\mathbf{Y}_\mathbf{x}$, i.e., we define

$$P(\mathbf{y}_\mathbf{x}) \triangleq P(\mathbf{Y}_\mathbf{x} = \mathbf{y}) \triangleq P(\mathbf{y} \mid do(\mathbf{x})).$$

We can interpret $\mathbf{Y}_\mathbf{x}$ as a counterfactual statement, which represents "the value that $\mathbf{Y}$ would have obtained, had $\mathbf{X}$ been $\mathbf{x}$". From the definition of the causal model we can observe that, if all the exogenous variables $\mathbf{U}$ are given, then $\mathbf{Y}_\mathbf{x}$ are no longer random variables but are fixed values. We denote the $\mathbf{Y}_\mathbf{x}$ under the context of $\mathbf{U} = \mathbf{u}$ by $\mathbf{Y}_\mathbf{x}(\mathbf{u})$. In the following we present several properties regarding the counterfactual statement, which are proved to be held in the context of Markovian model [10].

Property 1. For any variable Y, $Y_{\mathbf{Pa}_Y}$ is independent of the counterfactual statements of all Y's non-descendants.

Property 2. For any variable Y, we have

$$P(y_{\mathbf{pa}_Y}) = P(y \mid \mathbf{pa}_Y).$$

Property 3. For any set of endogenous variables $\mathbf{Y}$ and any set of endogenous variables $\mathbf{X}$ disjoint of $\{\mathbf{Y}, \mathbf{Pa}_\mathbf{Y}\}$, we have

$$P(\mathbf{y}_{\mathbf{pa}_\mathbf{Y},\mathbf{x}}) = P(\mathbf{y}_{\mathbf{pa}_\mathbf{Y}}).$$

Property 4. For any three sets of endogenous variables $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$,

$$\mathbf{Z}_\mathbf{x}(\mathbf{u}) = \mathbf{z} \quad \Longrightarrow \quad \mathbf{Y}_\mathbf{x}(\mathbf{u}) = \mathbf{Y}_{\mathbf{x},\mathbf{z}}(\mathbf{u}).$$

Property 1 reflects the local Markov condition. Property 2 renders every parent set $\mathbf{Pa}_Y$ exogenous relative to its child Y. Property 3 reflects the insensitivity of Y to any intervention once its direct causes are held constant. Property 4 states that, if we know the values that $\mathbf{Z}$ would have in certain situation, then the values of any other variables $\mathbf{Y}$ are equivalent to that if we perform an intervention to force $\mathbf{Z}$ to $\mathbf{z}$.

Next, we introduce an essential concept regarding to the unidentifiability of the path-specific effect by using the notion of counterfactual statement. Straightforwardly, by $\mathbf{Y}_\mathbf{x}(\mathbf{u})$ and $P(\mathbf{u})$, we can represent $P(\mathbf{y}_\mathbf{x})$ as

$$P(\mathbf{y}_\mathbf{x}) = \sum_{\mathbf{u}:\mathbf{Y}_\mathbf{x}(\mathbf{u})=\mathbf{y}} P(\mathbf{u}). \tag{8}$$

In the same way, we can define the joint distribution of multiple counterfactual statements (which cannot be defined by using the *do*-operator), i.e., $P(\mathbf{Y}_\mathbf{x} = \mathbf{y}, \mathbf{Y}_{\mathbf{x}'} = \mathbf{y}')$ or $P(\mathbf{y}_\mathbf{x}, \mathbf{y}'_{\mathbf{x}'})$, which represents the probability to "$\mathbf{Y}$ would be $\mathbf{y}$ if $\mathbf{X} = \mathbf{x}$ and $\mathbf{Y}$ would be $\mathbf{y}'$ if $\mathbf{X} = \mathbf{x}'$", given as

$$P(\mathbf{y}_\mathbf{x}, \mathbf{y}'_{\mathbf{x}'}) = \sum_{\{\mathbf{u}:\mathbf{Y}_\mathbf{x}(\mathbf{u})=\mathbf{y},\mathbf{Y}_{\mathbf{x}'}(\mathbf{u})=\mathbf{y}'\}} P(\mathbf{u}).$$

When $\mathbf{x} \neq \mathbf{x}'$, $\mathbf{Y}_\mathbf{x}$ and $\mathbf{Y}_{\mathbf{x}'}$ cannot be measured simultaneously. In fact, it is known that $P(\mathbf{y}_\mathbf{x}, \mathbf{y}'_{\mathbf{x}'})$ is unidentifiable from the observational data even in the Markovian model [37]. We will show that the unidentifiability of the $P(\mathbf{y}_\mathbf{x}, \mathbf{y}'_{\mathbf{x}'})$ is the source of the unidentifiability of the path-specific effect satisfying the recanting witness criterion. However, $P(\mathbf{y}_\mathbf{x}, \mathbf{y}'_{\mathbf{x}'})$ is certainly bounded by the following condition:

$$\sum_{\mathbf{y}'} P(\mathbf{y}_\mathbf{x}, \mathbf{y}'_{\mathbf{x}'}) = P(\mathbf{y}_\mathbf{x}). \tag{9}$$

4.2 Bounding Indirect Discrimination

Recalling the definition of the path-specific effect (Definition 3), in the π_i-specific effect, $P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i}))$ represents the probability of $E = e^+$ after the intervention of changing C from c^- to c^+ with the effect transmitted along π_i. By using the notation of the counterfactual statement, we can similarly denote the value of E after the intervention by E_{c^+}. However, keep in mind that different from the original counterfactual statement, here for E_{c^+} the effect of the intervention on C is transmitted only along π_i.

For any variable Y other than C, E, we can also denote their values that would be obtained after the intervention as counterfactual statement Y_{c^+}. Similar to E, the value of Y_{c^+} depends on whether it belongs to a path in π_i. If Y belongs to any path in π_i, then the value of Y_{c^+} will be affected by the intervention. If Y does not belong to any path in π_i, then the value of Y_{c^+} will not

be affected by the intervention and remain the same as if $C = c^-$. Based on the causal effect transmission, to obtain Y_{c^+}, we need to know the value of Y's each ancestor W affected by the intervention if there exists a path from W to Y that is a segment of a path in π_i; or we need to know the value of W not affected by the intervention if there exists a path from W to Y that is not a segment of any path in π_i. As can be seen, if W has two emanating edges where one belongs to a path in π_i and the other one does not belong to any path in π_i, we need to simultaneously know the value of W affected by the intervention as well as the one not affected by the intervention. To distinguish these two counterfactual situations, we denote the former by W_{c^+} and the latter by W_{c^-}. According to the definition of the recanting witness criterion (Definition 4), it can be easily shown that W is a node where both W_{c^+} and W_{c^-} are needed if and only if W is a witness for the recanting witness criterion. Here we call such node W a *witness variable/node*.

The above analysis shows that, for each witness variable W, we need to consider two sets of realizations, one obtained by W_{c^+} (denoted as w^+), and the other obtained by W_{c^-} (denoted as w^-). For each variable Y that is not a witness variable, we only consider one set of realizations obtained by Y_{c^+}.

In the following, we derive a general expression of $SE_{\pi_i}(c^+, c^-)$ and then develop its upper and lower bounds when subject to the recanting witness criterion. We first provide a property and a proposition that are needed for the derivation.

