Causally Fair Node Classification on Non-IID Graph Data

Anonymous authors
Paper under double-blind review

Abstract

Fair machine learning seeks to identify and mitigate biases in predictions against unfavorable populations characterized by demographic attributes, such as race and gender. Recent research has extended fairness to graph data, such as social networks, but many neglect the causal relationships among data instances. This paper addresses the prevalent challenge in fair ML algorithms, which typically assume Independent and Identically Distributed (IID) data, from the causal perspective. Particularly, this work targets the circumstance where nodes with different neighborhood structures follow different causal mechanisms, violating the invariance assumptions required for classical structural causal models and *do*-calculus. We base our research on the Network Structural Causal Model (NSCM) framework and develop a Message Passing Variational Autoencoder for Causal Inference (MPVA) to compute interventional distributions for causally fair node classification. Theoretical soundness is established under two conditions: Decomposability and Graph Independence. These conditions formalize when causal mechanism heterogeneity can be overcome by constructing a structural representation that restores invariance and facilitates the computation of interventional distributions using *do*-calculus in non-IID settings. Empirical evaluations on semi-synthetic and real-world datasets demonstrate that MPVA outperforms conventional methods by effectively approximating interventional distributions and mitigating bias. The implications of our findings underscore the potential of causality-based fairness in complex ML applications, setting the stage for further research into relaxing the initial assumptions to enhance model fairness.

1 Introduction

Amid the increasing prevalence of machine learning algorithms and models in the real world, ensuring unbiased predictions by identifying and mitigating biases is essential to maintaining equity and promoting the reliability of deploying machine learning to high-stake scenarios (Caton & Haas, 2020; Zafar et al., 2017a;b; Mehrabi et al., 2021; Pessach & Shmueli, 2023; Zliobaite, 2017; Quy et al., 2022; Wan et al., 2022). The past decade has seen the development of numerous fair learning algorithms designed to build fair machine-learning systems, which are grounded in various notions of fairness, including both statistical-based and causality-based ones (Pedreschi et al., 2009; Hardt et al., 2016; Zhang & Bareinboim, 2018b; Wen et al., 2019).

Despite significant progress in fair machine learning, many existing algorithms rely on the assumption of Independent and Identically Distributed (IID) (Caton & Haas, 2020; Zafar et al., 2017a; Hardt et al., 2016; Zafar et al., 2017b). In real-world scenarios, however, data instances are seldom independent and often exhibit connections, which are referred to as non-IID settings[1]. For instance, in predicting loan defaults, while each individual's data point appears independent, an individual's loan repayment behavior can be influenced by the experiences of their friends and family. To address the above issues, the research community has extended the fairness notions and studied fair machine learning in the context of graph mining. Group fairness notions like demographic parity, equalized odds, equal opportunity, and so on, have been extended to graph settings (Dong et al., 2022; Fan et al., 2021; Bose & Hamilton, 2019; Zhu et al., 2024b; Luo et al., 2024), and new fairness notions such as degree-related fairness and node pair distance-based fairness (Kang et al., 2020; Dong et al., 2021) have also been proposed.

[1]In this paper, the terms 'non-IID settings', 'graph settings', and 'network settings' are used interchangeably and refer to the situations where data instances are interconnected.

However, a significant gap in the current literature on fair graph learning is the lack of principled exploration of causality-based methods. Causal fairness plays a vital role in the fair machine learning field by modeling unfairness as the causal effect of the sensitive feature on the model prediction rather than relying solely on correlation. While causality-based fairness has been extensively studied in IID settings, with notions such as direct and indirect discrimination, counterfactual fairness, and path-specific counterfactual fairness being well-established (Zhang & Bareinboim, 2018a; Chiappa, 2019; Malinsky et al., 2019), its application in non-IID graph settings remains largely underexplored. Recent studies have highlighted that directly applying conventional fairness notions to non-IID settings without accounting for dependencies among individuals can yield biased outcomes (Zhang et al., 2022; Zhang, 2023). Although causal inference for non-IID settings has been studied in the context of interference (Hudgens & Halloran, 2008; Tchetgen & VanderWeele, 2012), the integration of these methods into machine learning workflows to measure causal unfairness in non-IID graph data presents considerable challenges, due to computation and estimation barriers such as the consistent interference assumption (Arbour et al., 2016a; Lee & Honavar, 2016; Arbour et al., 2016b). While preliminary efforts have explored formulating causal fairness in non-IID settings, e.g., (Agarwal et al., 2022; 2021; Ma et al., 2022; Yang et al., 2024), these studies lack a rigorous foundation for extending traditional inference techniques, such as *do*-calculus, to graph-structured data. A comprehensive theoretical and generalized framework for causal inference in such settings is still lacking.

In this work, we address a fundamental challenge in extending causal inference to graph data: the heterogeneity of causal mechanisms across nodes. In graph settings, the outcome of a node is typically influenced not only by its own attributes but also by the attributes of its neighbors. Crucially, the manner in which these influences are aggregated depends on the local network structure. As a result, nodes with different neighborhood structures effectively follow different causal mechanisms, violating the invariance assumptions that underlie classical structural causal models (SCMs) and the application of do-calculus. To overcome this challenge, our key insight is that, although node-level mechanisms may vary with local network structure, it is feasible to construct a representation of this structure that restores invariance. Building on the Network Structural Causal Model (NSCM), we leverage the Weisfeiler-Lehman (WL) graph isomorphism framework to encode local structural information through node colors. We then introduce two general conditions, Decomposability and Graph Independence, under which the networked causal process can be reduced to an equivalent causal model with a shared mechanism across nodes. Under these conditions, we show that interventional distributions in non-IID graph data can be expressed using standard do-calculus. In particular, we derive a representation of the interventional distribution that separates the effect of the sensitive attribute from the network structural variability. This formulation reveals that both observational and interventional distributions share a common latent functional component, which can be estimated from data and reused under interventions.

Motivated by this analysis, we propose a deep learning framework, the Message Passing Variational Autoencoder for Causal Inference (MPVA), which is designed to approximate this shared component. MPVA combines message passing neural networks (MPNNs) to capture neighborhood-level effects with conditional variational autoencoders (cVAEs) to model conditional distributions. Rather than directly enforcing fairness constraints, our approach estimates interventional distributions and incorporates them into a causal fairness regularization framework for node classification.

We evaluate our approach on both semi-synthetic and real-world datasets. The results demonstrate that MPVA more accurately estimates interventional quantities in graph settings and achieves improved fairness compared to existing methods that do not account for mechanism heterogeneity.

2 Related Work

Graph-based Causal Inference. Recently, the IID assumption in causal inference has been investigated and extended. A set of works extends the graphical causal modeling framework. Ogburn & VanderWeele (2014) extended DAGs for the interference relationships among individuals. Sherman & Shpitser (2018) modeled interference using chain graphs that permit modeling unknown interactions between individuals. Bhattacharya et al. (2019) proposed an interventional method for estimating causal effects under data dependence when the structure is known. In addition to the graphical modeling, researchers defined various effects to capture the relationships among variables and data points. Shpitser & Pearl (2008) defined the individual and group

average direct and indirect effects (a.k.a. spillover effect) in the interference situations. Ogburn & VanderWeele (2014) developed direct interference, interference by contagion and infectiousness, and avocational interference. In addition to modeling, recent years have witnessed a rise in papers on causal interference effect estimation. Hudgens & Halloran (2008) and VanderWeele & Tchetgen Tchetgen (2011) have developed new randomized procedures for unbiased estimands. Fatemi & Zheleva (2020) proposed a new experiment design approach to minimize interference bias and selection bias during estimation. Tchetgen Tchetgen et al. (2021) proposed a general g-computation method for causal interference.

Fairness on Graphs. Recently, algorithmic bias in machine learning has garnered significant attention from the research community, spawning a proliferation of methods and studies (Binns, 2020; Jiang et al., 2023; Verma & Rubin, 2018). Various notions of fairness have been proposed to define fairness formally, which can be categorized into two groups. The first category is *statistical parity*, which means the proportions of receiving favorable decisions for the protected and non-protected groups should be similar. The quantitative metrics derived from *statistical parity* include *risk difference*, *risk ratio*, *relative change*, and *odds ratio* (Wu et al., 2019; Hardt et al., 2016; Pedreschi et al., 2012). Recent works (Agarwal et al., 2021; Bose & Hamilton, 2019; Buyl & Bie, 2020; Dai & Wang, 2021; Dong et al., 2021; Kang et al., 2020) mitigate bias in node representation learning. Most of the works (Beutel et al., 2017; Zhang et al., 2018) are focused on adversarial learning, ensuring that the learned representations do not reliably predict the associated sensitive attribute. These works focus on eliminating the statistical dependency between the sensitive attribute and prediction from the learned representation but neglect the bias raised by the feature or graph structure due to the causal effect. Another category of fairness notion is counterfactual fairness, which is developed mainly under the structural causal model (SCMs) (Pearl, 2009). There are a few works (Ma et al., 2022; Agarwal et al., 2021; Yang et al., 2024) that extend the counterfactual fairness to graphs. However, most of these works ignore the potential biases introduced by the sensitive attributes of neighboring nodes and the causal effect of sensitive attributes on other nodes.

3 Preliminary

Structural Causal Model (SCM). Structural causal models (Pearl, 2009) provide a mathematical framework to understand causal relationships within a system. SCMs define the causal dynamics of a system through a collection of structural equations. Each variable X in the system is associated with a function f_X, such that $x = f_X(\mathsf{pa}_X, u_X)$. Here, pa_X denotes the values of other endogenous variables that directly influence X, and u_X represents the values of the exogenous variables impacting X. Each SCM is associated with a causal diagram consisting of a set of nodes for representing variables and a set of directed edges for representing the direct causal relations. We posit the causal Markovian model in this paper, i.e., all exogenous variables are mutually independent.

