Path-specific Causal Fair Prediction via Auxiliary Graph Structure Learning

Liuyi Yao
Alibaba Group
yly287738@alibaba-inc.com

Yaliang Li
Alibaba Group
yaliang.li@alibaba-inc.com

Bolin Ding
Alibaba Group
bolin.ding@alibaba-inc.com

Jingren Zhou
Alibaba Group
jingren.zhou@alibaba-inc.com

Jinduo Liu
Beijing University of Technology
liujinduo0607@emails.bjut.edu.cn

Mengdi Huai
Iowa State University
mdhuai@iastate.edu

Jing Gao
Purdue University
jinggao@purdue.edu

ABSTRACT

With ubiquitous adoption of machine learning algorithms in web technologies, such as recommendation system and social network, algorithm fairness has become a trending topic, and it has a great impact on social welfare. Among different fairness definitions, path-specific causal fairness is a widely adopted one with great potentials, as it distinguishes the fair and unfair effects that the sensitive attributes exert on algorithm predictions. Existing methods based on path-specific causal fairness either require graph structure as the prior knowledge or have high complexity in the calculation of path-specific effect. To tackle these challenges, we propose a novel casual graph based fair prediction framework which integrates graph structure learning into fair prediction to ensure that unfair pathways are excluded in the causal graph. Furthermore, we generalize the proposed framework to the scenarios where sensitive attributes can be non-root nodes and affected by other variables, which is commonly observed in real-world applications, such as recommendation system, but hardly addressed by existing works. We provide theoretical analysis on the generalization bound for the proposed fair prediction method, and conduct a series of experiments on real-world datasets to demonstrate that the proposed framework can provide better prediction performance and algorithm fairness trade-off.

CCS CONCEPTS

• **Information systems** → **Data mining**.

KEYWORDS

Fairness, Graph Structure Learning, Causality

WWW '23, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3583280

ACM Reference Format:
Liuyi Yao, Yaliang Li, Bolin Ding, Jingren Zhou, Jinduo Liu, Mengdi Huai, and Jing Gao. 2023. Path-specific Causal Fair Prediction via Auxiliary Graph Structure Learning. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3543507.3583280

1 INTRODUCTION

Nowadays, more and more people use web technologies, such as recommendation system and social network, to seek information and make decision. Such trend makes algorithm fairness critical, since machine learning algorithms are widely adopted in these web technologies and ensuring the fairness has a great impact on both the social welfare and the platform interests [2–4, 7, 10, 11, 36, 37, 45, 47]. Algorithm fairness aims to reduce or even eliminate *unjustified distinctions of individuals based on their sensitive attributes (e.g., gender and race) during the prediction* [41]. Unfortunately, machine learning models constructed from the raw data are vulnerable to the unfairness risk due to the historical prejudices in the data. It is crucial for model designers to take algorithm fairness into consideration for long-term social welfare.

In recent years, researchers have developed a variety of causal fairness definitions to help machine learning models make fair predictions [16, 17, 21, 26, 28, 31, 32, 35, 39, 40, 42, 43], and one of them, path-specific causal fairness [6, 26, 33], is adopted in this paper. Under the definition of path-specific causal fairness, unfairness is viewed as the presence of *unfair causal effect through the disallowed causal pathway* that the sensitive attributes exert on predictions. In other words, a fair prediction satisfies path-specific causal fairness if it eliminates the causal effect that the sensitive attributes assert on the prediction through disallowed causal pathways. Such a definition provides the flexibility of tracing the unfairness, because in some scenarios, the sensitive attributes affect the decision along multiple pathways, and not all pathways are unfair. For example, in the online marketing, shown in Figure 1, gender and race (denoted as sensitive attribute R) are only allowed to affect the promotion through the preference, since it is reasonable to decide whether sending the promotion according to the preference. Under this fairness rule, paths $R \rightarrow Y$ and $R \rightarrow Z \rightarrow Y$ are unfair paths and path $R \rightarrow Q \rightarrow Y$ is a fair path.

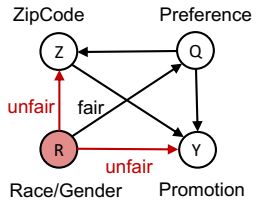
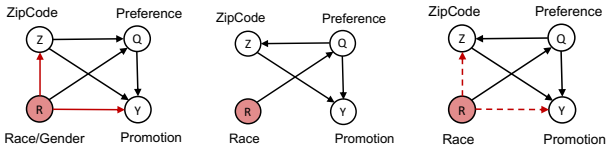


Figure 1: Example of Online Marketing.

To fulfill path-specific causal fairness, some existing works directly calculate the path-specific causal effect (PSE) [1, 27] along the unfair pathways, and minimize the effect simultaneously when maximizing prediction accuracy [26, 33]. Some other works correct the variables located on the unfair pathway by a latent inference-projection method [6]. However, these existing works still face the following challenges: (1) Most of them require a pre-defined graph as the prior knowledge to calculate PSE. (2) The calculation of the path-specific effect is complex, requiring the sequential ignorability assumption [18] to ensure the identification. (3) They all assume the sensitive attributes are root nodes in the causal graph. Namely, there are no other variables that affect the sensitive attributes. Few of them consider the case when the sensitive attributes are non-root nodes, which can be widely observed in real-world applications. For example, in the recommendation system, the item popularity is a sensitive attribute [10, 47], while this variable is a non-root node as it is affected by the item's characteristics.

In light of the above challenges, we propose a **C**ausal **G**raph based **F**airness Framework (**CGF**). To tackle the challenge of lacking the causal graph information, CGF integrates the causal graph structure learning and fair prediction, revealing the causal relationships among the observed variables. To simplify the PSE calculation, CGF imposes the fairness regularization at the graph level by restricting the existence of unfair edges in the learned causal structure. In this way, fair decisions are made based on the corrected observations reconstructed from the learned graph structure. Furthermore, the proposed CGF framework can straightforwardly generalize to the case where sensitive attributes are non-root nodes by introducing the latent variables to divide the fair and unfair effect flow. To the best of our knowledge, the proposed framework CGF is the first work to consider such non-root node case.

Generally speaking, the key of CGF framework is that the causal graph, which the model relies on to make predictions, reveals the fair causal paths of the original observation and eliminates the edges that are unfair by fairness regularization. To be specific, the proposed CGF framework contains three components including graph structure learning, fairness regularization and label prediction. The graph structure learning generates the causal graph that reveals the causal relations between observed variables. We use weighted adjacency to represent the causal graph and each element in the adjacency matrix indicates the effect strength through the edge. The adjacency matrix is learned by minimizing the difference between the observed data (i.e., the data recorded in the dataset) and the data reconstructed based on the adjacency matrix. The second component, fairness regularization, further constraints the adjacency matrix by reducing the weights of the unfair edges, which controls the effect flow through the unfair edges in the causal graph. In other words, the fairness regularization guides the graph



(a) Observed Graph. **(b) Fair Graph.** **(c) Model Graph.**

Figure 2: Causal Graphs of Online Marketing Example.

structure learning part by eliminating the unfair edges in the causal graph. With the above two components, the reconstructed data based on the learned causal graph is a correction of the original observed data with unfair effect eliminated. Then on top of the reconstructed data, the third component, label prediction, provides the fair predictions.