Similar to Property 4, in the path-specific effect, if we know the two realizations that witness variables $\mathbf{W}$ would have in both counterfactual situations, then the values of any other variable Y are equivalent to that if we perform an intervention to force $\mathbf{W}$ to these realizations. Thus, we obtain the following property that is directly extended from Property 4.

Property 5. For endogenous variables X, Y, W, assume that W is a witness variable, x, x' are two realizations of X, and w, w' are two realizations of W. For any π-specific effect of X we have

$$W_x(\mathbf{u}) = w,\ W_{x'}(\mathbf{u}) = w' \quad \Longrightarrow \quad Y_x(\mathbf{u}) = Y_{x,w^*}(\mathbf{u}),$$

where w^* means that its value is specified by w if there exists a path from W to Y that is a segment of a path in π, and specified by w' otherwise.

Based on Properties 3, 4 and 5, we can prove the following proposition.

Proposition 3. In π_i-specific effect $SE_{\pi_i}(c^+, c^-)$, for any endogenous variable Y, use $\mathbf{pa}_Y^+$ to denote the realization of Y's parents meaning that if $\mathbf{Pa}_Y$ contains any witness node W or C, its value is specified by w^+ or c^+ if edge $W \to Y$ belongs to a path in π_i, and specified by w^- or c^- otherwise; and use $\mathbf{pa}_Y^-$ to denote the realization of Y's parents meaning that if $\mathbf{Pa}_Y$ contains any witness node W or C, its value is specified by w^- or c^-. If Y is not a witness variable, we have

$$P(y_{c^+}, \cdots) = \begin{cases} P(y_{\mathbf{pa}_Y^+}, \cdots) & \text{if } Y \text{ belongs to any path in } \pi_i, \\ P(y_{\mathbf{pa}_Y^-}, \cdots) & \text{otherwise,} \end{cases} \tag{10}$$

and if Y is a witness variable, we have

$$P(y_{c^+}, \cdots) = P(y_{\mathbf{pa}_Y^+}, \cdots) \text{ and } P(y_{c^-}, \cdots) = P(y_{\mathbf{pa}_Y^-}, \cdots), \tag{11}$$

where $\cdots$ represents all other variables.

Please refer to the appendix for the proof.

For ease of representation, we divide all nodes on the causal paths from C to E (except C and E) into three disjoint subsets:

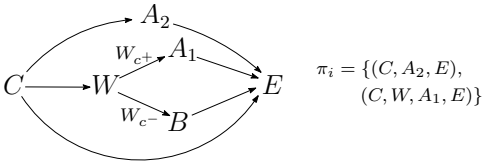


Fig. 4: π_i-specific effect satisfying recanting witness criterion.

the subset of witness nodes (denoted by $\mathbf{W}$), the subset of nodes not in $\mathbf{W}$ that belong to paths in π_i (denoted by $\mathbf{A}$), and the subset of nodes not in $\mathbf{W}$ that do not belong to any path in π_i (denoted by $\mathbf{B}$)[2]. An example is shown in Figure 4 where $\mathbf{W} = \{W\}$, $\mathbf{A} = \{A_1, A_2\}$, and $\mathbf{B} = \{B\}$. The notations on the edges represent the specification of the values of each node's parents.

In Theorem 5 we give the general expression of $SE_{\pi_i}(c^+, c^-)$. Since by definition we have $SE_{\pi_i}(c^+, c^-) = P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i})) - P(e^+|c^-)$ where the second term is trivial, we focus on the general expression of $P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i}))$.

Theorem 5. When subject to the recanting witness criterion, $P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i}))$ is given by

$$P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i})) = \sum_{\mathbf{a},\mathbf{b},\mathbf{w}^+,\mathbf{w}^-} P(e^+|c^-, \mathbf{q}) \prod_{A\in\mathbf{A}} P(a|\mathbf{pa}_A^+) \prod_{B\in\mathbf{B}} P(b|\mathbf{pa}_B^-) \prod_{W\in\mathbf{W}} P(w^+_{\mathbf{pa}_W^+}, w^-_{\mathbf{pa}_W^-}). \tag{12}$$

Proof: For simplicity and without loss of generality, assume that all nodes are along the causal paths from C to E. We can re-write distribution $P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i}))$ as the sum of the joint distribution as follows.

$$\begin{aligned} &P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i})) \triangleq P(E_{c^+} = e^+) \\ &= \sum_{\mathbf{a},\mathbf{b},\mathbf{w}^+,\mathbf{w}^-} P(E_{c^+} = e^+, \mathbf{A}_{c^+} = \mathbf{a}, \mathbf{B}_{c^+} = \mathbf{b}, \mathbf{W}_{c^+} = \mathbf{w}^+, \mathbf{W}_{c^-} = \mathbf{w}^-) \\ &\triangleq \sum_{\mathbf{a},\mathbf{b},\mathbf{w}^+,\mathbf{w}^-} P(e^+_{c^+}, \underbrace{a_{c^+}, \cdots}_{A\in\mathbf{A}}, \underbrace{b_{c^+}, \cdots}_{B\in\mathbf{B}}, \underbrace{w^+_{c^+}, w^-_{c^-}, \cdots}_{W\in\mathbf{W}}). \end{aligned}$$

By using Proposition 3, it follows that

$$P(e^+|do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i})) = \sum_{\mathbf{a},\mathbf{b},\mathbf{w}^+,\mathbf{w}^-} P(e^+_{c^-,\mathbf{q}}, \underbrace{a_{\mathbf{pa}_A^+}, \cdots}_{A\in\mathbf{A}}, \underbrace{b_{\mathbf{pa}_B^-}, \cdots}_{B\in\mathbf{B}}, \underbrace{w^+_{\mathbf{pa}_W^+}, w^-_{\mathbf{pa}_W^-}, \cdots}_{W\in\mathbf{W}}).$$

According to Property 1, the counterfactual statement of each variable is independent of all its non-descendants. Thus, we have

$$\begin{aligned} &P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i})) \\ &= \sum_{\mathbf{a},\mathbf{b},\mathbf{w}^+,\mathbf{w}^-} P(e^+_{c^-,\mathbf{q}}) \prod_{A\in\mathbf{A}} P(a_{\mathbf{pa}_A^+}) \prod_{B\in\mathbf{B}} P(b_{\mathbf{pa}_B^-}) \prod_{W\in\mathbf{W}} P(w^+_{\mathbf{pa}_W^+}, w^-_{\mathbf{pa}_W^-}). \end{aligned}$$

According to Property 2, it follows that

$$\begin{aligned} &P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i})) \\ &= \sum_{\mathbf{a},\mathbf{b},\mathbf{w}^+,\mathbf{w}^-} P(e^+|c^-, \mathbf{q}) \prod_{A\in\mathbf{A}} P(a|\mathbf{pa}_A^+) \prod_{B\in\mathbf{B}} P(b|\mathbf{pa}_B^-) \prod_{W\in\mathbf{W}} P(w^+_{\mathbf{pa}_W^+}, w^-_{\mathbf{pa}_W^-}). \end{aligned} \tag{13}$$

Hence the theorem is proven. □

We can see that Eq. (13) contains the joint distribution of counterfactual statements $P(w^+_{\mathbf{pa}_W^+}, w^-_{\mathbf{pa}_W^-})$ which is unidentifiable

2. Redlining attributes can be contained in $\mathbf{W}$ and $\mathbf{A}$ but cannot be contained in $\mathbf{B}$.

from the observational data, making $P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i}))$ and hence the π_i-specific effect $SE_{\pi_i}(c^+, c^-)$ unidentifiable.