Causality-based Fairness Notions. Defining causality-based fairness notions is facilitated with the *do*-operator (Pearl, 2009), which simulates the physical interventions that force some variable to take certain values. Formally, the intervention that sets the value of S, a sensitive demographic characteristic (i.e., race or gender), to s is denoted by $do(S = s)$. The distribution of a variable X after the intervention on s is called the interventional distribution, denoted as $P(x|do(s)) := P(x|do(S = s))$. Causality-based fairness notions are usually defined by the disparity in the interventional distribution across different demographic groups, such as the total effect (Zhang & Bareinboim, 2018a), direct discrimination (Zhang et al., 2017), indirect discrimination (Zhang et al., 2017), and counterfactual fairness (Kusner et al., 2017). In this paper, we consider the total effect of S on Y defined as $\mathbb{E}[Y|do(s^+)] - \mathbb{E}[Y|do(s^-)]$ where s^+ and s^- represent the favorable and unfavorable groups of the demographic characteristics.

4 Problem Formulation

4.1 Network Structural Causal Model (NSCM)

Traditional SCM assumes that data instances are IID. To deal with non-IID data, recent studies have proposed to extend the SCM to capture the interference between individuals (e.g., (Ogburn et al., 2022)). These extensions usually involve using interference graphs (as illustrated in Fig. 1a) to describe both the causal

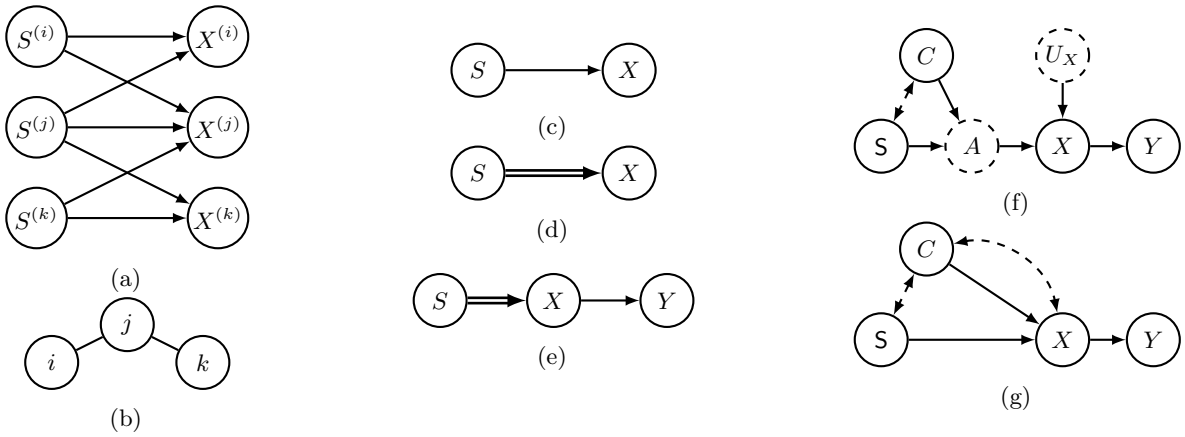

Figure 1: Graphs and diagrams. (a) The interference graph. (b) The network $\mathcal{G}$. (c) The causal diagram $\mathcal{C}$. (d) The networked causal diagram $\mathcal{N}$. (e) The networked causal diagram for node classification. (f) The causal diagram that is equivalent to the networked causal diagram in Fig. 1e. (g) The causal graph that violates Condition 2 Graph Independence.

relationship between features and the interference relationship between individuals. Inspired by this, in our work, we adopt a more general framework extended from traditional SCM, which we refer to as the Network Structural Causal Model (NSCM). In NSCM, we combine the network, which describes the existence of potential interference, with the graph, which describes the causal relationship between features. To distinguish between the two, NSCM explicitly considers two components: 1) network ($\mathcal{G}$), where each node represents an individual or instance and the edges depict the potential for interference between connected individuals; and 2) causal diagram ($\mathcal{C}$), where each node represents a feature and the edges depict the parent-child relationship between features determined by the structural equations. It is formally defined as follows.

Definition 1 (Network Structural Causal Model (NSCM))**.** An NSCM $\mathcal{M}$ is a quadruple $\mathcal{M} = \langle \mathcal{G}, \mathbf{U}, \mathbf{V}, \mathbf{F} \rangle$ where

1. $\mathcal{G}$ is a network that consists of a set of connected nodes.
2. $\mathbf{U}$ is a set of exogenous variables. For every node i in the graph, $\mathbf{u}^{(i)} \in \mathbf{U}$ is an instantiation of the exogenous variables received by node i.
3. $\mathbf{V}$ is a set of endogenous variables. For every node i in the graph, $\mathbf{v}^{(i)} \in \mathbf{V}$ is an instantiation of the endogenous variables received by node i.
4. $\mathbf{F}$ is a set of structural equations. For each variable $X \in \mathbf{V}$ and a node i, an equation $x^{(i)} = f(\mathsf{pa}_X^{(i)}, \{\mathsf{pa}_X^{(j)} : j \in \mathsf{ne}^{1:k}(i)\}, u_X^{(i)})$ determines the value of $x^{(i)}$, where $\mathsf{ne}^{1:k}(i)$ denotes the neighborhood of i within k hops.

An illustrative example of an NSCM with two variables S, X is shown in Figs. 1a, 1b and 1c, which show the interference graph, the network $\mathcal{G}$, and the causal diagram $\mathcal{C}$. In this example, the structural equations of this NSCM can be given by

$$x^{(i)} = f(s^{(i)}, \{s^{(j)} : j \in \mathsf{ne}^{1:k}(i)\}, u_X^{(i)}) \tag{1}$$

To further simplify representation, we introduce a hybrid graph–the networked causal diagram $\mathcal{N}$–that integrates network information with the causal diagram while omitting detailed interference structure, as shown in Fig. 1d. In this networked causal diagram, the solid line arrow represents the case where the causal effect is transmitted only from a variable of an individual to another variable of the same individual as seen in the traditional SCM (not showing in this example), while the double line arrow represents the existence of interference between different individuals in $\mathcal{G}$ when the causal effect is transmitted from one variable to another. For $k = 1$, the interference solely occurs between nodes that are immediate neighbors, while for $k > 1$, the interference can propagate through multiple hops in a message-passing fashion.

Given the NSCM formulation, the causal inference task is to infer the interventional distribution $P(x|do(s)) := P(x|do(S = s))$ from observational graph data.

4.2 Fair Node Classification

We will apply our causal inference techniques to the fair node classification problem as the downstream task. Denote the sensitive feature by S, the non-sensitive feature by X, and the decision by Y. We consider a general networked causal diagram as shown in Fig. 1e, where, for simplicity, we posit that interference only exists from S to X. However, our methods can be applied to deal with the interference between any pair of variables. Suppose that we are given a dataset $\mathcal{D} = \{s^{(i)}, x^{(i)}, y^{(i)}\}_{i=1}^{K}$ and a network $\mathcal{N}$ that connects individuals in $\mathcal{D}$ to reflect interference. The goal is to build a classifier $h : X \mapsto Y$ for predicting the label. We say that the classifier is causally fair if $\mathbb{E}[h(x)|do(s^+)] = \mathbb{E}[h(x)|do(s^-)]$, where $do(\cdot)$ represents the intervention that is conducted under the networked causal diagram.

5 Method

5.1 Causal Inference on Network Data

We use the networked causal diagram in Fig. 1e and the task of inferring the causal effect from S to X (i.e., computing $P(x|do(s))$) as a running example for elaborating our method. The *do*-calculus is an axiomatic system that is widely used for solving the causal inference problem ((Pearl, 2009), Ch. 3.4). However, as shown in recent studies (Zhang et al., 2022; Zhang, 2023), directly applying *do*-calculus in non-IID settings can lead to biased results. The fundamental challenge is that, in networked data, the structural equation associated with each node depends on its local neighborhood. Specifically, unlike standard structural equations in classical SCMs, Eq. 1 incorporates neighborhood information, implying that nodes with different local network structures follow different effective causal mechanisms. As a result, the invariance assumptions required for the validity of *do*-calculus no longer hold. This section presents a theoretical study aimed at extending the intervention *do* operator and symbolic notations to NSCM, starting by identifying conditions under which a shared causal mechanism can be recovered despite the presence of neighborhood-dependent interactions.

Let $\mathbf{s}^{(i)} := \{\, s^{(i)}, \{s^{(j)} : j \in \mathsf{ne}^{1:k}(i)\}\,\}$ denote the multiset of sensitive attributes associated with node i and its k-hop neighbors, capturing all neighborhood information of S that influences node i, including node i itself. We define the intervention $do(\mathbf{s}^{(i)} = s)$ as setting the sensitive attribute of every node in $\mathbf{s}^{(i)}$ to s. Let $do(\mathbf{s} = s)$ denote the global intervention that applies $do(\mathbf{s}^{(i)} = s)$ simultaneously for all nodes i. Under this global intervention, the interventional distribution $P(x|do(\mathbf{s}) = s)$ coincides with $P(x|do(s))$ in the graph setting, which is the target quantity we aim to compute. The following discussions focus on the computation of $P(x|do(\mathbf{s}) = s)$.