To validate the effectiveness of CGF framework, we conduct a series of experiments on real-world datasets. On the real-world dataset, we compare the proposed framework with several baselines, and experimental results show that CGF can provide a better utility and fairness trade-off. Further, we conduct experiments on a real-world recommendation dataset to evaluate the performance of CGF framework under the case of sensitive attributes as non-root nodes. The experimental results demonstrate that CGF framework can make comparable accurate recommendations while reducing the negative effect caused by sensitive attributes (i.e., item popularity), compared to existing recommendation methods.

2 BACKGROUND

Causal Graph. A causal graph is a directed acyclic graph (DAG) reflecting the causal relationships between variables. Let $\mathcal{G}$ denote a causal graph, and $\mathcal{G} = \langle V, E \rangle$, where V is the set of nodes representing all the variables, and E is the set of edges with each edge $V_i \rightarrow V_j$ describing the causal relation between variable V_i and V_j. The *parents nodes* of node V_i, denoted as $\Pi(V_i)$, and $V_j \in \Pi(V_i)$ if $V_j \rightarrow V_i$. A node is a root node if it has no parent nodes. A *path*, also named as causal pathway, is defined as a sequence of unique nodes with edges between each consecutive node. The *depth* of a node in the graph is the number of arrows in the longest path to the root nodes. In the rest of the paper, we use the term "node", "variable", and "attribute" interchangeably.

Path-specific Causal Fairness. Path-specific causal fairness ensures that the sensitive attributes are not allowed to affect the prediction along the unfair causal pathway. Path-specific causal fairness distinguishes the causal pathways that start from sensitive variables to predicted variables into fair paths and unfair paths, and the goal of fair prediction is to reduce the unfair paths. Path-specific causal fairness is closely related to other definitions of fairness. It is equivalent to remove the direct and indirect discrimination [43]. When all paths starting from the sensitive variables are unfair, achieving path-specific causal fairness is equal to demographic parity (i.e., removing disparate impact) [38].

Definition 2.1. *(Observed Graph). Observed graph is the causal graph of the observed data.*

Definition 2.2. *(Fair Graph). The causal graph satisfying the fairness criterion, and meanwhile, preserving the remaining structure of the observation graph, is the fair graph.*

Definition 2.3. *(Model Graph). Model graph is the causal graph that the decision model relies on.*

Figure 2 shows the observational graph, fair graph and model graph of the example of online marketing in Section 1. Figure 2a is the observational graph, which is the causal graph of the observed data. In the graph, the fair path $R \to Q \to Y$ represents that it is acceptable, in terms of preference, that some people with certain race/gender are not offered with promotion. While, the paths $R \to Y$ and $R \to Z \to Y$ are unfair, indicating that it is disallowed that the race/gender affects the promotion offering directly or indirectly through ZipCode. Figure 2b is the fair graph, which describes the ideal causal relations. Compared with the observational graph, it eliminates the unfair paths. By removing the unfair paths, the fair graph reflects that the difference of promotion offering results across different race/gender groups is explained by the different preference levels among those groups. The rightmost sub-figure is the model graph, which is the graph that the model relies on to predict. As shown in Figure 2c, the model takes R, Z, and Q as input, therefore, they all have directed arrows pointing to prediction Y.

From the above triple-graph perspective, under path-specific causal fairness, the model graph should be consistent with the fair graph, but it is not. Therefore, our objective is to exclude the unfair path (the red dashed arrow in Figure 2c) for the decision, and retain the remaining causal pathways.

Structure Causal Model (SCM). In SCM, each node in $\mathcal{G}$ is associated with a causal mechanism representing the generation of the current node by its parent nodes. It defined as: $\mathcal{F} = \{f_i : V_i = f_i(\Pi(V_i)) + \epsilon_i\}$, where $V_i \in V$ is the i-th node in the graph, $\Pi(V_i)$ is the set of parent nodes of V_i, and ϵ_i is the random noise.

Graph Structure Learning. The problem of graph structure learning is to infer a directed acyclic graph (DAG) from data that reflects the causal relationships among variables. In general, this can be summarized as solving the following problem:

$$\min_W S(G(W), D) \quad \text{subject to } G(W) \in \mathsf{DAG}, \tag{1}$$

where $W \in \mathbb{R}^{d\times d}$ is the adjacency matrix, $G(W)$ is the graph whose adjacency matrix is W, $S(G(W), D)$ is the scoring function measuring the fitness between graph $G(W)$ and the data D. Searching DAG from data is known to be an NP-hard problem [30]. Recently, a continuous optimization based approach called NOTEARS [46] has been proposed to handle this problem by introducing a matrix exponential based DAG constraint: $tr(e^{W\odot W}) = d$. The matrix exponential is given by the power series: $e^{W\odot W} = \sum_{k=0}^{\infty} \frac{1}{k!} (W \odot W)^k$, where the k-th term denotes the adjacency after k times walking on the graph, and $W \odot W$ makes the element of adjacency matrix non-negative. The trace of the k-th term ($k > 0$) should be zero if the graph is acyclic, because a node cannot go back to itself after k times walking. Therefore $tr(e^{W\odot W}) = d$ indicates that the graph is acyclic, where d is the trace of first term in the power series.

3 METHODOLOGY

To satisfy the path-specific causal fairness, we propose a causal graph learning based fairness framework CGF. The key of CGF framework is to make the causal graph and the data that the model relies on close to the ideal fair graph. The proposed framework contains three components: graph structure learning, fair regularization, and label prediction. The graph learning part aims to reveal the graph structure of the observation data, the fairness restriction targets at reducing the unfair paths, and the label prediction part outputs the fair predictions. These three components influence each other in that: These three components are interdependent, as the fairness restriction guides the graph learning by reducing the unfair edges, and the final prediction is made based on the values of its parent nodes which is detected by the graph learning component. Overall, the objective function is:

$$\mathcal{L} = L_{GL} + L_F + L_P, \tag{2}$$

where L_{GL} is the graph learning loss, L_F is the fairness restriction, L_P is the label prediction loss. In the following sections, we will first present the detailed implementation when the sensitive attributes are root nodes, provide the theoretical analysis about the generalization error, and then generalize the developed method to the case where sensitive attributes are non-root nodes.

3.1 CGF Framework

In the following three subsections, we will present the implementations of the three components in Eqn. (2).