Next, we show how to bound $P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i}))$ by scaling up and down certain terms in Eq. (13) and then eliminating $P(w^+_{\mathbf{pa}^+_W}, w^-_{\mathbf{pa}^-_W})$ using Eq. (9). For ease of representation, we further divide $\mathbf{A}$ into two disjoint subsets: (1) the set of nodes that are involved in the "kite pattern", i.e., it is contained in a path in π_i that also contains any node in $\mathbf{W}$, denoted by $\mathbf{A}_1$; (2) the complementary set, i.e., those not involved in the "kite pattern", denoted by $\mathbf{A}_2$. Then, we give the upper and lower bounds of $P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i}))$ as shown in Theorem 6.

Theorem 6. The upper bound of $P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i}))$ is given by

$$\sum_{\mathbf{a}_2,\mathbf{b},\mathbf{w}^-} \max_{\mathbf{a}_1,\mathbf{w}^+}\{P(e^+|c^-,\mathbf{q})\} \prod_{A\in\mathbf{A}_2} P(a|\mathbf{pa}^+_A) \prod_{B\in\mathbf{B}} P(b|\mathbf{pa}^-_B) \prod_{W\in\mathbf{W}} P(w^-|\mathbf{pa}^-_W), \tag{14}$$

and the lower bound of $P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i}))$ is given by

$$\sum_{\mathbf{a}_2,\mathbf{b},\mathbf{w}^-} \min_{\mathbf{a}_1,\mathbf{w}^+}\{P(e^+|c^-,\mathbf{q})\} \prod_{A\in\mathbf{A}_2} P(a|\mathbf{pa}^+_A) \prod_{B\in\mathbf{B}} P(b|\mathbf{pa}^-_B) \prod_{W\in\mathbf{W}} P(w^-|\mathbf{pa}^-_W). \tag{15}$$

Proof: It is straightforward that

$$P(e^+|c^-,\mathbf{q}) \le \max_{\mathbf{a}_1,\mathbf{w}^+}\{P(e^+|c^-,\mathbf{q})\}.$$

Thus, from Eq. (12) we have

$$P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i})) \le \sum_{\mathbf{a}_1,\mathbf{a}_2,\mathbf{b},\mathbf{w}^+,\mathbf{w}^-} \max_{\mathbf{a}_1,\mathbf{w}^+}\{P(e^+|c^-,\mathbf{q})\} \prod_{A\in\mathbf{A}_1} P(a|\mathbf{pa}^+_A) \prod_{A\in\mathbf{A}_2} P(a|\mathbf{pa}^+_A) \prod_{B\in\mathbf{B}} P(b|\mathbf{pa}^-_B) \prod_{W\in\mathbf{W}} P(w^+_{\mathbf{pa}^+_W}, w^-_{\mathbf{pa}^-_W}).$$

We can identify three properties for any node $A \in \mathbf{A}_1$: (1) A cannot be the parent of any node A' in $\mathbf{A}_2$. If not so, we have a path that contains C, A, A', E and any node $W \in \mathbf{W}$. This path must belong to π_i, otherwise A is contained in both a path in π_i and a path not in π_1, making A a witness node. Thus, A' is also involved in the "kite pattern". (2) A cannot be the parent of any node in $\mathbf{B}$. Otherwise, A belongs to a path in π_i and also a path not in π_i, making A a witness node. (3) A cannot be the parent of any node in $\mathbf{W}$, otherwise A also becomes a witness node. Based on the three properties, the RHS of above inequality equals to

$$\begin{aligned}
&\sum_{\mathbf{a}_2,\mathbf{b},\mathbf{w}^+,\mathbf{w}^-} \max_{\mathbf{a}_1,\mathbf{w}^+}\{P(e^+|c^-,\mathbf{q})\} \prod_{A\in\mathbf{A}_2} P(a|\mathbf{pa}^+_A) \prod_{B\in\mathbf{B}} P(b|\mathbf{pa}^-_B) \\
&\quad \prod_{W\in\mathbf{W}} P(w^+_{\mathbf{pa}^+_W}, w^-_{\mathbf{pa}^-_W}) \sum_{\mathbf{a}_1} \prod_{A\in\mathbf{A}_1} P(a|\mathbf{pa}^+_A) \\
&= \sum_{\mathbf{a}_2,\mathbf{b},\mathbf{w}^+,\mathbf{w}^-} \max_{\mathbf{a}_1,\mathbf{w}^+}\{P(e^+|c^-,\mathbf{q})\} \prod_{A\in\mathbf{A}_2} P(a|\mathbf{pa}^+_A) \prod_{B\in\mathbf{B}} P(b|\mathbf{pa}^-_B) \\
&\quad \prod_{W\in\mathbf{W}} P(w^+_{\mathbf{pa}^+_W}, w^-_{\mathbf{pa}^-_W}).
\end{aligned}$$

Then, we can similarly identify two properties for any node $W \in \mathbf{W}$ and its realization w^+: (1) w^+ cannot be involved in $\mathbf{pa}^+_A$ for any $A \in \mathbf{A}_2$, otherwise there exists a path in π_i that contains W, A, making A be involved in the "kite pattern"; (2) w^+ cannot be involved in $\mathbf{pa}^-_B$ for any $B \in \mathbf{B}$, which is by the definition of $\mathbf{B}$. Thus, the above expression further becomes

$$\begin{aligned}
&\sum_{\mathbf{a}_2,\mathbf{b},\mathbf{w}^-} \max_{\mathbf{a}_1,\mathbf{w}^+}\{P(e^+|c^-,\mathbf{q})\} \prod_{A\in\mathbf{A}_2} P(a|\mathbf{pa}^+_A) \prod_{B\in\mathbf{B}} P(b|\mathbf{pa}^-_B) \\
&\quad \sum_{\mathbf{w}^+} \prod_{W\in\mathbf{W}} P(w^+_{\mathbf{pa}^+_W}, w^-_{\mathbf{pa}^-_W}) \\
&= \sum_{\mathbf{a}_2,\mathbf{b},\mathbf{w}^-} \max_{\mathbf{a}_1,\mathbf{w}^+}\{P(e^+|c^-,\mathbf{q})\} \prod_{A\in\mathbf{A}_2} P(a|\mathbf{pa}^+_A) \prod_{B\in\mathbf{B}} P(b|\mathbf{pa}^-_B) \prod_{W\in\mathbf{W}} P(w^-|\mathbf{pa}^-_W).
\end{aligned}$$

By using $P(e^+|c^-,\mathbf{q}) \ge \min_{\mathbf{a}_1,\mathbf{w}^+}\{P(e^+|c^-,\mathbf{q})\}$, similarly we can prove the lower bound. □

From Theorem 6 we can directly obtain the upper bound $ub(SE_{\pi_i}(c^+, c^-))$ and lower bound $lb(SE_{\pi_i}(c^+, c^-))$ of $SE_{\pi_i}(c^+, c^-)$.