The key challenge in computing $P(x \mid do(\mathbf{s} = s))$ arises from the heterogeneity of node-level causal mechanisms induced by varying local network structures. To address this, we aim to decouple the causal effect of S from the structural variability introduced by neighborhood interactions. To this end, we leverage the notion of node colors from the Weisfeiler-Lehman (WL) graph isomorphism test to encode local structural information. The node color serves as a representation of the local computation tree of each node. In particular, two nodes are assigned the same color if and only if they have identical computation trees under the WL procedure. Let $c^{(i)}$ denote the color of node i obtained from the WL algorithm with identical initial node colors. As shown in (Jegelka, 2022), we have the following result:

Lemma 1 ((Jegelka, 2022)). *For two different nodes i, j, $c^{(i)} = c^{(j)}$ if and only if nodes i and j have identical computation trees in the WL graph isomorphism test.*

Based on this representation, we introduce conditions under which the heterogeneous node-level mechanisms can be reduced to a shared functional form.

Condition 1 (Decomposability)**.** The structural equation in the NSCM can be decomposed into a message-passing mechanism that aggregates neighborhood information and an internal mechanism that governs the node-level causal effect.

Under this condition, the structural equation $x^{(i)} = f(s^{(i)}, \{s^{(j)} : j \in \mathsf{ne}^{1:k}(i)\}, u_X^{(i)})$ can be decomposed into two equations:

$$a^{(i)} = f^{MP}(\mathsf{s}^{(i)}, c^{(i)}), \qquad x^{(i)} = f^{INT}(a^{(i)}, u_X^{(i)})$$

where function f^{MP} represents the message-passing mechanism, $a^{(i)}$ is an intermediate variable summarizing neighborhood effects, and function f^{INT} represents the internal node-level causal mechanism.

Combining Lemma 1 and Condition 1, we obtain the following result, which characterizes when the aggregated neighborhood effect can be represented by a shared mapping.

Proposition 2. *For any two nodes i, j, if $\{\mathsf{s}^{(i)}, c^{(i)}\} = \{\mathsf{s}^{(j)}, c^{(j)}\}$, then we have $a^{(i)} = a^{(j)}$.*

Proof. By Lemma 1, nodes i and j have identical computation trees. Under Condition 1, the message-passing function f^{MP} operates over these computation trees. Therefore, identical inputs and identical structural contexts imply identical outputs, i.e., $a^{(i)} = a^{(j)}$. □

Condition 2 (Graph Independence)**.** Exogenous variable U_X is independent of node color C, i.e., $U_X \perp C$.

The Graph Independence condition ensures that the exogenous variation affecting X does not depend on the network structure. Consequently, the intermediate variable A provides a sufficient summary of all neighborhood-dependent effects relevant for determining X under intervention.

Building on the above results, we now show that, under Conditions 1 and 2, the NSCM can be reduced to an equivalent causal model that admits standard causal inference via do-calculus. Under Condition 1, the neighborhood-dependent structural equation can be rewritten by dropping the node index as

$$a = g(\mathsf{s}, c), \qquad x = f^{INT}(a, u_X),$$

where c denotes the node color encoding local network structure, and g is a deterministic mapping induced by the message-passing mechanism f^{MP}. This reduction is critical as it transforms the original neighborhood-dependent mechanism into a two-stage process in which all structural variability arising from the network is captured by (S, C) through A, while the downstream mechanism f^{INT} is shared across nodes. In other words, Condition 1 enables the recovery of a shared functional form for the causal mechanism of X.

Given this representation, the resulting causal model consists of variables (S, C, A, X) with structural equations

$$c \sim P(C), \qquad \mathsf{s} \sim P(\mathsf{S} \mid C), \qquad a = g(\mathsf{s}, c), \qquad x = f^{INT}(a, u_X),$$

where the dependence between S and C allows for potential confounding induced by the network structure. These structural equations correspond to the reduced causal diagram as shown in Fig. 1f, which is a standard causal diagram without explicit interference between instances.

Condition 2 plays a critical role in ensuring identifiability as the assumption $U_X \perp C$ guarantees that all graph-dependent effects on X are mediated through A, preventing additional confounding paths from C to X via the exogenous variable. In fact, note that in the reduced causal diagram, the only backdoor path from S to X is $\mathsf{S} \leftrightarrow C \rightarrow A \rightarrow X$, which arises due to the dependence between S and the structural variable C. This backdoor path can be blocked by conditioning on C, making C a valid adjustment set.

We then examine the case where Condition 2 is violated. In this setting, the dependence between U_X and C induces latent confounding between C and X, which can be represented as a bidirected edge $C \leftrightarrow X$ in the causal diagram. To analyze identifiability, we consider the latent projection of the model onto the observed variables $\{\mathsf{S}, C, X\}$ by marginalizing out the latent variables A and U_X (see Fig. 1g). In the resulting graph, the directed edges reduce to $\mathsf{S} \rightarrow X$ and $C \rightarrow X$, while latent confounding induces bidirected edges $\mathsf{S} \leftrightarrow C$ and $C \leftrightarrow X$. Consequently, the variables S, C, and X belong to a single c-component. Under this structure, the Hedge Criterion (Shpitser & Pearl, 2008) can be applied to assess identifiability. In particular, consider the two c-forests: (i) the larger forest $F' = \{\mathsf{S}, C, X\}$, which is rooted at X and contains the treatment variable S, and (ii) the smaller forest $F = \{C, X\}$, which is also rooted at X but excludes S. Since $F \subset F'$ and both forests share the same root, the hedge criterion is satisfied. By the completeness of the hedge

criterion, the existence of such a structure implies that the causal effect $P(x \mid do(\mathsf{s}) = s)$ is not identifiable from observational data in general.

As a result, we have the following proposition.

Proposition 3. *Under Conditions 1 and 2, the causal effect $P(x \mid do(\mathsf{s}) = s)$ is identifiable via adjustment on C, provided that C blocks all backdoor paths from S to X and $P(\mathsf{S} = s \mid C = c) > 0$ for all c with $P(C = c) > 0$.*

5.2 From Identification to Estimation

While Section 5.1 establishes that the causal effect $P(x \mid do(\mathsf{s}) = s)$ is identifiable via adjustment on the structural variable C, directly estimating the conditional distribution $P(x \mid \mathsf{s}, c)$ for the adjustment remains challenging in practice. In particular, the node color C encodes fine-grained local graph structure and may take a large number of distinct values, leading to high-dimensional and sparsely supported conditioning.

To address this issue, we leverage the decomposable structure implied by Condition 1. Specifically, under this condition, the effect of (S, C) on X is mediated through an intermediate variable $A = g(\mathsf{S}, C)$, which provides a compact representation of neighborhood influence. This suggests replacing direct conditioning on (S, C) with a learned representation A that captures the relevant structural information for predicting X.

Motivated by this observation, we derive a representation of the interventional distribution that separates the contribution of structural variability from the downstream causal mechanism. This formulation not only provides a principled route for estimating $P(x \mid do(\mathsf{s}) = s)$, but also directly informs the design of our model architecture, as described below.

Theorem 4. *Suppose Conditions 1 and 2 hold. Then the interventional distribution of X under $do(\mathsf{s}) = s$ admits the following representation:*

$$P(x \mid do(\mathsf{s}) = s) = \sum_{c} P(c) \sum_{a} P(x \mid a)\, P(g(\mathsf{s}, c) = a)\,, \tag{2}$$

where $A = g(\mathsf{S}, C)$ denotes the intermediate variable induced by the message-passing mechanism.

Theorem 4 provides a representation of the interventional distribution through the intermediate variable $A = g(\mathsf{S}, C)$. To motivate estimation, we next relate this expression to the observational distribution.

By marginalizing over S, C, and A, the observational distribution of X can be written as

$$P(x) = \sum_{\mathsf{s},c,a} P(\mathsf{s}, c) P(g(\mathsf{s}, c) = a) P(x \mid a) = \sum_{\mathsf{s},c} P(\mathsf{s}, c) \sum_{a} P(x \mid a) P(g(\mathsf{s}, c) = a). \tag{3}$$

Comparing Eq. 2 from Theorem 4 with Eq. 3, we observe that both the interventional and observational distributions share the same inner quantity $Q(x; \mathsf{s}, c) := \sum_a P(x \mid a)\, P(g(\mathsf{s}, c) = a)$. The key implication here is that the difference between the observational and interventional distributions lies only in how the shared component $Q(x; \mathsf{s}, c)$ is averaged. In the observational setting, $Q(x; \mathsf{s}, c)$ is averaged with respect to the joint distribution $P(\mathsf{s}, c)$. Under intervention, the value of s is fixed and the averaging is performed only over the structural variable C according to $P(c)$.

This observation admits a natural Monte Carlo interpretation. Suppose that a model is learned to approximate the shared component $Q(x; \mathsf{s}, c)$. Then the observational distribution $P(x)$ can be estimated by sampling pairs $(\mathsf{s}, c) \sim P(\mathsf{s}, c)$ and averaging the resulting values of $Q(x; \mathsf{s}, c)$, whereas the interventional distribution $P(x \mid do(\mathsf{s}) = s)$ can be estimated by fixing s, sampling $c \sim P(c)$, and averaging the same quantity $Q(x; \mathsf{s}, c)$. Therefore, once the shared component is learned, intervention amounts to a reweighting or resampling operation rather than learning a new model.

Based on this observation, we state the following estimation principle.

Corollary 5. *Suppose a model $Q_\theta(x; \mathsf{s}, c)$ is learned from observational data to approximate the shared component $\sum_a P(x \mid a)\, P(g(\mathsf{s}, c) = a)$ in Eq. 3. Assume further that the learned mapping remains valid under*

interventions on S *in the sense that the intervention changes only the weighting over* (s, c)*, while preserving the structural mechanism encoded by* Q_θ*. Then the interventional distribution can be approximated by*

$$P(x \mid do(\mathsf{s}) = s) \approx \sum_c P(c)\, Q_\theta(x; \mathsf{s}, c) = \mathbb{E}_{c \sim P(c)}[Q_\theta(x; \mathsf{s}, c)]. \tag{4}$$

Corollary 5 does not require explicitly recovering the true latent values of A. Instead, it suggests learning a representation that captures the shared functional component relating (S, C) to X, and then reusing this representation under intervention through a modified averaging scheme. This provides a practical route for estimating interventional distributions while remaining consistent with the decomposable structure in Condition 1.