3.1.1 Graph Structure Learning. The objective of graph structure learning is to find the optimal causal graph that fits the observed data best. Motivated by the continuous optimization of causal graph structure learning [46], the loss of graph structure learning is:

$$L_{GL} = \beta||D - \tilde{D}||_2^2 + \gamma_1 \left(tr(e^{W\odot W}) - (d_A + d_X + 1)\right)^2 + \gamma_2||W||_1, \tag{3}$$

where $W \in \mathbb{R}^{(d_A+d_X+1)\times(d_A+d_X+1)}$ is the adjacency matrix, and if its element $w_{i,j} \neq 0$, there exists an edge $V_j \to V_i$ with weight $w_{i,j}$ indicating the effect strength. d_A is the dimension of sensitive attributes, and d_X is the dimension of other features. D is the observed data, and $\tilde{D}$ is the reconstructed data based on W and D according to the causal graph. $\odot$ is the element-wise matrix multiplication operator, $e^{W\odot W}$ denotes the matrix exponential of $W \odot W$, $tr(\cdot)$ is the matrix trace. β, γ_1, and γ_2 are the hyper-parameters.

The first term in Eqn. (3) measures the fitness of the causal graph by calculating the difference between the observed data and the data reconstructed from the graph. The second term is the directed acyclic graph (DAG) constraint, which ensures the learned graph does not contain any cycle [46]. The third term is the ℓ_1 norm of the adjacency matrix which makes the learned graph to be sparse. The details of data reconstruction (i.e., $\tilde{D}$) and the DAG constraint are described as follows.

Cascade Data Reconstruction. In data reconstruction, each node is reconstructed based on its parents' reconstructed values. Eqn. (4) shows the reconstruction of node V_i:

$$\hat{V}_i = f_i(\hat{\Pi}(V_i)W[i_\Pi, i]), \tag{4}$$

where $f_i(\cdot)$ is the causal mechanism of node V_i, $\hat{\Pi}(V_i)$ is V_i's parent nodes after reconstruction. i_Π is the index set of V_i's parent nodes. $W[i_\Pi, i]$ is the elements in the adjacency matrix W whose row indices are in i_Π and column indices are i. It is noticed that the reconstruction of a node is based on its parents' reconstructed values, instead of the observed values, because the observed values

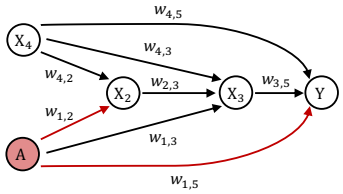


Figure 3: Cascade Data Reconstruction Example.

contain unfair effect if the parent nodes locate in the unfair paths. To satisfy this, the parent nodes should be reconstructed before the child nodes, thereby the data reconstruction follows a cascade reconstruction procedure with ascending order of depth.

We use a graph with five nodes (Figure 3), to illustrate the cascade data reconstruction. In the figure, A is the sensitive attribute, X_2, X_3, X_4 are regular features, and Y is the class label. The red arrows denote the unfair edges. The reconstruction order decided by the depth is: A & X_4, X_2, X_3, Y and the reconstruction procedure is:

$$\begin{aligned} &\text{Root Nodes:} && \hat{A} = A;\ \hat{X_4} = X_4; \\ &\text{Depth 1 node:} && \hat{X_2} = w_{1,2}A + w_{4,2}X_4 + b_2; \\ &\text{Depth 2 node:} && \hat{X_3} = w_{1,3}A + w_{4,3}X_4 + w_{2,3}\hat{X_2} + b_3; \\ &\text{Depth 3 node:} && \hat{Y} = w_{1,5}A + w_{4,5}X_4 + w_{3,5}\hat{X_3} + b_Y, \end{aligned} \tag{5}$$

where $w_{i,j}$ is the element of the i-th row and the j-th column in adjacency matrix W, b_i is the intercept term, and $\tilde{D} = [\hat{A}, \hat{X}_2, \hat{X}_3, \hat{X}_4, \hat{Y}]$. In this example, we adopt the linear function as the causal mechanism, and it can generalize to the more complex functions, such as neural network by changing $f_i(\cdot)$ in Eqn. (4).

DAG Constraint. The second term in Eqn. (3) is the directed acyclic graph (DAG) constraint, which ensures that there is no cycle in the learned graph [46], as a node cannot affect itself. The trace of adjacency matrix's exponential is adopted in the second term to measure the graph acyclic.

3.1.2 Fairness Regularization. The goal of fairness regularization is to reduce the unfair edges, so that the sensitive attributes pass less effect through the unfair path. As mentioned previously, the element in the adjacency matrix not only represents the causal direction, but also indicates the effect strength along this edge. To reduce the unfair edges, the elements in the adjacency matrix associate with unfair edges should be close to zero. Therefore, we apply the following fairness regularization on the adjacency matrix:

$$L_F = \alpha||W \odot M_F||_1, \tag{6}$$

where $\odot$ is the element-wise matrix multiplication, W is the adjacency matrix, and M_F is the fairness mask with the same dimension as W. The element of the j-th row, i-th column of M_F is set as 1 if edge $V_i \to V_j$ is unfair. More details related to the construction of the fairness mask are in the appendix. $||\cdot||_1$ is the ℓ_1 norm. The fairness regularization L_F minimizes the total strength of effect on the unfair edges, which reduces the effect flow along the unfair paths. By regularizing on the adjacency matrix, fairness regularization is able to eliminate the unfair path $A \to Y$ and $A \to X_2$ in Figure 3. Therefore, with the fairness regularization, the reconstructed data $\tilde{D}$ is a correction of the original data D with unfair effect reduced.

Fairness Mask Construction. Our proposed method only requires prior knowledge of what attributes are allowed or not allowed to construct the fairness mask M_F. For example, in online marketing example, Figure 2, the prior knowledge is that Race/Gender R is only allowed to affect Y through Preference Q. Therefore, in the fairness mask, in the i-th column, only the j-th row is set to be 0, and all others in the i-th column are set as 1, if Race/Gender is the i-th variable, and Preference is the j-th variable. It means that except Race/Gender $R \to Q$ Preference, all other paths including Race/Gender $R \to Z$ or $R \to Y$ are all unfair paths.

3.1.3 Label Prediction. Since the unfair effect has been reduced in the reconstruction data, the prediction based on the obtained reconstruction is fair. The label Y is predicted as: $\hat{Y} = \tilde{D}W_Y$, where W_Y is the last column of W, which indicates the existence of the edges and their effect strength starting from other nodes to Y and $\tilde{D}$ is the reconstructed data. Accordingly, the prediction loss is: $L_P = ||Y - \hat{Y}||_2^2$.