4.3 Algorithms for Unidentifiable Situation

Based on the derived bounds of the indirect discrimination, we can refine the proposed discovery algorithm *PSE-DD* to better deal with the unidentifiable situation, as shown in *PSE-DD** (Algorithm 5). On the other hand, we can also refine the proposed removal algorithm *PSE-DR* by replacing $SE_{\pi_i}(c^+, c^-)$ and $SE_{\pi_i}(c^-, c^+)$ in the constraints of the quadratic programming with $ub(SE_{\pi_i}(c^+, c^-))$ and $ub(SE_{\pi_i}(c^-, c^+))$. We refer to this new quadratic programming as the adjusted quadratic programming problem. The refined removal Algorithm *PSE-DR** is shown in Algorithm 6.

Algorithm 5: *PSE-DD**

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

$\mathcal{G}$ = *buildCausalNetwork*($\mathcal{D}$);
$judge_d = judge_i = false$;
Compute $SE_{\pi_d}(\cdot)$ according to Eq. (4);
if $SE_{\pi_d}(c^+, c^-) > \tau \,\|\, SE_{\pi_d}(c^-, c^+) > \tau$ **then**
  $judge_d = true$;
Call subroutine $[\mathbf{S}_{\pi_i}, \bar{\mathbf{S}}_{\pi_i}]$ = *DivideChildren*($\mathcal{G}, C, E, \mathbf{R}$);
if $\mathbf{S}_{\pi_i} \cap \bar{\mathbf{S}}_{\pi_i} \neq \emptyset$ **then**
  Compute $ub(SE_{\pi_i}(c^+, c^-))$, $lb(SE_{\pi_i}(c^+, c^-))$, $lb(SE_{\pi_i}(c^-, c^+))$, $ub(SE_{\pi_i}(c^-, c^+))$ according to Eq. (14), (15);
  if $ub(SE_{\pi_i}(c^+, c^-)) \le \tau$ & $ub(SE_{\pi_i}(c^-, c^+)) \le \tau$ **then**
    $judge_i = false$;
  else if $lb(SE_{\pi_i}(c^+, c^-)) > \tau \,\|\, lb(SE_{\pi_i}(c^-, c^+)) > \tau$ **then**
    $judge_i = true$;
  else
    $judge_i = unknown$;
  return $[judge_d, judge_i]$;
Compute $SE_{\pi_i}(\cdot)$ according to Eq. (5);
if $SE_{\pi_i}(c^+, c^-) > \tau \,\|\, SE_{\pi_i}(c^-, c^+) > \tau$ **then**
  $judge_i = true$;
return $[judge_d, judge_i]$;

Algorithm 6: *PSE-DR**

Input : Historical dataset $\mathcal{D}$, protected attribute C, decision attribute E, redlining attributes $\mathbf{R}$, threshold τ.
Output: Modified dataset $\mathcal{D}^*$.

$[judge_d, judge_i]$ = *PSE-DD**($\mathcal{D}, C, E, \mathbf{R}, \tau$);
if $[judge_d, judge_i] == [false, false]$ **then**
  return $\mathcal{D}$;
$\mathcal{G}$ = *buildCausalNetwork*($\mathcal{D}$);
if $judge_i == unkonwn$ **then**
  Obtain the modified CPT of E by solving the adjusted quadratic programming problem;
else
  Obtain the modified CPT of E by solving the original quadratic programming problem;
Calculate $P^*(\mathbf{v})$ using the modified CPTs and generate $\mathcal{D}^*$;
return $\mathcal{D}^*$;

The following proposition shows that, the adjusted quadratic programming will at least produce an equivalently good solution as the quadratic programming after performing subroutine

GraphPreprocess. This implies that *PSE-DR** performs at least as good as *PSE-DR* in term of the data utility preserving. Our experiments in Section 6 show that *PSE-DR** outperforms *PSE-DR* in the practical situations.

Proposition 4. The modified CPT of E obtained from the quadratic programming after performing *GraphPreprocess* is a feasible solution of the adjusted quadratic programming problem.

Proof: Firstly consider algorithm *PSE-DR*. Denote by $\mathcal{G}'$ the causal graph obtained after the *GraphPreprocess* subroutine, denote by $\mathbf{Q}^*$ ($\mathbf{Q}^* \subseteq \mathbf{Q}$) the parents of E in $\mathcal{G}'$, and denote by $P^*(e|c,\mathbf{q}^*)$ the modified CPT of E obtained by solving the quadratic programming problem. Note that in $\mathcal{G}'$, based on the local Markov condition, $P^*(e|c,\mathbf{q}^*) = P^*(e|c,\mathbf{q})$ for all $\mathbf{q}$ that $\mathbf{q}^* \subseteq \mathbf{q}$. According to the constraints in the quadratic programming, the indirect discrimination based on the modified CPT of E is bounded by τ.

Now consider the original causal graph $\mathcal{G}$ with E's CPT $P^*(e|c,\mathbf{q}) = P^*(e|c,\mathbf{q}^*)$ for all $\mathbf{q}$ that $\mathbf{q}^* \subseteq \mathbf{q}$. We can see that causal graph $\mathcal{G}$ is actually equivalent to causal graph $\mathcal{G}'$, hence the indirect discrimination measured should also be the same[3]. In the following, we show that the indirect discrimination measured in $\mathcal{G}$ based on $P^*(e|c,\mathbf{q})$ equals to its upper bound given in Theorem 6, which means that $P^*(e|c,\mathbf{q})$ satisfies the constraints of the adjusted quadratic programming, and hence is a feasible solution of the adjusted quadratic programming problem.

As shown in Theorem 5, the first term in Eq. (12) is given by

$$\sum_{\mathbf{a},\mathbf{b},\mathbf{w}^+,\mathbf{w}^-}\left(P^*(e^+|c^-,\mathbf{q})\prod_{A\in\mathbf{A}}P(a|\mathbf{pa}_A^+)\prod_{B\in\mathbf{B}}P(b|\mathbf{pa}_B^-)\prod_{W\in\mathbf{W}}P(w^+_{\mathbf{pa}_W^+},w^-_{\mathbf{pa}_W^-})\right)$$
$$=\sum_{\mathbf{a},\mathbf{b},\mathbf{w}^+,\mathbf{w}^-}\left(P^*(e^+|c^-,\mathbf{q}^*)\prod_{A\in\mathbf{A}}P(a|\mathbf{pa}_A^+)\prod_{B\in\mathbf{B}}P(b|\mathbf{pa}_B^-)\prod_{W\in\mathbf{W}}P(w^+_{\mathbf{pa}_W^+},w^-_{\mathbf{pa}_W^-})\right).$$

Similar to Theorem 6, set $\mathbf{A}$ can be divided into two subsets $\mathbf{A}_1$ and $\mathbf{A}_2$. In addition to the properties shown in the proof of Theorem 6, we further identify two properties that appear after executing *GraphPreprocess*: (1) any node $A \in \mathbf{A}_1$ cannot belong to $\mathbf{Q}^*$, otherwise the "kite pattern" still exists, contradicting to that *GraphPreprocess* removes the "kite pattern"; (2) for similar reason w^+ of any $W \in \mathbf{W}$ cannot be involved in $\mathbf{q}^*$. Thus, the above expression becomes

$$\sum_{\mathbf{a}_2,\mathbf{b},\mathbf{w}^-}P^*(e^+|c^-,\mathbf{q}^*)\prod_{A\in\mathbf{A}_2}P(a|\mathbf{pa}_A^+)\prod_{B\in\mathbf{B}}P(b|\mathbf{pa}_B^-)\prod_{W\in\mathbf{W}}P(w^-|\mathbf{pa}_W^-). \tag{16}$$