Motivated by this principle, we design the Message Passing Variational Autoencoder (MPVA) for causal inference in networks, which combines a message-passing neural network (MPNN) with a conditional variational autoencoder (cVAE), as detailed in the next subsection.

5.3 Message Passing Variational Autoencoder for Causal Inference (MPVA)

We develop the MPVA framework to directly implement the estimation principle established in Theorem 4. The architecture is designed to approximate the shared component $Q(x; \mathsf{s}, c)$ between observational and interventional distributions, which then enables computing interventional distributions through reweighting. For generalization, we further consider the existence of independent variables Z other than S that directly affect X, and Corollary 5 readily applies to this case. In this framework, we first use an MPNN to learn the intermediate representation A from Condition 1, which captures the aggregated causal effect of S on X from each node's neighborhood through the message-passing mechanism $g(\mathsf{s}, c)$, denoted as $\hat{a} = \hat{g}^{MPNN}(\mathsf{s})$. Then, we use the estimated intermediate representation $\hat{A}$ along with the variables Z as inputs to a multilayer perceptron (MLP) to predict the node feature X, denoted as $\hat{x} = MLP(\hat{a}, z)$. After that, we use a cVAE for learning the conditional distribution $P(x|\hat{a}, z)$, which corresponds to $P(x|a)$ in Theorem 4, where the encoder $v = EN(x, \hat{a}, z)$ takes x as the input with $\hat{a}, z$ as conditions, and the decoder $\hat{x} = DE(v, \hat{a}, z)$ attempts to reconstruct x. The architecture of MPVA is illustrated in Fig. 2.

In the training phase, we first train MPNN and MLP with the dataset $\mathcal{D}$ and network $\mathcal{N}$. Note that in this process, we do not require actual values of A as supervised signals. Then, we freeze the parameters of MPNN and MLP and train the cVAE. In the inference phase, for each node i, we first use the MPNN to compute $\hat{a}^{(i)} = \hat{g}^{MPNN}(\mathsf{s}^{(i)})$ and use the encoder of the cVAE to compute $v^{(i)} = EN(x^{(i)}, \hat{a}^{(i)}, z^{(i)})$. Then, we perform the intervention $do(\mathsf{s}) = s$ to change the value of S to s for all nodes. After that, we use the MPNN to compute the value of A again under the intervention, i.e., $\tilde{a}_s = \hat{g}^{MPNN}(do(\mathsf{s}^{(i)}) = s)$, and use the cVAE to reconstruct X using $\tilde{a}_s$ and $v^{(i)}$, i.e., $\tilde{x}_s^{(i)} = DE(v^{(i)}, \tilde{a}_s, z^{(i)})$. As a result, $\tilde{x}_s^{(i)}$ represents the estimated outcome for node i under the population-level intervention $do(\mathsf{s}) = s$. This inference process follows an Abduction-Action-Prediction structure: we first encode the observed outcome (abduction), then compute the effect of intervention on the intermediate representation (action), and finally decode to obtain the interventional outcome (prediction). Critically, the identifiability of these interventional distributions is guaranteed by Conditions 1 and 2, as established in Section 5.1.

5.4 Causally Fair Node Classification

Having established a method for estimating interventional distributions, we now apply it to achieve causally fair node classification. We conceptualize causal fair node classification as a regularized optimization problem that explicitly minimizes the causal effect of sensitive attributes on predictions. To this end, we utilize the interventional distributions estimated by MPVA to develop a causal fairness regularization term, which is appended to the traditional classification loss. Specifically, we generate the interventional outcomes using the classifier h and the learned MPVA, denoted as $\tilde{y}_{s^+}$ and $\tilde{y}_{s^-}$, where $\tilde{y}_s = h(\tilde{x})|do(s)$. To address fairness in the classification task, we construct a regularization term aimed at minimizing the causal discrepancy between

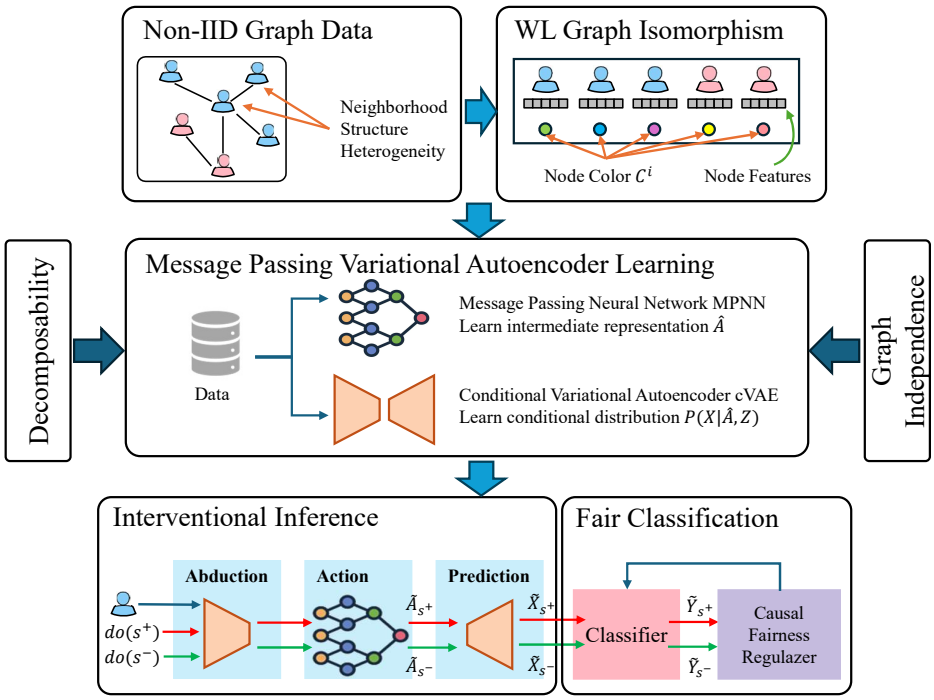

Figure 2: Overview of the MPVA framework. The MPNN component learns node-level representations by aggregating causal influences from neighboring nodes. The cVAE models the conditional distribution necessary to estimate interventional outcomes. Together, the MPNN and cVAE enable inference of the interventional distribution of Y and act as a regularizer to improve causally fair classification.

two interventional variants. This term is defined as follows:

$$\begin{aligned}\ell_f &= \mathbb{E}[h(\tilde{x})|do(s^+)] - \mathbb{E}[h(\tilde{x})|do(s^-)] \\ &= \mathbb{E}[\mathbb{1}_{\tilde{y}_{s^+}=1}] - \mathbb{E}[\mathbb{1}_{\tilde{y}_{s^-}=1}] = \mathbb{E}[\mathbb{1}_{\tilde{y}_{s^+}=1}] + \mathbb{E}[\mathbb{1}_{\tilde{y}_{s^-}=-1}] - 1,\end{aligned} \tag{5}$$

where $\mathbb{1}$ is the indicator function. Following (Wu et al., 2019), the indicator function can be further replaced with the differentiable surrogate function $u(\cdot)$. It is noteworthy that this differentiability ensures that the regularization term can be effectively incorporated into the classic loss functions used for training the node classifier. To sum up, this regularization term can be seamlessly incorporated with the classification loss: $\ell = \frac{1}{n}\Sigma_{i\in[K]}\ell_c(h(x^{(i)}), y^{(i)}) + \lambda\ell_f$, where ℓ_c is the empirical loss function and λ is a hyper-parameter to balance model performance and causal fairness.

6 Experiment

We evaluate the proposed method and comparisons on both semi-synthetic and real-world graphs. To ensure reproducibility of our work, we provide comprehensive implementation details and experimental setup information in the appendix. The complete source code for MPVA, including the Message Passing Variational Autoencoder architecture and training procedures, is available at an anonymous repository[2]. All hyperparameter settings, network architectures, and optimization details are specified in the appendix. All baseline implementations follow their original papers with publicly available code, and we report means and standard deviations across five independent runs to ensure statistical reliability.

[2] https://1drv.ms/u/c/c09afa5d46da1993/IQDTMFsPrOzfRJTA-xjZiBvNAVIRWo8FafDuiX1Go46eAxs?e=C7Qkyi

6.1 Datasets

In our experiments, we adopt both semi-synthetic datasets, which allow full control over the data-generation process, and real-world datasets, which evaluate the external generalizability of our methods. Semi-synthetic data are commonly used in the causal inference and fair machine learning fields since the ground truth, e.g., causal effects and bias, cannot be directly observed in real-world settings. To create semi-synthetic datasets, we adapt the Credit Dataset (Yeh, 2016) by introducing network structures and causal relationships using the Network Structural Causal Model. This approach allows us to precisely control data generation and accurately derive ground-truth interventional distributions for arbitrary interventions on the sensitive attribute. Specifically, we generate two semi-synthetic datasets, denoted as D1 and D2, for evaluation purposes. The synthetic data generation process following the Network Structural Causal Model is fully described in the appendix. We also conduct experiments on widely used real-world datasets, namely Credit Defaulter (Yeh, 2016) and German (Hofmann, 1994). For real-world datasets (Credit and German), we provide detailed preprocessing steps and train/validation/test splits in the supplementary code repository. Since the underlying mechanisms for the real-world dataset are unknown, we evaluate our proposed framework's ability to estimate non-IID causal interventions and compare it against baseline methods, including IID-based causal fairness approaches. We use the learned MPVA model to measure the interventional quantities and evaluate the performance in terms of non-IID causal fairness. The detailed statistics of these datasets are included in the appendix, including the number of nodes, the number of edges, and the dimension of features.