3.1.4 Generalization to Nonlinear Causal Mechanism. In Eqn. (5) and label prediction, the adopted linear causal mechanism can generalize to more advanced functions by modifying the cascade data reconstruction and the adjacency matrix. Assume we choose neural network (NN) as the causal mechanism, the reconstruction of node V_i is: $\hat{V}_i = f_i^{NN}(DW_i^{NN})$, where $f_i^{NN}(\cdot)$ represents the neural network, $W_i^{NN} \in \mathbb{R}^{(d_A+d_X)\times d_{NN}}$ is the parameter of the first linear layer in $f_i^{NN}(\cdot)$, d_{NN} is the dimension of the first hidden layer. Namely, to use the NN mechanism, replace the linear model in Eqn. (5) with the NN whose first layer is the linear layer and all those NNs share the common hidden layers, as suggested in [22]. The adjacency matrix is constructed based on the parameter in the first linear layer. Specifically, each element in the adjacency matrix W is calculated as: $w_{i,j} = ||W_i^{NN}[j,:]||_2^2$, where $W_i^{NN}[j,:]$ is the j-th row in W_i^{NN}.

3.1.5 Initialization and Optimization. When the sensitive attributes are root nodes, the overall loss function is:

$$\begin{aligned} \mathcal{L} = &||Y - \hat{Y}||_2^2 + \beta||D - \tilde{D}||_2^2 + \gamma_1 \left(tr(e^{W\odot W}) - (d_A + d_X + 1)\right)^2 \\ &+\gamma_2||W||_1 + \alpha||W \odot M_F||_1, \end{aligned} \tag{7}$$

where $\hat{Y}$ is defined in Eqn. (6), and $\tilde{D}$ is the reconstruction of D via the cascade data reconstruction presented in Section 4.1.1.

Initialization. As mentioned previously, the cascade data reconstruction requires acyclic graph. To satisfies this, we can initialize the adjacency matrix by the following two ways: (1) adopt the prior knowledge about the basic acyclic graph; (2) pre-train the parameter by the following objective function:

$$||Y - \hat{Y}||_2^2 + \beta||D - \tilde{D}'||_2^2 + \gamma_1 \left(tr(e^{W\odot W}) - (d_A + d_X + 1)\right)^2, \tag{8}$$

where $\tilde{D}'$ is the data reconstructed by the observed data. Each node V_i in $\tilde{D}'$ is calculated as: $\hat{V}_i' = f_i\left(\Pi(V_i)W[i_\pi, i]\right)$, where $\Pi(V_i)$ is the node V_i's observed value, and $W[i_\pi, i]$ is the same as the one in Eqn. (4). Eqn. (8) replaces the cascade data reconstruction $\tilde{D}$ in Eqn. (7) with regular data reconstruction $\tilde{D}'$, which not strictly requires acyclic graph.

Optimization. We adopt the Adam [19] to optimize both Eqn. (8) and Eqn. (7). Besides, at each iteration of optimizing Eqn. (7), the adjacency matrix W is forced to be acyclic.

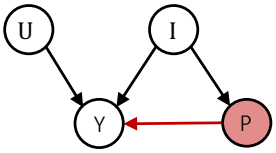


Figure 4: Sensitive Attributes as Non-root Nodes.

3.2 Theoretical Analysis

Here we provide the theoretical analysis about the generalization error. Following the results in [22, 24], we have:

Theorem 3.1. *Suppose the data D follow the Gaussian distribution, the generalization error of the proposed fair classifier h_F on the observed dataset, which is denoted as $\epsilon_{h_F}^{\mathcal{D}^{ob}}$, satisfies the following inequality with probability $1-\delta, \forall \delta > 0$:*

$$\epsilon_{h_F}^{\mathcal{D}^{ob}} \leq \left(||D - \tilde{D}||_{fro}^2 + \tau\sqrt{\frac{2\log\frac{2}{\delta}}{m}} + \frac{C_1}{\sqrt{m\delta}} + \frac{C_2}{\sqrt[4]{m\delta}} + C_3 \right)^{\frac{1}{2}} + 4\hat{\epsilon}_{h_F}^{\mathcal{D}^F} + \frac{1}{m}\left[\mathcal{R}_{dag} + C_4(\mathcal{R}_{l_1} + \mathcal{R}_F) + \log(\frac{8}{\delta})\right] + C_5, \tag{9}$$

where $\mathcal{D}^{ob}$ is the distribution of the observed data and D is its observed sample. $\tilde{D}$ is the reconstructed data by cascade reconstruction, and its distrubution is denoted as $\mathcal{D}^F$. $\epsilon_h^{\mathcal{D}} = \int_{\mathcal{D}} \ell_h(a,x)p^{\mathcal{D}}(a,x)dadx$, is the expected error on the underlying space $\mathcal{D}$, and $\ell_h(a,x) = \int_Y L(Y, h(a,x))p(Y|a,x)dY$ is the expected error on a single point (a,x). $\hat{\epsilon}_{h_F}^{\mathcal{D}^F}$ is the empirical error of classifier h_F on $\mathcal{D}^F$. $\mathcal{R}_{dag}$ and $\mathcal{R}_{l_1}$ are the value of DAG constrain, ℓ_1 regularization, which are the last two terms in Eqn. (3). $\mathcal{R}_F$ is the value of fairness regularization defined in Eqn. (6). $\tau, C_1 \sim C_5$ are constants.

Theorem 3.1 give an upper bound of the generalization error of the fair classifier on the observed dataset. The upper bound shows that the generalization error is related to the qualities of the reconstruction and the classifier trained on the fair dataset, which is exactly the two terms in our objective function. The reconstruction part in Theorem 3.1 also represents the fairness level, since the fairer the data is, the fairer the dataset is, the smaller the reconstruction error. Detailed proof of Theorem 3.1 is in the Appendix.

3.3 Generalization to Non-Root Node Case

Most of the existing works consider sensitive attributes such as age, gender, and race that can only be the root nodes in the causal graph. However, in some real-world applications, sensitive attributes are affected by other variables. For example, in the recommendation system, the item popularity should not affect whether this item to be recommended [10, 47], for the purpose of recommendation diversity. Figure 4 shows causal graph of the recommendation example where U and I represent user and item respectively, P denotes the item popularity and Y denotes whether the user click the item. In this example, the item popularity P is the sensitive attribute, and it is the non-root node as it is affected by the item's characteristics.

Challenge. When the sensitive attributes A are non-root nodes, their parent nodes $\Pi(A)$ also contain the information about those sensitive nodes. If the parent nodes have other causal pathways to the label node Y not passing the sensitive node, the information related to the sensitive attributes can still reach the label node through those paths. Therefore, the proposed framework in Section 3.1 requires slight modification to handle this case.

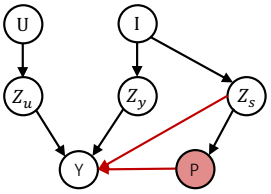


Figure 5: Causal Graph with Effect Diversion.

Effect Diversion. To address the challenge, two latent nodes, Z_s and Z_y, are added between the parent nodes and the sensitive nodes to divert the effect flow. This allows the fairness regularization proposed in the previous section to be applied to the unfair flow. Z_s controls the effect from the sensitive attributes' parent nodes to the label node passing through the sensitive attributes, while Z_y controls the effect from the parent node to the label node not passing through the sensitive attributes. Thus, the paths from the parent nodes $\Pi(A)$ to the label node Y are divided into two categories, one containing only Z_y and the other containing only Z_s. By separating Z_y and Z_s, the fairness regularization can be directly applied to the paths containing Z_s, as no other paths are exposed to information leakage risk. When there are multiple non-root sensitive attributes, the same procedure is applied to each one of them.