Now back to the upper bound. Consider the first term of Eq. (14), which is given by

$$\sum_{\mathbf{a}_2,\mathbf{b},\mathbf{w}^-}\max_{\mathbf{a}_1,\mathbf{w}^+}\{P(e^+|c^-,\mathbf{q}^*)\}\prod_{A\in\mathbf{A}_2}P(a|\mathbf{pa}_A^+)\prod_{B\in\mathbf{B}}P(b|\mathbf{pa}_B^-)\prod_{W\in\mathbf{W}}P(w^-|\mathbf{pa}_W^-). \tag{17}$$

As stated, $\mathbf{a}_1$ and $\mathbf{w}^+$ cannot be involved in $\mathbf{q}^*$. Thus, the maximization operation on $P(e^+|c^-,\mathbf{q}^*)$ has no effect, making Eq. (16) and (17) equivalent. Hence, the the proposition is proved. □

5 Extensions to Realistic Scenarios

5.1 Dealing with Multiple Protected Attributes and Domain Values

For simplicity, in this paper we assume a single protected attribute with binary values. However, in realistic scenarios we may encounter multiple domain values or even multiple protected attributes. For example, in a university admission system, the protected attributes may include the applicant's gender, race and age, and each protected attribute can have multiple values, such as white/black/asian for race. In this case, one may need to ensure that there is no discrimination against each of the protected attribute in term of any domain value. In this subsection, we show how our approach can easily extend to multiple protected attributes and domain values.

Suppose that we have a protected attribute C with n domain values $c^1,\cdots,c^n$, where each value can be specified to denote the protected group. Without loss of generality, we assume c^1 is the protected group, and our objective is to discover whether there is discrimination against c^1 in terms of all other groups, and then remove all the biases that are discovered. For discovery, we can compute $SE_{\pi_d}(c^j,c^1)$ and $SE_{\pi_i}(c^j,c^1)$ for each non-protected group j. If any one is larger than τ, then it indicates discrimination. For removal, the challenge here is that the modification in term of one non-protected group may change the discriminatory effect in term of another non-protected group. This means that, suppose that we have removed the discrimination based on $SE_{\pi_d}(c^j,c^1)$ and $SE_{\pi_i}(c^j,c^1)$ for group c^j, this discrimination may reappear if we continue to remove the discrimination based on $SE_{\pi_d}(c^{j'},c^1)$ and $SE_{\pi_i}(c^{j'},c^1)$ for another group $c^{j'}$. The solution here is including the discrimination constraints into the quadratic programming problem in terms of all non-protected groups for removing all biases at once. The quadratic programming problem is guaranteed to be solvable, since there exists a trivial solution such that letting $P^*(e|c,\mathbf{q}) = P(e)$.

When there are multiple protected attributes $C_1,\cdots,C_m$ (assume that the protected groups are $c_1^1,\cdots,c_m^1$ respectively), two similar methods can be applied. First, we can consider discrimination $SE_{\pi_d}(c_k^{j_k},c_k^1)$ and $SE_{\pi_i}(c_k^{j_k},c_k^1)$ for each protected attributed C_k and its non-protected group $c_k^{j_k}$ as different constraints, i.e., we require that $\forall k,j_k$, $SE_{\pi_d}(c_k^{j_k},c_k^1) \le \tau$ and $SE_{\pi_i}(c_k^{j_k},c_k^1) \le \tau$. On the other hand, we can consider the combination of all protected attributes, i.e., we require that $\forall j_1,\cdots,j_m$, $SE_{\pi_d}(c_1^{j_1}\cdots c_m^{j_m},c_1^1\cdots c_m^1) \le \tau$ and $SE_{\pi_i}(c_1^{j_1}\cdots c_m^{j_m},c_1^1\cdots c_m^1) \le \tau$. We leave the comparison of the two methods to the future work.

5.2 Dealing with Numerical Decision

In some scenarios instead of the categorical decision, we may encounter numerical decisions. For example, in the loan application, the decision can be the amount of loan granted to the applicant. In this case, although we can discretize the numerical decision and turn it into a categorical attribute with multiple domain values, in fact our framework can be naturally extended to deal with numerical decisions directly. If we change the definition of the total causal effect (Definition 2) from the difference of probabilities to the different of expectations, i.e., $TE(x_2,x_1) = \mathbb{E}[Y|do(x_2)] - \mathbb{E}[Y|do(x_1)]$, then we can measure the total causal effect of X on Y even if Y is numerical. Similarly, we can change the definition of the path-specific effect (Definition 3) to $SE_\pi(x_2,x_1) = \mathbb{E}[Y \mid do(x_2|_\pi, x_1|_{\bar{\pi}})] - \mathbb{E}[Y \mid do(x_1)]$ to handle the numerical Y.

The challenge here is how to represent the conditional probability of the numerical decision in the causal graph. A possible way is to employ the Conditional Linear Gaussian (CLG) distribution which is used in the Bayesian network to deal with the mixture of discrete and numerical variables [38]. We denote the conditional

3. In fact, it can be easily shown that the indirect discrimination measured in $\mathcal{G}'$ based on Eq. (5) is equivalent to the indirect discrimination measured in $\mathcal{G}$ based on Eq. (13).

distribution of decision E given its parents $C, \mathbf{Q}$, i.e., $P(e|c, \mathbf{q})$, by a Gaussian distribution $\mathcal{N}(\mu_{c,\mathbf{q}}, \sigma^2_{c,\mathbf{q}})$ where $\mu_{c,\mathbf{q}}$ and $\sigma^2_{c,\mathbf{q}}$ are the mean and variance depending on E's parents. Then, we can calculate $TE(x_2, x_1)$ and $SE_\pi(x_2, x_1)$ based on the truncated factorization formula and Theorem 2.

5.3 Relaxing Assumptions of Protected Attribute and Decision

In Section 3.2 we have made two assumptions: C has no parent and E has no child. Based on the two assumptions, the causal graph is simplified as there cannot be any confounder between C and E, so that we can obtain the concise formulas for computing the discriminatory effects as shown in Theorem 4. In the general situation where confounders exist, the computation of the causal effect including the total effect and the path-specific effect is facilitated by the well-known graphical test called the *back-door criterion* [10]. It has been proved that if a set of nodes $\mathbf{S}$ satisfies the back-door criterion relative to X, Y, then the causal effect of X on Y can be computed under the adjustment for $\mathbf{S}$. For example, let $\mathbf{S}$ satisfy the back-door criterion relative to C, E. Then, the π_d-specific effect $SE_{\pi_d}(c^+, c^-)$ is given by

$$SE_{\pi_d}(c^+, c^-) = \sum_{\mathbf{q},\mathbf{s}} P(e^+|c^+, \mathbf{q}, \mathbf{s})P(\mathbf{q}|c^-, \mathbf{s})P(\mathbf{s}) - \sum_{\mathbf{s}} P(e^+|c^-, \mathbf{s})P(\mathbf{s}).$$

When $\mathbf{S} = \emptyset$, the above equation becomes Eq. (4) in Theorem 4. Similarly, we can obtain the adjusted formulas for the π_i-specific effect as well as its upper and lower bounds. In the Markovian model, it is guaranteed that we can find a set of nodes satisfying the back-door criterion relative to any two nodes.