6.2 Experiment Settings

Fairness Metrics: We evaluate the performance of the proposed framework in terms of prediction accuracy and fairness. For non-causal fairness notions, we use demographic parity, which is a widely used fairness notion in the fairness-aware learning field, to evaluate the fairness performance at the group level. Demographic parity requires the decision made by the classifier to be independent of the sensitive attribute. Usually, it is quantified with regard to *risk difference* (**RD**), i.e., the difference in the positive predictions between the sensitive group and the non-sensitive group. It can be expressed as $\left|\mathbb{E}_{X|S=s^+}[\hat{Y}] - \mathbb{E}_{X|S=s^-}[\hat{Y}]\right|$. For causal fairness notions, we consider both the IID and the non-IID (i.e., graph-based) causal fairness notions. We denote the IID causal fairness as **CF** whose calculation and estimation approaches are described in the appendix. On the other hand, we denote the graph-based causal fairness notion as **gCF**, which is described in Eq. 5.

Mitigation Baselines: We compare the proposed framework **MPVA** with several state-of-the-art non-IID bias mitigation methods and the conventional IID constraint-based methods. **GCN-RD** and **GCN-IID** are the conventional Graph Convolutional Networks (GCN) with the risk difference and the causal fairness constraints developed for the IID data. The constraint formulation is described in the appendix. **FairGNN** (Dai & Wang, 2021) employs a covariance-based adversarial discriminator to predict the sensitive attribute into the conventional GNN node classifier. **GEAR** (Ma et al., 2022) utilizes a variational auto-encoder to synthesize counterfactual samples to achieve counterfactual fairness for node classification. **NIFTY** (Agarwal et al., 2021) enhances fairness and stability in GNNs by introducing a novel objective function and layer-wise weight normalization based on the Lipschitz constant. **FairINV** (Zhu et al., 2024a) trains fair GNNs by eliminating spurious correlations between labels and sensitive attributes within a single training session. For the implementation details, please refer to Sec. C.2 in the appendix.

Table 1: Fairness measurement of conventional GCN using various metrics on semi-synthetic datasets.

Data	Acc	RD	CF	gCF	True-gCF
Semi-synthetic D1	$0.9674_{\pm 0.0017}$	$0.0580_{\pm 0.0015}$	$0.0319_{\pm 0.0001}$	$0.1792_{\pm 0.0344}$	$0.1960_{\pm 0.0338}$
Semi-synthetic D2	$0.9715_{\pm 0.0011}$	$0.0832_{\pm 0.0013}$	$0.0461_{\pm 0.0003}$	$0.6721_{\pm 0.0132}$	$0.6884_{\pm 0.0084}$

Table 2: Evaluation of mitigation methods on semi-synthetic datasets.

Data	Metric	Own Metric	Acc	gCF	True-gCF
Semi-synthetic D1	MPVA	-	$0.9259_{\pm 0.0051}$	$\mathbf{0.0023_{\pm 0.0012}}$	$\mathbf{0.0010_{\pm 0.0006}}$
	GCN-RD	$0.0074_{\pm 0.0074}$	$0.8743_{\pm 0.0179}$	$0.4150_{\pm 0.0961}$	$0.4219_{\pm 0.0969}$
	GCN-IID	$0.0031_{\pm 0.0015}$	$0.9024_{\pm 0.0071}$	$0.1787_{\pm 0.0437}$	$0.1896_{\pm 0.0456}$
Semi-synthetic D2	MPVA	-	$0.9490_{\pm 0.0009}$	$\mathbf{0.0019_{\pm 0.0018}}$	$\mathbf{0.0073_{\pm 0.0048}}$
	GCN-RD	$0.0091_{\pm 0.0055}$	$0.8828_{\pm 0.0026}$	$0.1634_{\pm 0.0744}$	$0.1868_{\pm 0.0900}$
	GCN-IID	$0.0069_{\pm 0.0010}$	$0.9094_{\pm 0.0021}$	$0.6818_{\pm 0.0310}$	$0.6977_{\pm 0.0292}$

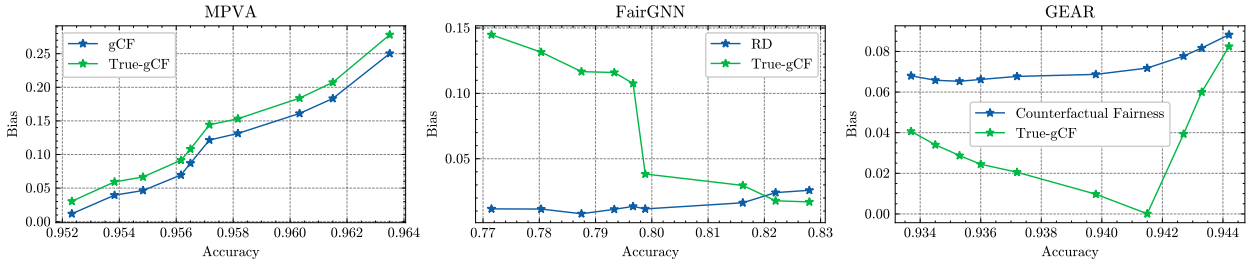

Figure 3: Comparison of measured bias and true gCF bias on D2 with various mitigation methods.

6.3 Results on Semi-synthetic Data

We first generate two network structures with different generating parameters for the semi-synthetic datasets. To show the biases contained in the generated graphs in terms of the proposed metrics, as well as the effectiveness of MPVA for estimating gCF, we train the classic GCN models without any bias mitigation considerations. Given the conventional GCN node classification models, we measure the bias and report the results as shown in Tab. 1, which present the empirical node prediction accuracy, the estimated RD, CF, gCF (highlighted in blue), as well as the ground truth gCF (highlighted in green) that is directly computed by performing interventions on the true causal model. As we can see, the node prediction accuracy is high, meaning the models are well-trained and able to make accurate predictions. Comparing our estimated gCF (in blue) and the ground truth of gCF (in green), we see that our method can accurately estimate the causal fairness in the graph data. We also observe that RD and CF differ substantially from gCF, indicating that one cannot simply use RD and CF to estimate gCF.

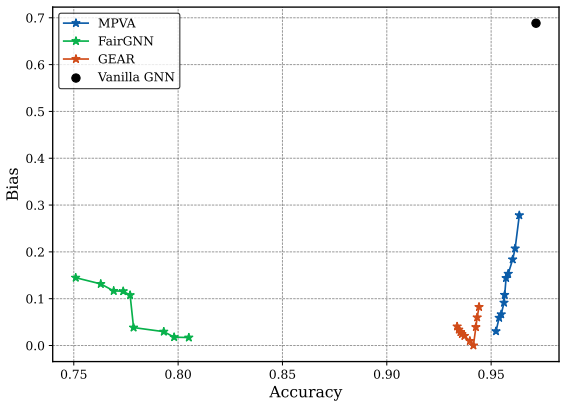

Figure 4: Trade-off between fairness and accuracy for MPVA, FairGNN, and GEAR on the semi-synthetic dataset D2.

Next, we build fair node classification models on the generated graph data using the proposed method and baselines. The performance of classification prediction and fairness is shown in Tab. 2. For a fair comparison, all the models are trained to be fair based on the fairness metrics used by their own (i.e., GCN-RD uses RD and GCN-IID uses CF as their fairness metrics). As illustrated in the **Own Metric** column, all the models are well-trained and fair (for MPVA, its own metric is shown in the gCF column). Then, we present the accuracy,

estimated gCF, and the ground truth gCF of all methods. As can be seen, although the baselines, including GCN-RD and GCN-IID, are considered fair based on their own fair metrics, they exhibit significant bias from *graph-based* causal fairness (i.e., gCF) perspective. In addition, the baseline methods neglect the potential effect of the graph structure while attempting to address the bias, resulting in a compromise of accuracy. On the other hand, our proposed MPVA achieves the best performance in terms of both accuracy and graph-based causal fairness gCF. We further compare our proposed MPVA with the state-of-the-art graph-based bias mitigation algorithms, FairGNN and GEAR. FairGNN aims to mitigate statistical bias in graph data, while GEAR aims to alleviate counterfactual bias in graph data. For a comprehensive comparison, we compare the trade-off between model bias and performance for MPVA, FairGNN, and GEAR. As shown in Fig. 3, we tune each model multiple times to obtain various bias-accuracy trade-offs and plot the corresponding fairness/bias measurement used by the models and the true gCF derived from the data generation process at various accuracy levels in each subplot. For our method MPVA, the estimated fairness aligns with the true causal fairness gCF at every accuracy level, thanks to our graph causal inference technique. However, for FairGNN, and GEAR, the measured fairness is significantly different from the true fairness, implying that they cannot guarantee to obtain a fair model by fine-tuning models to balance the bias-accuracy trade-off. In addition, we compare the trade-off between fairness and accuracy for each method on the dataset D2, as shown in Fig. 4. The figure demonstrates that MPVA achieves the most favorable trade-off, attaining high accuracy while maintaining low gCF bias. In contrast, FairGNN and GEAR struggle to simultaneously achieve both high fairness and accuracy, further validating the effectiveness of our causal framework in balancing both objectives.