Figure 5 shows the causal graph with effect diversion in the recommendation example, where the latent node Z_u is user embedding, Z_s is the item popularity related embedding and Z_y is the clicking variable Y related embedding. The parent node I affects item popularity only through Z_s, and other paths from I to Y all pass through Z_y. The red arrows indicate the unfair paths where the fairness regularization is applied on.

Objective Function. In Eqn. (3), the reconstruction part requires the variables' observed values, while in this case latent nodes lack that. To address this issue, we notice that to fit the graph with effect diversion, Z_y and Z_s should have less overlapped information. To fulfill this, the orthogonal regularizations on Z_y and Z_s are adopted in the graph structure learning part to ensure the separation of the effect flow. Overall, the objective function is:

$$\mathcal{L} = ||Y - \hat{Y}||_2^2 + \beta||D - \tilde{D}||_2^2 + \beta_z \sum_{i=1}^{\mathcal{I}} cos(Z_s^{(i)}, Z_y^{(i)}) + \gamma_1 \left(tr(e^{W\odot W}) - (d_A + d_X + 1)\right)^2 + \gamma_2||W||_1 + \alpha||W \odot M_F||_1, \tag{10}$$

where $\mathcal{I}$ is the number of total items. $cos(\cdot,\cdot)$ denotes cosine similarity, which ensures the orthogonality between $Z_s^{(i)}$ and $Z_y^{(i)}$, and other types of correlation measure such as HSIC [12] can be adopted. Compared with Eqn. (2), the graph structure learning part is modified to satisfy the effect diversion design. In the recommendation system example, $||Y - \hat{Y}||_2^2$ is the clicking prediction error, and $||D - \tilde{D}||_2^2$ is the error of predicting the item popularity.

4 EXPERIMENT

We experiment on two real-world datasets, Adult Dataset (one of the commonly adopted datasets when evaluating algorithm fairness) and MovieLens Dataset (one of the popular datasets in recommendation area) to confirm: (1) The proposed CGF framework works on both cases where sensitive attributes are root or non-root nodes. (2) Our proposed framework provides a better trade-off between

utility and fairness. More details regarding implementations are listed in the Appendix.

4.1 Experiment on Adult Dataset

Dataset. Adult dataset[1], is a commonly adopted dataset for fairness evaluation. In this dataset, there are total 48842 individuals and each has 14 attributes regarding their demographic information, jobs, and level of education. The class label is binary indicating whether an individual's income is above or below $50k$. The objective is to predict the income class given an individual's attributes.

Unfair Paths. As suggested in [26], the direct path "Gender" $\rightarrow$ "Income", and the paths containing edge "Gender" $\rightarrow$ "Married" are all unfair. Namely, the gender should not direct affect the income and meanwhile it is not allowed to affect income through marital status.

Baselines. The logistic regression model (LR) and neural network (NN) constructed from raw data is adopted as the baseline. We also adopt the Fair Inference (FIO) [26] and PSE-DR [43] as the baselines. FIO method directly minimizes the path-specific causal effect from sensitive attributes to the label nodes through the unfair paths. The causal graph used in FIO is the same as the one in the original paper [26], and is shown in the appendix. PSE-DR is a two-step method which first learns the causal graph and then trains the classifier based on the corrected data generated from the learned causal graph. Our proposed models are denoted as LR-CGF and NN-CGF, which take the linear logistic regression model and neural network as the causal mechanism function, respectively.

Data Pre-processing for Baseline PSE-DR. To run the code of baseline PSE-DR[2] provided by the authors in [43], the Adult datasets requires additional stratification step to reduce the number of categories of each variable. The procedure of each variable is: higher_edu: higher_edu: $\lfloor$higher_edu/10$\rfloor$; high_hours: $\lfloor$high_hours/20$\rfloor$; managerial_occ: $\lfloor$managerial_occ/5$\rfloor$; gov_jobs: $\lfloor$gov_jobs/5$\rfloor$; age: $\lfloor$age/20$\rfloor$; native_country: $\lfloor$native_country/5$\rfloor$, married: $\lfloor$married/3$\rfloor$, where $\lfloor x \rfloor$ denotes the floor of the scalar x, which is the largest integer i, such that $i <= x$.

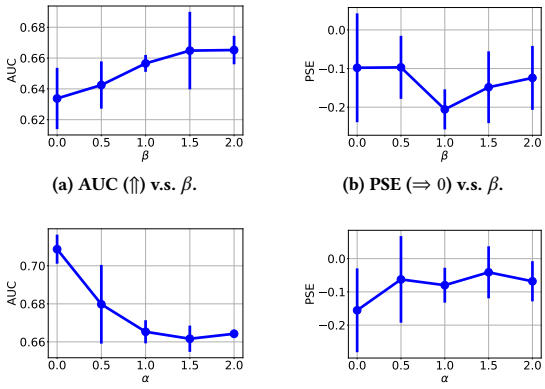
Evaluation Metrics. Due to label imbalance, Area Under the ROC Curve (AUC) is adopted as the utility metric. The higher the AUC value, the better the utility. Following [26], we adopt the path-specific causal effect (PSE) [27, 29] to measure the fairness. The PSE value may have a negative value indicating the negative effect. The closer the PSE value is to 0, the fairer the model is.

4.1.1 Result Analysis. Table 1 summarizes the results of different methods on Adult dataset with 5-fold cross validation. It is observed that compared with baseline methods, our proposed methods are fairer and meanwhile have better utility. On this dataset, when evaluated on the test set randomly split from the original dataset, there is a trade-off between fairness and utility. The baseline LR has the highest AUC but it is unfair with the highest PSE value among all methods. Compared with LR, the methods with fairness design sacrifice the utility for fairness. Among the methods with fairness design, our proposed methods LR-CGF and NN-CGF have a better trade-off between utility and fairness, which validates the effect of regularizing the fairness at the graph level. We also notice that the method with nonlinear causal mechanism performs best in terms of both utility and fairness. The reason is that compared with linear causal mechanism, the neural network can reconstruct the data better while ensuring fairness. This observation also confirms Theorem 3.1 that the better the reconstruction is, the better the performance is.

Method	AUC ($\Uparrow$)	PSE ($\Rightarrow 0$)
LR	0.712 ± 0.005	3.508 ± 0.005
NN	**0.721 ± 0.012**	2.068 ± 0.223
FIO	0.505 ± 0.007	1.048 ± 0.003
PSE-DR	0.686 ± 0.018	0.450 ± 0.151
LR-CGF	0.507 ± 0.099	0.925 ± 0.073
NN-CGF	**0.689 ± 0.012**	**−0.198 ± 0.109**

Table 1: Results on Adult Dataset. $\Uparrow$: the higher the better, and $\Rightarrow 0$: the closer to 0, the better.