6 Experiments

In this section, we conduct experiments using two real datasets: the Adult dataset [39] and the Dutch census of 2001 [40]. We evaluate our discovery and removal algorithms under both identifiable and unidentifiable situations. For comparison, we involve the local massaging (*LMSG*) and local preferential sampling (*LPS*) algorithms proposed in [6] and disparate impact removal algorithm (*DI*) proposed in [5], [41]. The causal graphs are constructed and presented by utilizing an open-source software TETRAD [42]. We employ the original PC algorithm [27] and set the significance threshold 0.01 for conditional independence testing in causal graph construction. The quadratic programming is solved using CVXOPT [43]. All experiments were conducted with a PC workstation with 16GB RAM and Intel Core i7-4770 CPU. By default, the discrimination threshold τ is set as 0.05.

6.1 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 and the convention in collecting features by social-platforms [44], we binarize each attribute's domain values into two classes to reduce the domain sizes. We use three tiers in the partial order for temporal priority: `sex`, `age`, `native_country`, `race` are defined in the first tier, `edu_level` and `marital_status` are defined in the second tier, and all other attributes are defined in the third tier. The constructed causal graph is shown in Figure 5a. We treat `sex` as the protected attribute, `income` as the decision, and `marital_status` as the redlining attribute. Then set π_d contains the edge pointing from `sex` to `income`, and set π_i contains all the causal paths from `sex` to `income` that pass through `marital_status`. As can be seen, the π_i-specific effect does not satisfy the recanting witness criterion. By computing the path-specific effects, we obtain that $SE_{\pi_d}(c^+, c^-) = 0.025$ and $SE_{\pi_i}(c^+, c^-) = 0.175$. By setting $\tau = 0.05$, the results indicate no direct discrimination but significant indirect discrimination against females according to our criterion. In [6], 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.

The Dutch census consists of 60,421 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 5b. We treat `sex` as the protected attribute, `occupation` as the decision, and `marital_status` as the redlining attribute. In this case, the recanting witness criterion is also not satisfied. For this dataset, we obtain $SE_{\pi_d}(c^+, c^-) = 0.220$ and $SE_{\pi_i}(c^+, c^-) = 0.001$, indicating significant direct discrimination but no indirect discrimination against females.

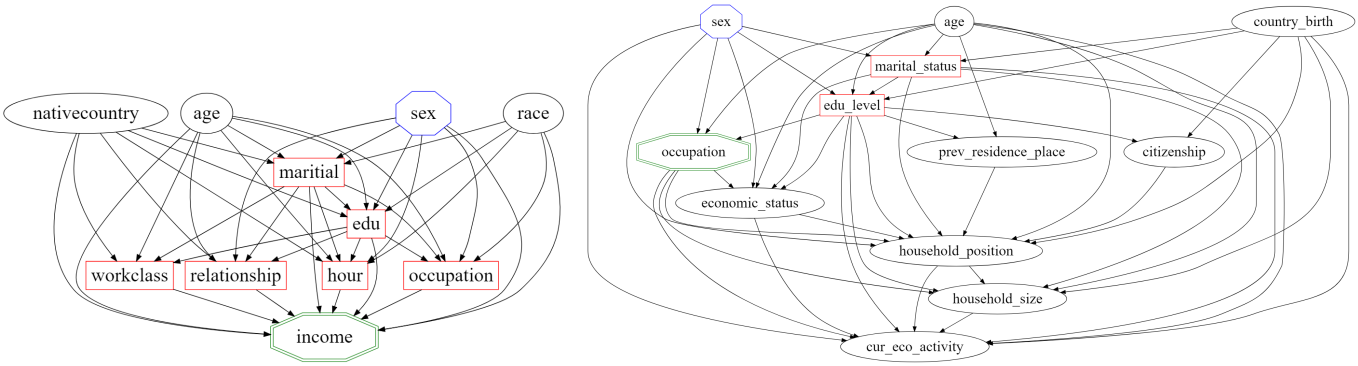
TABLE 1: Discrimination in the modified data ($\tau = 0.05$), and comparison of utility with varied τ values for Adult dataset.

	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

6.2 Discrimination Removal

We run the removal algorithm *PSE-DR* to remove discrimination from both datasets, 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) for the Adult dataset, and in Table 2 (left) for the Dutch census. As can be seen, our method *PSE-DR* completely removes direct and indirect discrimination from both datasets. In addition, *PSE-DR* produces relatively small data utility loss in term of χ^2. For *LMSG* and *LPS*, indirect discrimination is not removed from the Adult dataset, and in both datasets 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 Tables 1 and 2 (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*.



(a) Adult (b) Dutch census

Fig. 5: Constructed causal graphs: the blue octagon node represents the protected attribute, the green double-octagon node represents the decision, and the red rectangle nodes represent represent the (potential) redlining attributes.

TABLE 2: Discrimination in the modified data ($\tau = 0.05$), and comparison of utility with varied τ values for Dutch census.

	Remove Algorithm				τ			
	PSE-DR	*DI*	*LMSG*	*LPS*	0.025	0.05	0.075	0.1
Direct	0.049	0.000	-0.081	-0.100	0.022	0.049	0.073	0.099
Indirect	0.001	-0.001	0.001	0.001	0.001	0.001	0.001	0.001
$\chi^2(\times 10^4)$	1.104	4.604	4.084	1.742	1.279	1.104	1.099	0.934

TABLE 3: Discrimination in prediction for Adult dataset.

		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

TABLE 4: Discrimination in prediction for Dutch census.

		LR	*DT*	*RF*	*SVM*
Direct		0.059	0.103	0.098	0.099
Indirect		0.001	0.001	0.001	0.001
Accuracy(%)	Original	83.45	82.46	83.12	83.70
	Modified	81.93	81.36	81.57	82.10

TABLE 5: Discrimination measured and bounded under unidentifiable situation for Adult dataset.

		edu	occupation	hours	workclass	relationship
Direct				0.025		
Indirect	*lb*	-0.114	-0.069	-0.027	-0.014	-0.086
	ub	0.361	0.039	0.072	0.016	0.015

We also examine whether the predictive models built from the data 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-DD* 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. The results are shown in Tables 3 and 4. As can be seen, for the Adult dataset, the predictions of all classifiers do not incur direct or indirect discrimination, with the accuracy only slightly decreased. However, for the Dutch census, the predictions contain direct discrimination, which is smaller than that in the original data yet significant. Some recent works imply that, even if discrimination is removed from the training data, it can still appear in the predictions of classifiers [45], [46]. How to ensure non-discrimination in the prediction is a future direction of our work.

6.3 Unidentifiable Situation

In this subsection, we examine the proposed methods for handling the unidentifiable situation when measuring and removing the indirect discrimination. We consider each of attribute other than `marital_status` that is on the causal paths from the protected attribute to the decision as the redlining attribute and see whether the recanting witness criterion is satisfied, i.e., π_i forms the "kite pattern". For the Adult dataset, these attributes include `edu_level`, `occupation`, `hours_per_week`, `workclass` and `relationship`, each of which creates the "kite pattern" if it is treated as the redlining attribute. For the Dutch census, only `edu_level` is on the causal paths from the protected attribute to the decision, and treating it as the redlining attribute will not create the "kite pattern". Thus, the remaining of this subsection focus on the Adult dataset.