6.4 Sensitivity Analysis

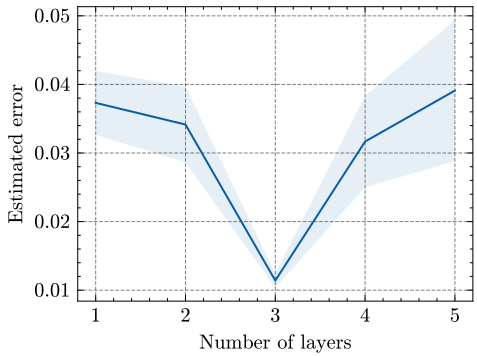

Figure 5: The impact of the number of MPNN layers against estimated error.

Table 3: gCF Difference under different dependency strengths.

Violation Strength	gCF Difference
0.0	$0.0075_{\pm 0.0007}$
0.5	$0.0592_{\pm 0.0012}$
1.0	$0.1142_{\pm 0.0016}$

6.4.1 Multi-hop Causal Effect Analysis

We further demonstrate the proposed MPVA framework is capable of capturing multiple-hop causal effects in graph data. We generate the influence of neighbors' sensitive attributes of a certain node within the range of three-hop, based on the dataset D2. As shown in Fig. 5, when the MPNN module has the same number of layers as the neighborhood hops of the generating model (which is 3 in Fig. 5), it can achieve the best performance in estimating the interventional distribution. This occurs because using too few layers leads to underfitting, while using too many layers results in overfitting. We also observe that the variance of the results is minimal when the layer number matches the neighborhood hops. This observation guides the selection of an appropriate number of layers in practice.

6.4.2 Violation of Proposed Conditions

Additionally, we simulate a scenario where the proposed conditions are violated, based on the synthetic dataset D2. To facilitate the analysis, we introduce a dependency between U_X and A by incorporating a weighted multiplicative interaction (e.g., $c \cdot A \cdot U_X$) during the data generation process, where a coefficient (c) is introduced to control the violation or dependency strength. Then, we measure the absolute difference

between our estimated gCF and the True-gCF. As shown in Tab. 3, although both methods minimize their respective fairness metrics, they fail to achieve graph-based causal fairness, as reflected in the gCF and True-gCF columns, which are consistent with the observations reported in our paper.

Table 4: Results of various methods on real-world datasets.

Dataset	Method	Acc	RD	CF	gCF
Credit	GCN	$0.8192_{\pm 0.0005}$	$0.0195_{\pm 0.0014}$	$0.0049_{\pm 0.0001}$	$0.0705_{\pm 0.0123}$
	GCN-RD	$0.7988_{\pm 0.0062}$	$0.0057_{\pm 0.0044}$	$0.0055_{\pm 0.0001}$	$0.0540_{\pm 0.0145}$
	FairGNN	$0.7930_{\pm 0.0086}$	$0.0047_{\pm 0.0012}$	$0.0043_{\pm 0.0008}$	$0.0404_{\pm 0.0251}$
	GCN-IID	$0.8065_{\pm 0.0008}$	$0.0100_{\pm 0.0023}$	$0.0010_{\pm 0.0003}$	$0.1360_{\pm 0.0833}$
	NIFTY	$0.7933_{\pm 0.0146}$	$0.0543_{\pm 0.0068}$	$0.0079_{\pm 0.0006}$	$0.0981_{\pm 0.0165}$
	GEAR	$\mathbf{0.8075_{\pm 0.0005}}$	$0.0260_{\pm 0.0108}$	$0.0055_{\pm 0.0003}$	$0.0278_{\pm 0.0111}$
	FairINV	$0.7720_{\pm 0.0205}$	$0.0111_{\pm 0.0139}$	$0.0087_{\pm 0.0003}$	$0.0295_{\pm 0.0226}$
	MPVA	$0.8054_{\pm 0.0033}$	$0.0142_{\pm 0.0046}$	$0.0075_{\pm 0.0007}$	$\mathbf{0.0036_{\pm 0.0033}}$
German	GCN	$0.9758_{\pm 0.0027}$	$0.0771_{\pm 0.0083}$	$0.0704_{\pm 0.0004}$	$0.6080_{\pm 0.2596}$
	GCN-RD	$0.9608_{\pm 0.0054}$	$0.0046_{\pm 0.0023}$	$0.0354_{\pm 0.0005}$	$0.2657_{\pm 0.0600}$
	FairGNN	$0.8183_{\pm 0.0162}$	$0.0051_{\pm 0.0038}$	$0.0536_{\pm 0.0012}$	$0.8263_{\pm 0.0526}$
	GCN-IID	$0.7643_{\pm 0.0118}$	$0.1719_{\pm 0.0133}$	$0.0047_{\pm 0.0011}$	$0.2994_{\pm 0.0317}$
	NIFTY	$0.8113_{\pm 0.0224}$	$0.0975_{\pm 0.0347}$	$0.0566_{\pm 0.0011}$	$0.9987_{\pm 0.0019}$
	GEAR	$\mathbf{0.9717_{\pm 0.0187}}$	$0.0689_{\pm 0.0275}$	$0.0359_{\pm 0.0040}$	$0.3667_{\pm 0.3769}$
	FairINV	$0.8200_{\pm 0.0433}$	$0.0428_{\pm 0.0508}$	$0.0607_{\pm 0.0063}$	$0.5844_{\pm 0.3669}$
	MPVA	$0.9283_{\pm 0.0353}$	$0.1136_{\pm 0.0332}$	$0.0667_{\pm 0.0014}$	$\mathbf{0.0030_{\pm 0.0032}}$

6.5 Results on Real Data

We further conduct extensive experiments on real-world data. We first train a naive Graph Convolutional Network (GCN) without any bias mitigation methods, run fairness-aware methods, and repeat five independent experiments. The results are shown in Tab. 4. As can be seen, in the original GCN, there is a big gap between gCF and CF, which implies that the IID causal metric is not accurate for measuring non-IID causal fairness in the graph. In both datasets, MPVA **outperforms** all other baselines in terms of gCF with a mild accuracy decrease compared with the classic GCN model. Although other methods achieve fairness regarding their own metrics, they fail to meet the gCF requirements. FairGNN neglects the causality-based bias, which results in compromising accuracy in order to achieve fairness. The baseline GEAR achieves good accuracy performance. However, it fails to eliminate the bias effectively. For example, in the German dataset, GEAR is unable to completely remove the significant bias present. The results show that existing fairness methods cannot guarantee fairness for gCF. In summary, the results are consistent with those in the semi-synthetic datasets, demonstrating the superiority of the proposed method.

7 Conclusions

This paper has addressed a fundamental challenge in extending causal fairness to graph data that the heterogeneity of causal mechanisms across nodes due to varying neighborhood structures, which violates the invariance assumptions underlying classical structural causal models. We focused on graph settings where data instances are interconnected and proposed a principled solution based on the Network Structural Causal Model (NSCM) framework. Our key insight is that, although node-level mechanisms vary with local network structure, it is feasible to construct a structural representation that restores invariance. We introduced two conditions, Decomposability and Graph Independence, that formalize when interventional distributions can be identified using *do*-calculus in non-IID settings by separating the effect of sensitive attributes from network structural variability. Building on this theoretical foundation, we developed the Message Passing Variational Autoencoder for Causal Inference (MPVA), which estimates the shared functional component between observational and interventional distributions. We further integrated MPVA with a causal fairness regularization framework that explicitly minimizes the causal effect of sensitive attributes on predictions through interventional distribution estimation, enabling causally fair node classification in non-IID

graph settings. Empirical evaluations on semi-synthetic and real datasets have shown that our approach surpasses baseline methods by more accurately approximating interventional distributions and reducing bias. Additionally, sensitivity analysis demonstrates that MPVA can effectively capture multi-hop causal effects and maintain performance even under varying conditions.

8 Limitations and Future Work

While our proposed method has shown promising results, several limitations warrant discussion, along with potential directions for future work. Our method assumes that the causal graph structure is known a priori. In practice, the true causal graph is often unknown and must be inferred from data. Future work could integrate causal discovery algorithms to automatically learn the underlying causal structure, or develop robust methods that are less sensitive to misspecification of the causal graph. In addition, the current framework is specifically designed for graph-structured data where dependencies between instances are explicitly modeled through edges. Extending this approach to other settings, such as tabular data with latent dependencies or temporal data with sequential dependencies, represents an important direction. This could involve developing analogous principles for decomposability and independence in these alternative settings. Our current formulation primarily focuses on scenarios with a single sensitive attribute. Real-world fairness problems often involve multiple intersecting sensitive attributes (e.g., race and gender). Extending the framework to handle multiple sensitive attributes by leveraging the idea of the intersectional fairness (Foulds et al., 2020) is an important avenue for future research.

References

Chirag Agarwal, Himabindu Lakkaraju, and Marinka Zitnik. Towards a unified framework for fair and stable graph representation learning. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July 2021*, Proceedings of Machine Learning Research, pp. 2114–2124. AUAI Press, 2021. URL https://proceedings.mlr.press/v161/agarwal21b.html.

Chirag Agarwal, Marinka Zitnik, and Himabindu Lakkaraju. Probing GNN explainers: A rigorous theoretical and empirical analysis of GNN explanation methods. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, Proceedings of Machine Learning Research, pp. 8969–8996. PMLR, 2022. URL https://proceedings.mlr.press/v151/agarwal22b.html.

David T. Arbour, Dan Garant, and David D. Jensen. Inferring network effects from observational data. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 715–724. ACM, 2016a. doi: 10.1145/2939672.2939791.

David T. Arbour, Katerina Marazopoulou, and David D. Jensen. Inferring causal direction from relational data. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, UAI 2016, June 25-29, 2016, New York City, NY, USA*. AUAI Press, 2016b. URL http://auai.org/uai2016/proceedings/papers/217.pdf.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. *FAT ML Workshop*, 2017. URL http://arxiv.org/abs/1707.00075.