We further experimentally explore the relationship between the reconstruction and fairness regularization. We fix one part's hyper-parameter and tune the other one. Figure 6 reports the results of NN-CGF. From Figure 6a and 6b, it is observed that, with the increasing strength of reconstruction, it improves the accuracy but reduces the fairness. The fairness regularization has an opposite effect with reconstruction part. As shown in Figure 6c and 6d, the fairness regularization improves the model fairness but reduces utility. Overall, the reconstruction part and the fairness regularization, together, control the trade-off between model utility and fairness.



(a) AUC ($\Uparrow$) v.s. β. **(b) PSE ($\Rightarrow 0$) v.s. β.**

(c) AUC ($\Uparrow$) v.s. α. **(d) PSE ($\Rightarrow 0$) v.s. α.**

Figure 6: Effects of Reconstruction and Fairness Regularization.

4.2 Experiment on Recommendation Dataset

In this section, we conduct the experiment on real-world recommendation data to show that, in the case when sensitive attributes are non-root nodes, our proposed framework is able to provide fair predictions with high utility.

Dataset. The MovieLens dataset [13] is adopted to validate the performance of CGF. Following the settings in [10], the sensitive attribute *item popularity* is added to each item, and for each item, the value of item popularity is 1 if its total exposure is top 20%, otherwise 0. The item popularity is a non-root node since it is affected by item characteristics. For each user, we sort their interactions according to the timestamp, and the last interaction is put into the test set, and others are in the training set. The validation set is the last interaction of each user in the training set.

[1] https://archive.ics.uci.edu/ml/datasets/adult
[2] https://www.yongkaiwu.com/publication/zhang-2017-causal/zhang-2017-causal.zip

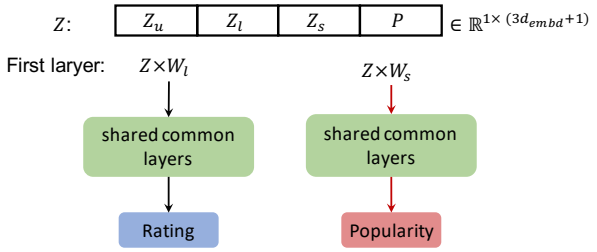


Figure 7: Neural network structure of MLP-CGF.

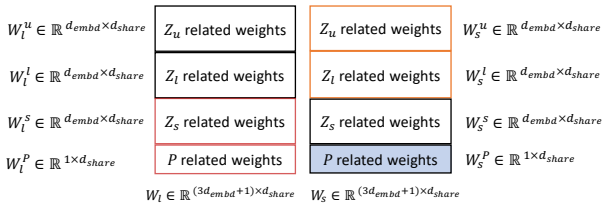


Figure 8: Details of W_l and W_s in the first layer.

Baselines. We adopt Matrix Factorization (MF) [20], Generalized Matrix Factorization (GMF) and Multiple Layer Perceptron (MLP) [14] as our baselines. GMF and MLP are also the base models of our proposed framework, and we name our methods as GMF-CGF and MLP-CGF accordingly.

Neural Network Structure of CGF. Figure 7 shows the neural network structure of MLP-CGF, where Z is the concatenation of user embedding Z_u, rating related item embedding Z_l, popularity related item embedding Z_s and popularity P. The weight W_l and W_s in the first layer control the flow of information into rating prediction and popularity prediction, separately. After the first layer, the ratings and popularity tasks shared several common layers, followed by their specific prediction layers.

The weighting matrices W_l and W_s control the information flow from Z to rating and popularity prediction, respectively. The details of W_l and W_s are shown in Figure 8. The dimension of W_l and W_s are both $(3d_{embd} + 1) \times d_{share}$, where d_{embd} is the embedding size, d_{share} is the dimension of the first layer in shared common layers. Each of the weights contains four parts that are Z_u related weights, Z_l related weights, Z_s related weights and P related weights. the P related weights W_s^l in W_s is zero matrix because in popularity prediction, ground-truth popularity value should be the input. Notice that Z_l and X_u should not affect item popularity, we also minimize the norm of W_s^u and W_s^l. Since the paths $Z_s \rightarrow Y$ and $P \rightarrow Y$ are unfair as shown in Figure 5, the fairness regularization is: $L_F = \alpha \left(||W_l^s||_1 + ||W_l^P||_1 \right)$.

Evaluation Metrics. In terms of utility/accuracy measure, Top-k ranking metrics hit rate (HR) and normalized discounted cumulative gain (NDCG) are adopted to measure the recommendation performance. Following [10], the Gini Index and Popularity Rate (PR) are also adopted to measure the fairness. Given the item impression list $\mathcal{K} = [k_1, k_2, \cdots, k_{\mathcal{I}}]$, where k_i represents the number of exposures of the i-th item, the Gini Index is defined as: $\text{Gini Index}(\mathcal{K}) = \frac{1}{2|\mathcal{I}|^2 \bar{k}} \sum_{i=1}^{\mathcal{I}} \sum_{j=1}^{\mathcal{I}} |k_i - k_j|$, where $\mathcal{I}$ is the number of total items, $\bar{k}$ is the mean of item impression list $\mathcal{K}$. Gini Index measures the statistical dispersion of the item exposure. Popularity rate is the ratio of popular items among the total items recommended to the users, and is defined as: $\text{PR}(\mathcal{K}) = \sum_{i=1}^{\mathcal{I}} P_i k_i / \sum_{i=1}^{\mathcal{I}} k_i$, where P_i is binary denoting the i-th item's value of item popularity. For HR and NDCG, the higher the value is, the better the performance is. For Gini Index and PR, the lower the value is, the fairer the model is.