Upon selecting the redlining attribute, we execute algorithm *PSE-DD** to compute the π_d-specific effect $SE_{\pi_d}(c^+, c^-)$ as well as the upper and lower bounds of the π_i-specific effect $ub(SE_{\pi_i}(c^+, c^-))$ and $lb(SE_{\pi_i}(c^+, c^-))$. The results are shown in Table 5. As can be seen, for all attributes the π_d-specific effect is the same. This is reasonable since treating different attribute as the redlining attribute should not affect the direct discrimination. On the other hand, the upper and lower bounds imply that we can ensure no indirect discrimination if either `occupation`, `workclass` or `relationship` onsidered as the redlining attribute, and we are uncertain about indirect discrimination if either treating `edu_level` or `hours_per_week` as the redlining attribute.

We use `edu_level` as an example to show the results of discrimination removal. The subgraph shown in Figure 6 presents the "kite pattern" formed when treating `edu_leve` as the redlining attribute. The π_i-specific effect satisfies the recanting witness criterion with `marital_status` as the witness. We evaluate the two removal algorithms: *PSD-DR* and *PSD-DR**. For *PSD-DR*,

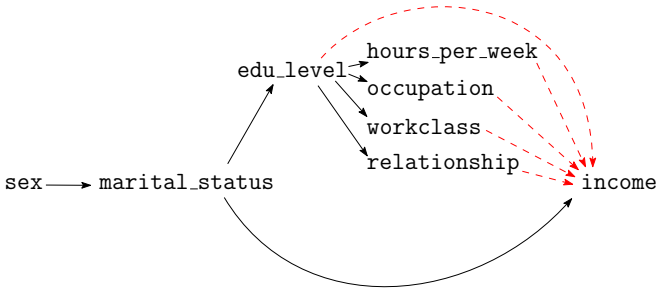


Fig. 6: The "kite pattern" when treating `edu_leve` as redlining. Red dashed edges are to be deleted by *GraphPreprocess*.

TABLE 6: Discrimination in the modified data when treating `edu_level` as redlining.

	PSE-DR	*PSE-DR**
Direct	0.038	0.033
Indirect (*ub*)	0	0.050
$\chi^2(\times 10^4)$	1.499	1.106

subroutine *GraphPreprocess* needs to cut off all causal paths passing through the redlining attribute in order to remove the "kite pattern", which means that it should delete all the edges highlighted by the red dashed edges. The discrimination in the modified data is shown in Table 6. As can be seen, both algorithms guarantee no direct discrimination as well as no indirect discrimination based on its upper bound. However, the utility of the modified data produced by *PSE-DR** is better than that produced *PSE-DR*, which is consistent with our theoretical result. A more straightforward explanation for this example can be that, since all the causal paths in π_i are involved in the "kite pattern", *GraphPreprocess* must cut off all these paths, resulting a total elimination of all indirect discriminatory effect. However, *PSE-DR** can utilize the threshold $\tau = 0.05$, achieving a better balance between non-discrimination and utility preserving.

7 Conclusions and Future Work

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 graph 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. For the situation where indirect discrimination cannot be exactly measured due to the unidentifiability of the path-specific effects, we derived the upper and lower bounds for the unidentifiable indirect discrimination, and developed the refined discovery algorithm *PSE-DD** and removal algorithm *PSE-DR**. 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. Under the unidentifiable situation, the refined algorithms *PSE-DR** produced smaller utility loss than *PSE-DR* that directly deletes edges to remove the unidentifiability.

In the future work, we will extend our work from acquiring discrimination-free dataset to constructing discrimination-free predictive models. Recent works [45], [46] show that, even if the discrimination in the training data is completely removed, the discrimination in prediction can still exist due to the bias in the classifier. Several notions of fair classifiers have been proposed, such as equal opportunity/equal odds [45], and error bias [46], in terms of the balance in the miss-classification rates between protected and non-protected groups. We will study how our discrimination removing technique can be combined with these notions to achieve non-discrimination in the prediction.

Appendix

Proof of Proposition 3

Proof: To prove Eq. (10), denote Y's parents by $\mathbf{Z}$, i.e., $\mathbf{X} = \mathbf{Pa}_Y$. Assume that $\mathbf{X}$ contains no witness node or C. Then $P(y_{c^+}, \cdots)$ can be written as $P(y_{c^+}, \mathbf{x}_{c^+}, \cdots)$. According to Eq. (8), we have

$$P(y_{c^+}, \mathbf{x}_{c^+}, \cdots) = \sum_{\{\mathbf{u}: Y_{c^+}(\mathbf{u})=y, \mathbf{X}_{c^+}(\mathbf{u})=\mathbf{x}, \cdots\}} P(\mathbf{u}).$$

Based on Property 4, we have

$$\mathbf{X}_{c^+}(\mathbf{u}) = \mathbf{x} \quad \Longrightarrow \quad Y_{c^+}(\mathbf{u}) = Y_{c^+,\mathbf{x}}(\mathbf{u}).$$

Since $\mathbf{X} = \mathbf{Pa}_Y$, according to Property 3 we have

$$Y_{c^+,\mathbf{x}}(\mathbf{u}) = Y_{\mathbf{x}}(\mathbf{u}).$$

Therefore, it follows that

$$P(y_{c^+}, \mathbf{x}_{c^+}, \cdots) = \sum_{\{\mathbf{u}: Y_{\mathbf{x}}(\mathbf{u})=y, \mathbf{X}_{c^+}(\mathbf{u})=\mathbf{x}, \cdots\}} P(\mathbf{u}) = P(y_{\mathbf{x}}, \mathbf{x}_{c^+}, \cdots),$$

which can be re-written as $P(y_{\mathbf{pa}_Y^+}, \cdots)$ according to the definition of $\mathbf{pa}_Y^+$.

Assume that $\mathbf{X}$ contains any witness node W or C. Then by applying Property 5, we can similarly obtain $P(y_{c^+}, \cdots) = P(y_{\mathbf{x}^*}, \cdots)$, where $\mathbf{x}^*$ means that if any witness node W or C connects Y with a segment of a path in π_i then its value is specified by w^+ or c^+, and specified by w^- or c^- otherwise. According to the definition of $\mathbf{pa}_Y^+$ and $\mathbf{pa}_Y^-$, $P(y_{\mathbf{x}^*}, \cdots)$ can be re-written as $P(y_{\mathbf{pa}_Y^+}, \cdots)$ if Y belongs to any path in π_i, and $P(y_{\mathbf{pa}_Y^-}, \cdots)$ otherwise.

If Y is a witness node, then the first case and second case of Eq. (11) can be proved similarly to the first case and second case of Eq. (10) respectively. □

Acknowledgments

This paper is a significant extension of the 7-page IJCAI'17 paper [47]. This work was supported in part by NSF 1646654.

References

[1] S. Hajian and J. Domingo-Ferrer, "A methodology for direct and indirect discrimination prevention in data mining," *JKDE*, vol. 25, no. 7, pp. 1445–1459, 2013.