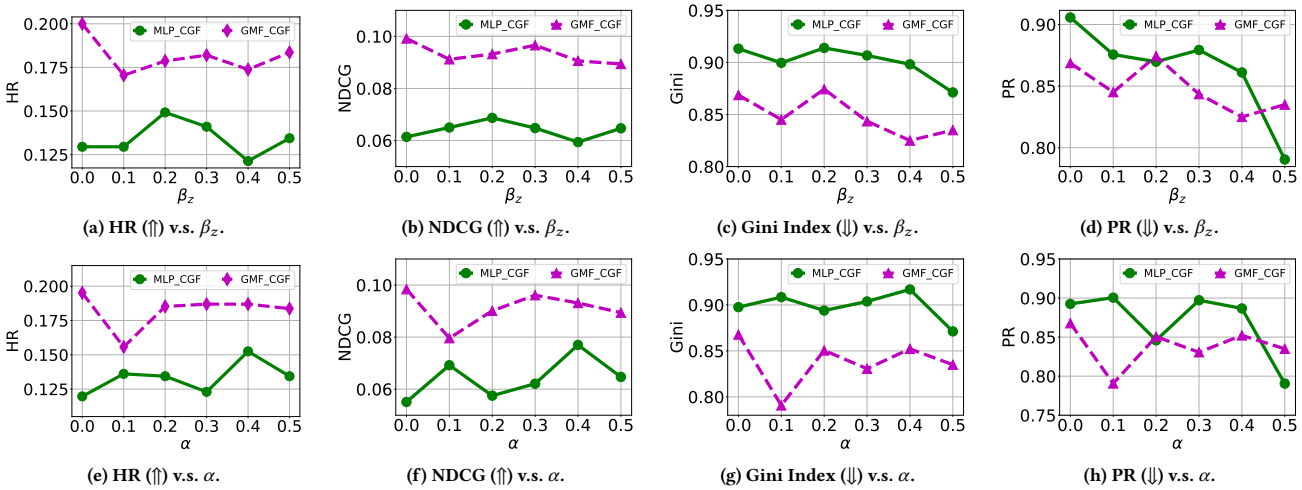
Method	HR (⇑)	NDCG (⇑)	GINI (⇓)	PR (⇓)
MF	**0.197**	**0.103**	0.878	0.861
GMF	0.195	0.098	0.875	0.857
MLP	0.149	0.077	0.919	0.925
GMF-CGF (ours)	0.166	0.084	**0.837**	**0.773**
GMF-CGF w/o ord	0.174	0.089	0.853	0.820
GMF-CGF w/o fairness	0.166	0.089	0.857	0.808
GMF-CGF w/o ord & fairness	0.195	0.102	0.881	0.860
MLP-CGF (ours)	0.116	0.059	0.882	0.844
MLP-CGF w/o ord	0.112	0.054	0.902	0.891
MLP-CGF w/o fairness	0.141	0.066	0.923	0.932
MLP-CGF w/o ord & fairness	0.139	0.066	0.903	0.863

Table 2: Results on the Movielens Dataset. ⇓ indicates the lower, the better.

4.2.1 Results Analysis. Table 2 shows the results of baselines and our proposed methods with Top 10 rankings metrics. We also list the results of our methods' variants. $*$-CGF w/o rec denotes the CGF without the orthogonal regularization part, i.e., $\beta_z = 0$ in Eqn. (10). $*$-CGF w/o fairness and denotes CGF without the fairness regularization part ($\alpha = 0$). $*$-CGF w/o ord & fairness is CGF without both of these two parts.

In terms of fairness, our proposed methods recommend more diverse items, and meanwhile, have the comparable recommendation accuracy to the baselines. This observation verifies that our proposed method makes the base model to be fair without scarifying too much utility. It is worth to mention that the results measured by GINI and PR are also the indirect indicator of how good the disentanglement of the effect from sensitive attributes' parent nodes to label node. The better it disentangles, the fairer the model. Furthermore, the ablation results shown in the Table 2 indicate that the orthogonal regularization and the fairness regularization both contribute to the model fairness.

To further analyze the effect of orthogonal regularization and fairness regularization, in Figure 9, we plot the four metrics with respect to different regularization strengths by tuning one hyper-parameter and fixing the others. From this figure, we can observe that the stronger the regularization strength is, the fairer the model is, and the more utility is sacrificed. Furthermore, the utility and fairness trade-off can be controlled by tuning the values of two regularizations' hyper-parameters. We also notice that MLP-CGF performs slightly different in terms of HR and NDCG: The stronger the orthogonal regularization and the fairness regularization, the better the performance. The reason is that compared with GMF-CGF, MLP-CGF has more learnable parameters in the neural network, and adding those regularizations would prevent the over-fitting.



(a) HR (⇑) v.s. β_z. (b) NDCG (⇑) v.s. β_z. (c) Gini Index (⇓) v.s. β_z. (d) PR (⇓) v.s. β_z.

(e) HR (⇑) v.s. α. (f) NDCG (⇑) v.s. α. (g) Gini Index (⇓) v.s. α. (h) PR (⇓) v.s. α.

Figure 9: Effects of the Orthogonal Regularization and Fairness Regularization.

5 RELATED WORK

Most of the existing path-specific causal fairness works restrict the unfair pathways by reducing their path-specific effect. In [25, 26], the prediction error and the path-specific effect along with unfair causal pathways are jointly minimized. The work proposed in [43] designs a two-step algorithm, by first learning the graph structure and then minimizing the prediction error with PSE regularization. In [33], the authors adopt the response-function variable to bound the path-specific causal fairness. Instead of directly minimizing the path-specific effect, a latent inference-projection based method is proposed in [6] to correct the variables that are the descent nodes of sensitive attributes. In [15], the CEVAE framework [23] is adopted to infer the causal mechanism based on the pre-defined causal graph, and then the auxiliary prediction model is constructed based on the selected causal relation along with the fairness requirement.

Relation to Existing Works. Most of the above existing works require the prior knowledge about causal graph to calculate the PSE or to correct sensitive variables' descent variables, which is hard to be satisfied in real-world applications. Compared with the work in [43] that has a separated time-consuming causal structure learning step, our work applies the fairness constraint on the continuous-optimization based graph structure learning, which can efficiently obtain the causal graph and simplify the PSE calculation. Furthermore, it is worth mentioning that all the above existing works assume that the sensitive attributes are root nodes. The proposed framework is the first work that generalizes to the case when sensitive attributes are non-root nodes under path-specific causal fairness. Additionally, our proposed framework is motivated by the work of utilizing the causal graph discovery to enhance the machine learning generalization ability [22]. Compared with [22], the proposed CGF framework contains the cascade reconstruction step, which is the major difference. With the cascade reconstruction step, the unfairness contained in the original data can be corrected. Besides, CGF also has the fairness regularization in our proposed method, which reduces the unfair paths in the causal graph and meanwhile assures that the data correction follows the fair graph.

We also notice that in the recommendation system domain, there are some existing works that handle the popularity bias by examining the causal link between the popularity and the item [8, 9, 44]. Compared with works in this line, in our work, we adopt the structural causal model and estimate the variable's generation function by neural work, so that we can directly recover the data under the fair graph (there is no path from popularity variable P to Y), and use the corrected data to train the recommendation model. This is the key difference between our work and other existing work.

6 CONCLUSIONS AND FUTURE WORK

In this work, we propose a novel causal graph based fair prediction framework under path-specific causal fairness. The core of the proposed framework is to ensure that the graph adopted by the prediction model should be close to the fair graph. To fulfill this, we integrate the graph structure learning and the fairness regularization in an interactive way. The learned graph structure reveals the causal graph of the original observations with unfair edges eliminated, and the data reconstructed from the learned graph is close to the original observations with unfair effect corrected. Based on the corrected causal graph and its associated data, the prediction model achieves the path-specific causal fairness. Experimental results confirm that the proposed framework ensures fair predictions and meanwhile retains the comparable utility. We also generalize the proposed framework to the case of sensitive attributes being non-root nodes by effect redividing, which is further validated by experiments on a real-world recommendation dataset.

In this paper, we assume that there are no latent confounders in the dataset. When this assumption is not satisfied, the causal graph may not be identified from the observation data. Recently, some causal discovery works that target to recover the causal graph in the presence of latent confounders [5, 34] have been developed. We also have a strong assumption that the data follows the Gaussian distribution in the theoretical analysis. Relaxing the above two assumptions and generalizing our work to the latent confounders case will be the future work.

ACKNOWLEDGMENTS

The work was partially supported by the National Science Foundation under Grant NSF IIS-2226108 and IIS-2141037.

REFERENCES

[1] Chen Avin, Ilya Shpitser, and Judea Pearl. 2005. Identifiability of Path-Specific Effects. In *Proc. of IJCAI'05*. 357–363.
[2] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna M. Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proc. of ACL'20*. 5454–5476.
[3] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv preprint arXiv:1607.06520* (2016).
[4] Robin Burke. 2017. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093* (2017).
[5] Ruichu Cai, Feng Xie, Clark Glymour, Zhifeng Hao, and Kun Zhang. 2019. Triad Constraints for Learning Causal Structure of Latent Variables. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. 12863–12872.
[6] Silvia Chiappa. 2019. Path-specific counterfactual fairness. In *Proc. of AAAI'19*, Vol. 33. 7801–7808.
[7] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proc. of FAT'19*. 120–128.
[8] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. 2022. A survey of research on fair recommender systems. *arXiv preprint arXiv:2205.11127* (2022).
[9] Chen Gao, Yu Zheng, Wenjie Wang, Fuli Feng, Xiangnan He, and Yong Li. 2022. Causal Inference in Recommender Systems: A Survey and Future Directions. *CoRR* abs/2208.12397 (2022). https://doi.org/10.48550/arXiv.2208.12397 arXiv:2208.12397
[10] Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, and Yongfeng Zhang. 2021. Towards Long-term Fairness in Recommendation. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*. ACM, 445–453.
[11] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862* (2019).
[12] Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alexander J Smola. 2007. A Kernel Statistical Test of Independence. In *NIPS*.
[13] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5, 4 (2015), 1–19.
[14] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proc. of WWW'17*. 173–182.
[15] Rik Helwegen, Christos Louizos, and Patrick Forré. 2020. Improving Fair Predictions Using Variational Inference In Causal Models. *arXiv preprint arXiv:2008.10880* (2020).
[16] Yaowei Hu, Yongkai Wu, Lu Zhang, and Xintao Wu. 2020. Fair Multiple Decision Making Through Soft Interventions. In *Proc. of NeurIPS'20*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.).
[17] Wen Huang, Yongkai Wu, Lu Zhang, and Xintao Wu. 2020. Fairness through Equality of Effort. In *Proc. of WWW'20*. 743–751.
[18] Kosuke Imai, Luke Keele, and Dustin Tingley. 2010. A general approach to causal mediation analysis. *Psychological methods* 15, 4 (2010), 309.
[19] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[20] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.
[21] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Proc. of NeurIPS'17*. 4066–4076.
[22] Trent Kyono, Yao Zhang, and Mihaela van der Schaar. 2020. CASTLE: Regularization via Auxiliary Causal Graph Discovery. In *Proc. of NeurIPS'20*.
[23] Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. In *Proc. of NeurIPS'17*. 6449–6459.
[24] Andreas Loukas. 2017. How Close Are the Eigenvectors of the Sample and Actual Covariance Matrices?. In *Proc. of ICML'17*, Doina Precup and Yee Whye Teh (Eds.), Vol. 70. 2228–2237.
[25] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. 2019. Optimal training of fair predictive models. *arXiv preprint arXiv:1910.04109* (2019).
[26] Razieh Nabi and Ilya Shpitser. 2018. Fair Inference on Outcomes. In *Proc. of AAAI'18*. AAAI Press, 1931–1940.
[27] Judea Pearl. 2001. Direct and indirect effects. In *Proc. of UAI'01*. 411–420.
[28] Chris Russell, Matt J. Kusner, Joshua R. Loftus, and Ricardo Silva. 2017. When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness. In *Proc. of NeurIPS'17*. 6414–6423.
[29] Ilya Shpitser. 2013. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive science* 37, 6 (2013), 1011–1035.
[30] Tomi Silander and Petri Myllymaki. 2012. A simple approach for finding the globally optimal Bayesian network structure. *arXiv preprint arXiv:1206.6875* (2012).
[31] Yongkai Wu, Lu Zhang, and Xintao Wu. 2018. On discrimination discovery and removal in ranked data using causal graph. In *Proc. of KDD'18*. 2536–2544.
[32] Yongkai Wu, Lu Zhang, and Xintao Wu. 2019. Counterfactual Fairness: Unidentification, Bound and Algorithm. In *Proc. of IJCAI'19*. 1438–1444.
[33] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. 2019. PC-Fairness: A Unified Framework for Measuring Causality-based Fairness. In *Proc. of NeurIPS'19*. 3399–3409.
[34] Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. 2020. Generalized Independent Noise Condition for Estimating Latent Variable Causal Graphs. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
[35] Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2019. Achieving causal fairness through generative adversarial networks. In *Proc. of IJCAI'19*.
[36] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A Survey on Causal Inference. *ACM Trans. Knowl. Discov. Data* 15, 5 (2021), 74:1–74:46.
[37] Sirui Yao and Bert Huang. 2017. Beyond Parity: Fairness Objectives for Collaborative Filtering. In *Proc. of NeurIPS'17*. 2921–2930.
[38] Muhammad Bilal Zafar, I. Valera, M. G. Rodriguez, and K. Gummadi. 2015. Learning Fair Classifiers. *arXiv: Machine Learning* (2015).
[39] Junzhe Zhang and Elias Bareinboim. 2018. Equality of Opportunity in Classification: A Causal Approach. In *Proc. of NeurIPS'18*. 3675–3685.
[40] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
[41] Lu Zhang and Xintao Wu. 2017. Anti-discrimination learning: a causal modeling-based framework. *International Journal of Data Science and Analytics* 4, 1 (2017), 1–16.
[42] Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. Situation Testing-Based Discrimination Discovery: A Causal Inference Approach. In *Proc. of IJCAI'16*, Subbarao Kambhampati (Ed.). 2718–2724.
[43] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In *Proc. of IJCAI'17*. 3929–3935.
[44] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal Intervention for Leveraging Popularity Bias in Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 11–20. https://doi.org/10.1145/3404835.3462875
[45] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proc. of the EMNLP'17*.
[46] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. 2018. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Proc. of NeurIPS'18*. 9492–9503.
[47] Ziwei Zhu, Xia Hu, and James Caverlee. 2018. Fairness-aware tensor-based recommendation. In *Proc. of the CIKM'18*. 1153–1162.