[2] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *KAIS*, vol. 33, no. 1, pp. 1–33, 2012.

[3] S. Ruggieri, D. Pedreschi, and F. Turini, "Data mining for discrimination discovery," *TKDD*, vol. 4, no. 2, p. 9, 2010.

[4] A. Romei and S. Ruggieri, "A multidisciplinary survey on discrimination analysis," *The Knowledge Engineering Review*, vol. 29, no. 05, pp. 582–638, 2014.

[5] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *KDD*. ACM, 2015, pp. 259–268.

[6] I. Žliobaitė, F. Kamiran, and T. Calders, "Handling conditional discrimination," in *ICDM 2011*, 2011, pp. 992–1001.

[7] L. Zhang, Y. Wu, and X. Wu, "Situation testing-based discrimination discovery: a causal inference approach," in *Proceedings of IJCAI 2016*, 2016.

[8] ——, "On discrimination discovery using causal networks," in *Proceedings of SBP-BRiMS 2016*, 2016.

[9] ——, "Achieving non-discrimination in data release," in *Proceedings of SIGKDD 2017*, 2017.

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

[11] C. Avin, I. Shpitser, and J. Pearl, "Identifiability of path-specific effects," in *IJCAI'05*, 2005, pp. 357–363.

[12] I. Shpitser, "Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding," *Cognitive science*, vol. 37, no. 6, pp. 1011–1035, 2013.

[13] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *KDD*. ACM, 2008, pp. 560–568.

[14] D. Pedreschi, S. Ruggieri, and F. Turini, "Measuring discrimination in socially-sensitive decision records," in *SIAM SDM*. Society for Industrial and Applied Mathematics, 2009, p. 581.

[15] K. Mancuhan and C. Clifton, "Combating discrimination using bayesian networks," *Artificial Intelligence and Law*, vol. 22, no. 2, pp. 211–238, 2014.

[16] B. T. Luong, S. Ruggieri, and F. Turini, "k-nn as an implementation of situation testing for discrimination discovery and prevention," in *KDD*. ACM, 2011, pp. 502–510.

[17] Y. Wu and X. Wu, "Using loglinear model for discrimination discovery and prevention," in *Proceedings of DSAA*. IEEE, 2016, pp. 110–119.

[18] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination aware decision tree learning," in *ICDM*. IEEE, 2010, pp. 869–874.

[19] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, 2010.

[20] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware learning through regularization approach," in *ICDMW*. IEEE, 2011, pp. 643–650.

[21] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Artificial Intelligence and Statistics*, 2017, pp. 962–970.

[22] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, 2012, pp. 214–226.

[23] F. Bonchi, S. Hajian, B. Mishra, and D. Ramazzotti, "Exposing the probabilistic causal structure of discrimination," *International Journal of Data Science and Analytics*, vol. 3, no. 1, pp. 1–21, 2017.

[24] N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," in *Advances in Neural Information Processing Systems*, 2017, pp. 656–666.

[25] R. Nabi and I. Shpitser, "Fair inference on outcomes," in *Proceedings of AAAI*, 2018.

[26] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[27] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT press, 2000, vol. 81.

[28] R. E. Neapolitan *et al.*, *Learning bayesian networks*. Prentice Hall Upper Saddle River, 2004, vol. 38.

[29] D. Colombo and M. H. Maathuis, "Order-independent constraint-based causal structure learning," *JMLR*, vol. 15, no. 1, pp. 3741–3782, 2014.

[30] M. Kalisch and P. Bühlmann, "Estimating high-dimensional directed acyclic graphs with the pc-algorithm," *The Journal of Machine Learning Research*, vol. 8, pp. 613–636, 2007.

[31] J. Pearl, "Direct and indirect effects," in *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2001, pp. 411–420.

[32] I. Tsamardinos, C. F. Aliferis, A. Statnikov, and L. E. Brown, "Scaling-up bayesian network learning to thousands of variables using local learning techniques," *Vanderbilt University DSL TR-03-02*, 2003.

[33] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos, "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*, vol. 11, no. Jan, pp. 171–234, 2010.

[34] D. Heckerman and J. S. Breese, "A new look at causal independence," in *Proceedings of UAI*. Morgan Kaufmann Publishers Inc., 1994, pp. 286–292.

[35] ——, "Causal independence for probability assessment and inference using bayesian networks," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 26, no. 6, pp. 826–831, 1996.

[36] M. K. Kozlov, S. P. Tarasov, and L. G. Khachiyan, "The polynomial solvability of convex quadratic programming," *USSR Computational Mathematics and Mathematical Physics*, vol. 20, no. 5, pp. 223–228, 1980.

[37] J. Tian and J. Pearl, "Probabilities of causation: Bounds and identification," in *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2000, pp. 589–598.

[38] S. L. Lauritzen and F. Jensen, "Stable local computation with conditional gaussian distributions," *Statistics and Computing*, vol. 11, no. 2, pp. 191–203, 2001.

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

[40] S. Netherlands, "Volkstelling," https://sites.google.com/site/faisalkamiran/, 2001.

[41] P. Adler, C. Falk, S. A. Friedler, G. Rybeck, C. Scheidegger, B. Smith, and S. Venkatasubramanian, "Auditing black-box models for indirect influence," in *Proceedings of ICDM 2016*, 2016.

[42] C. Glymour *et al.*, "The TETRAD project," http://www.phil.cmu.edu/tetrad, 2004.

[43] J. Dahl and L. Vandenberghe, "Cvxopt: a python package for convex optimization," in *Proc. eur. conf. op. res*, 2006.

[44] T. Speicher, M. Ali, G. Venkatadri, F. N. Ribeiro, G. Arvanitakis, F. Benevenuto, K. P. Gummadi, P. Loiseau, and A. Mislove, "Potential for discrimination in online targeted advertising," in *Conference on Fairness, Accountability and Transparency*, 2018, pp. 5–19.

[45] M. Hardt, E. Price, N. Srebro *et al.*, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, 2016, pp. 3315–3323.

[46] L. Zhang, Y. Wu, and X. Wu, "Achieving non-discrimination in prediction," in *Proceedings of IJCAI 2018*, 2018.

[47] ——, "A causal framework for discovering and removing direct and indirect discrimination," in *Proceedings of IJCAI 2017*, 2017.

Lu Zhang is an Assistant Professor in the Computer Science and Computer Engineering Department, University of Arkansas. He received the BEng degree from the University of Science and Technology of China, in 2008, and the PhD degree in computer science from Nanyang Technological University, Singapore in 2013. His research interests include fairness-aware data mining, causal modeling and inference, and distributed computing.

Yongkai Wu is working toward the PhD degree in computer science at University of Arkansas. He received the BE degree in Electronic Engineering from Tsinghua University in 2014. His main research interests include data mining and machine learning.

Xintao Wu is a Professor in the Department of Computer Science and Computer Engineering at the University of Arkansas. He held a faculty position at the College of Computing and Informatics at the University of North Carolina at Charlotte from 2001 to 2014. He got his Ph.D. degree in Information Technology from George Mason University in 2001. He received his BS degree in Information Science from the University of Science and Technology of China in 1994, an ME degree in Computer Engineering from the Chinese Academy of Space Technology in 1997. His major research interests include data mining and knowledge discovery, bioinformatics, data privacy and security.