Rohit Bhattacharya, Daniel Malinsky, and Ilya Shpitser. Causal inference under interference and network uncertainty. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, Proceedings of Machine Learning Research, pp. 1028–1038. AUAI Press, 2019. URL http://proceedings.mlr.press/v115/bhattacharya20a.html.

Reuben Binns. On the apparent conflict between individual and group fairness. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pp. 514–524. ACM, 2020. doi: 10.1145/3351095.3372864.

Avishek Joey Bose and William L. Hamilton. Compositional fairness constraints for graph embeddings. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Proceedings of Machine Learning Research, pp. 715–724. PMLR, 2019. URL http://proceedings.mlr.press/v97/bose19a.html.

Maarten Buyl and Tijl De Bie. DeBayes: A bayesian method for debiasing network embeddings. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, pp. 1220–1229. PMLR, 2020. URL http://proceedings.mlr.press/v119/buyl20a.html.

Simon Caton and Christian Haas. Fairness in Machine Learning: A Survey, October 2020. URL http://arxiv.org/abs/2010.04053.

Silvia Chiappa. Path-specific counterfactual fairness. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, the Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, the Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 7801–7808. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.33017801.

Enyan Dai and Suhang Wang. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *WSDM '21, the Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*, pp. 680–688. ACM, 2021. doi: 10.1145/3437963.3441752.

Yushun Dong, Jian Kang, Hanghang Tong, and Jundong Li. Individual fairness for graph neural networks: A ranking based approach. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pp. 300–310. ACM, 2021. doi: 10.1145/3447548.3467266.

Yushun Dong, Ninghao Liu, Brian Jalaian, and Jundong Li. EDITS: Modeling and mitigating data bias for graph neural networks. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pp. 1259–1269. ACM, 2022. doi: 10.1145/3485447.3512173.

Wei Fan, Kunpeng Liu, Rui Xie, Hao Liu, Hui Xiong, and Yanjie Fu. Fair graph auto-encoder for unbiased graph representations with wasserstein distance. In *IEEE International Conference on Data Mining, ICDM 2021, Auckland, New Zealand, December 7-10, 2021*, pp. 1054–1059. IEEE, 2021. doi: 10.1109/ICDM51629.2021.00122.

Zahra Fatemi and Elena Zheleva. Minimizing interference and selection bias in network experiment design. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pp. 176–186. AAAI Press, 2020. URL https://ojs.aaai.org/index.php/ICWSM/article/view/7289.

James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness. In *36th IEEE International Conference on Data Engineering, ICDE 2020, Dallas, TX, USA, April 20-24, 2020*, pp. 1918–1921. IEEE, 2020. doi: 10.1109/ICDE48307.2020.00203. URL https://doi.org/10.1109/ICDE48307.2020.00203.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3315–3323, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html.

Hans Hofmann. Statlog (german credit data). UCI Machine Learning Repository, 1994.

Michael G. Hudgens and M. Elizabeth Halloran. Toward Causal Inference With Interference. *Journal of the American Statistical Association*, (482):832–842, June 2008. ISSN 0162-1459. doi: 10.1198/016214508000000292. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2600548/.

Stefanie Jegelka. Theory of graph neural networks: Representation and learning. *CoRR*, 2022. doi: 10.48550/ARXIV.2204.07697.

Xiangyu Jiang, Yucong Dai, and Yongkai Wu. Fair selection through kernel density estimation. In *International Joint Conference on Neural Networks, IJCNN 2023, Gold Coast, Australia, June 18-23, 2023*, pp. 1–8. IEEE, 2023. doi: 10.1109/IJCNN54540.2023.10191616.

Jian Kang, Jingrui He, Ross Maciejewski, and Hanghang Tong. InFoRM: Individual fairness on graph mining. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pp. 379–389. ACM, 2020. doi: https://dl.acm.org/doi/10.1145/3394486.3403080.

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 4066–4076, 2017. URL `http://papers.nips.cc/paper/6995-counterfactual-fairness`.

Sanghack Lee and Vasant G. Honavar. On learning causal models from relational data. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pp. 3263–3270. AAAI Press, 2016. URL `http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11972`.

Renqiang Luo, Huafei Huang, Shuo Yu, Xiuzhen Zhang, and Feng Xia. Fairgt: A fairness-aware graph transformer. *arXiv preprint arXiv:2404.17169*, 2024.

Jing Ma, Ruocheng Guo, Mengting Wan, Longqi Yang, Aidong Zhang, and Jundong Li. Learning fair node representations with graph counterfactual fairness. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, pp. 695–703. ACM, 2022. doi: 10.1145/3488560.3498391.

Daniel Malinsky, Ilya Shpitser, and Thomas S. Richardson. A potential outcomes calculus for identifying conditional path-specific effects. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, Proceedings of Machine Learning Research, pp. 3080–3088. PMLR, 2019. URL `http://proceedings.mlr.press/v89/malinsky19b.html`.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, (6):115:1–115:35, 2021. doi: 10.1145/3457607.

Elizabeth L. Ogburn and Tyler J. VanderWeele. Causal Diagrams for Interference. *Statistical Science*, (4):559–578, November 2014. ISSN 0883-4237, 2168-8745. doi: 10.1214/14-STS501. URL `https://projecteuclid.org/journals/statistical-science/volume-29/issue-4/Causal-Diagrams-for-Interference/10.1214/14-STS501.full`.

Elizabeth L. Ogburn, Oleg Sofrygin, Iván Díaz, and Mark J. van der Laan. Causal Inference for Social Network Data. *Journal of the American Statistical Association*, pp. 1–15, December 2022. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2022.2131557.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009. ISBN 0-521-89560-X 978-0-521-89560-6.

Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Measuring discrimination in socially-sensitive decision records. In *Proceedings of the SIAM International Conference on Data Mining, SDM 2009, April 30 - May 2, 2009, Sparks, Nevada, USA*, pp. 581–592. SIAM, 2009. doi: 10.1137/1.9781611972795.50.

Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. A study of top-k measures for discrimination discovery. In *Proceedings of the ACM Symposium on Applied Computing, SAC 2012, Riva, Trento, Italy, March 26-30, 2012*, pp. 126–131. ACM, 2012. doi: 10.1145/2245276.2245303.

Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Comput. Surv.*, (3): 51:1–51:44, 2023. doi: 10.1145/3494672.

Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *WIREs Data Mining Knowl. Discov.*, (3), 2022. doi: 10.1002/widm.1452.

Eli Sherman and Ilya Shpitser. Identification and estimation of causal effects from dependent data. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pp. 9446–9457, 2018. URL `http://papers.nips.cc/paper/8153-identification-and-estimation-of-causal-effects-from-dependent-data`.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, pp. 1941–1979, 2008. URL `https://dl.acm.org/citation.cfm?id=1442797`.

Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. *Statistical Methods in Medical Research*, (1):55–75, February 2012. ISSN 0962-2802, 1477-0334. doi: 10.1177/0962280210386779.

Eric J. Tchetgen Tchetgen, Isabel R. Fulcher, and Ilya Shpitser. Auto-G-Computation of Causal Effects on a Network. *Journal of the American Statistical Association*, (534):833–844, April 2021. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2020.1811098.

Tyler J. VanderWeele and Eric J. Tchetgen Tchetgen. Effect partitioning under interference in two-stage randomized vaccine trials. *Statistics & Probability Letters*, (7):861–869, July 2011. ISSN 01677152. doi: 10.1016/j.spl.2011.02.019. URL `https://linkinghub.elsevier.com/retrieve/pii/S0167715211000654`.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018, Gothenburg, Sweden, May 29, 2018*, pp. 1–7. ACM, 2018. doi: 10.1145/3194770.3194776.

Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. In-Processing Modeling Techniques for Machine Learning Fairness: A Survey. *ACM Transactions on Knowledge Discovery from Data*, pp. 3551390, July 2022. ISSN 1556-4681, 1556-472X. doi: 10.1145/3551390.

Min Wen, Osbert Bastani, and Ufuk Topcu. Fairness with Dynamics. *arXiv:1901.08568 [cs, stat]*, January 2019. URL `http://arxiv.org/abs/1901.08568`.

Yongkai Wu, Lu Zhang, and Xintao Wu. On convexity and bounds of fairness-aware classification. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pp. 3356–3362. ACM, 2019. doi: 10.1145/3308558.3313723.

Wenjing Yang, Haotian Wang, Haoxuan Li, Hao Zou, Ruochun Jin, Kun Kuang, and Peng Cui. Your neighbor matters: Towards fair decisions under networked interference. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pp. 3829–3840. ACM, 2024.

I-Cheng Yeh. Default of credit card clients. UCI Machine Learning Repository, 2016. DOI: https://doi.org/10.24432/C55S3H.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, Proceedings of Machine Learning Research, pp. 962–970. PMLR, 2017a. URL `http://proceedings.mlr.press/v54/zafar17a.html`.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pp. 1171–1180. ACM, 2017b. doi: 10.1145/3038912.3052660.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, pp. 335–340. ACM, 2018. doi: 10.1145/3278721.3278779.

Chi Zhang. *Causal Analysis for Generalized Interference Problems*. PhD thesis, UCLA, 2023. URL https://escholarship.org/uc/item/0kd896dg.

Chi Zhang, Karthika Mohan, and Judea Pearl. Causal inference with non-iid data using linear graphical models. *Advances in Neural Information Processing Systems*, pp. 13214–13225, 2022.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making - the causal explanation formula. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 2037–2045. AAAI Press, 2018a. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16949.

Junzhe Zhang and Elias Bareinboim. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 3675–3685, 2018b. URL https://proceedings.neurips.cc/paper/2018/hash/ff1418e8cc993fe8abcfe3ce2003e5c5-Abstract.html.

Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pp. 3929–3935. ijcai.org, 2017. doi: 10.24963/ijcai.2017/549.

Yuchang Zhu, Jintang Li, Yatao Bian, Zibin Zheng, and Liang Chen. One fits all: Learning fair graph neural networks for various sensitive attributes. In Ricardo Baeza-Yates and Francesco Bonchi (eds.), *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pp. 4688–4699. ACM, 2024a. doi: 10.1145/3637528.3672029. URL https://doi.org/10.1145/3637528.3672029.

Yuchang Zhu, Jintang Li, Zibin Zheng, and Liang Chen. Fair graph representation learning via sensitive attribute disentanglement. In *Proceedings of the ACM Web Conference 2024*, pp. 1182–1192, 2024b.

Indre Zliobaite. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, (4):1060–1089, 2017. doi: 10.1007/s10618-017-0506-1.

Appendix

A Proof of Theorem 4 in the main paper

For any two variables X, Y in an SCM, let $y(u)$ be the value of Y of an instance whose exogenous variable is u, and $y_x(u)$ be the value of Y under the intervention $x(u) = x$. According to Pearl (2009), we have the following lemmas.

Lemma 6. *Given the causal diagram in Fig. 1f in the main paper, we have that $X_a \perp A$ for any a.*

Lemma 7. *Given the causal diagram in Fig. 1f in the main paper, for any node u, we have that $x_{\mathsf{s}}(u) = x_{\mathsf{s},c,a}(u) = x_a(u)$ if $a_{\mathsf{s}}(u) = a$ and $c(u) = c$.*

Lemma 8. *Given the causal diagram in Fig. 1f in the main paper, for any node u, we have that $a_{\mathsf{s}}(u) = a_{\mathsf{s},c}(u)$ if $c(u) = c$.*

Theorem 3. *Given the causal diagram in Fig. 1f in the main paper, we have*

$$P(x|do(\mathsf{s}) = s) = \sum_c P(c) \sum_a P(x|a)P(g(\mathsf{s}, c) = a).$$

Proof. According to the formula of the conditional probability, we directly have

$$\begin{aligned} P(x|do(\mathsf{s}) = s) &= \sum_{c,a} P(x|c, a, do(\mathsf{s}) = s)P(c, a|do(\mathsf{s}) = s) \\ &= \sum_{c,a} P(x|c, a, do(\mathsf{s}) = s)P(c|do(\mathsf{s}) = s)P(a|c, do(\mathsf{s}) = s). \end{aligned}$$

Since S is not a descendent of C in the causal diagram, it follows that

$$P(x|do(\mathsf{s}) = s) = \sum_{c,a} P(x|c, a, do(\mathsf{s}) = s)P(c)P(a|c, do(\mathsf{s}) = s).$$

According to Lemma 8, we have $P(a|c, do(\mathsf{s}) = s) = P(a|do(c), do(\mathsf{s}) = s)$ which can be rewritten as $P(g(\mathsf{s}, c) = a)$ below using the mapping g. By similarly applying Lemma 7, we have $P(x|c, a, do(\mathsf{s}) = s) = P(x|do(c), do(a), do(\mathsf{s}) = s) = P(x|do(a))$. As a result, we have

$$P(x|do(\mathsf{s}) = s) = \sum_{c,a} P(x|do(a))P(c)P(g(\mathsf{s}, c) = a).$$

Then, we rewrite the above equation as

$$P(x|do(\mathsf{s}) = s) = \sum_{c,a} \sum_{a'} P(x|a', do(a))P(a')P(c)P(g(\mathsf{s}, c) = a)$$

According to Lemma 6, we have $P(x|a', do(a)) = P(x|a, do(a))$, which is equal to $P(x|a)$ according to the Composition Axiom. Finally, we have that

$$\begin{aligned} P(x|do(\mathsf{s}) = s) &= \sum_{c,a} P(x|a) \sum_{a'} P(a')P(c)P(g(\mathsf{s}, c) = a) \\ &= \sum_c P(c) \sum_a P(x|a)P(g(\mathsf{s}, c) = a). \end{aligned}$$

Hence, the theorem is proved. □

B Discussion of Proposed Conditions

B.1 Decomposability

The Decomposability condition states that when a node's outcome is influenced by both its own attributes and its neighbors' attributes, the influence can be decomposed into two steps: first aggregating information from neighbors, then combining it with the node's own attributes. This condition helps make causal inference tractable in graph settings by providing a structured way to model how information flows through the network. With this condition, we can decompose the causal effect of the neighbors into the message passing mechanism, denoted as $f^{MP}(\cdot)$, and the internal causal mechanism within the node, denoted as $f^{INT}(\cdot)$.

For example, consider a social media platform that uses an algorithm to deliver purchase discounts to users. We posit that a user may choose to subscribe to the supplier (i.e., X), and if he/she decides to subscribe to the supplier, the chance of receiving the discount will increase. Due to the connections in the social network, we posit that a user's decision to subscribe is influenced by his/her own situation (i.e., S) as well as his/her neighbors. Then, Decomposition posits that the user will first aggregate the situations from all the neighbors and then combine them with his/her own situation when making the decision. This two-step process allows us to separately model the network effects and individual effects while still capturing their joint influence on the final outcome.

B.2 Graph Independence

The Graph Independence condition posits that the graph structure (captured by the node color C) is independent of the exogenous variable (denoted as U_X), which ensures that all relevant information from the neighborhood regarding the interventional value of X can be effectively summarized in the intermediate variable A. As a result, the implication of Proposition 2 and Condition 2 together implies that we can convert the networked causal diagram in Fig. 1e (from the main paper) to an equivalent causal diagram as shown in Fig. 1f (from the main paper).

In the example of a social network where users are connected based on shared interests or demographics, the Graph Independence condition means that the network connections themselves (who is friends with whom) are not influenced by exogenous factors that also affect the non-sensitive attributes. This condition allows us to treat the network structure as fixed and unaffected by interventions, thereby preventing cyclic dependencies in the causal diagram (i.e., Fig 1f). In other words, when we intervene on a node's attributes, we posit this intervention does not alter the underlying network topology.

C Experiments

C.1 Datasets

In the semi-synthetic dataset, we leverage the Credit Dataset (Yeh, 2016). To get full control over the data generation, we define the data generation mechanism as follows. Denoting the original features in the dataset as C, we first build a classifier to predict S. We use this classifier to soft label the prediction of S to obtain the probability distribution of generating S. Then, we similarly build another classifier to estimate Y in order to obtain the probability distribution of generating Y. After that, we randomly initialize a GNN $g(\cdot) : \mathcal{S} \to \mathcal{A}$ to mimic the influence of neighbors' sensitive attributes of a certain node on its own attributes. Finally, we generate our semi-synthetic dataset as follows:

$$
\begin{aligned}
S_i^g &\sim Bernoulli(p) \\
X_i^g &= g(S^g) + C_i + \xi \\
Y_i^g &= f_y(X_i^g)
\end{aligned}
$$

where the sensitive attribute is sampled from a Bernoulli distribution, $p = f_s(C_i)$ is the probability of $S_i^g = 1$, and ξ is the random noise that is sampled from Gaussian distribution. We simulate the probability of each edge (i, j) based on the similarity between X_i^g and X_j^g. We generate ground truth counterfactual data by setting all S^g to 1 and 0 to obtain positive and negative interventional distributions, respectively.

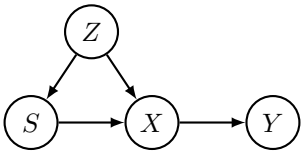

Figure 6: The networked causal diagram for node classification.

For the real-world graphs, we conduct experiments on widely used real-world datasets, namely Credit Defaulter (Yeh, 2016) and German (Hofmann, 1994). The dataset details are as follows.

Credit Defaulter: the nodes in the dataset are used to represent the credit card users, and the edges are formed based on the similarity of the payment information. The task is to classify the default payment method with the sensitive attribute "Sex". We treat "Education", "Marriage", and "Age" as Z, i.e., variables other than S that directly affect X.

German: The German credit network consists of 1,000 nodes, which represent clients of a German bank. These nodes are interconnected based on the similarity of their credit accounts. The objective is to categorize clients as either excellent or bad credit risks, taking into consideration the clients' gender as the sensitive attribute. We treat "YearsAtCurrentJob" and "JobClassIsSkilled" as Z.

C.2 Implementation

We use a one-layer message-passing neural network to aggregate the sensitive causal effect of one-hop neighbors. We train MPNN with 0.01 learning rate and 500 epochs. The models are all implemented using PyTorch 1.12.0 and PyG 2.4.0 and evaluated in a Linux server with an Intel(R) Core(TM) 19-10900X CPU and an NVIDIA GeForce RTX 3070 GPU. The memory consumption is about 2000 MiB. We use cVAE to reconstruct features conditional on $\hat{a}$ and c. For training cVAE, the learning rate is 0.01, and the number of epochs is 800. Experimental results are averaged over five repeated executions. We use the Adam optimizer for both components of our proposed framework and implement our method with PyTorch. For the constraint-based method, we train a multilayer perceptron (MLP) with corresponding fairness regularization terms to achieve fairness.

Risk difference on IID data: The risk difference usually refers to the difference of the positive predictions between the favorable group and the unfavorable group. It is easy to compute the possibility of output given certain sensitive attribute $P(y \mid s) = \mathbb{E}_{x|s}P(y \mid x)$. Then, we design the regularization term as $P(y \mid s^+) - P(y \mid s^-)$.

Causal inference on IID data: For IID data, we usually use a structural causal model to describe the causal relationship between two variables. For example, the causal relationship between S and X is given by:

$$x = f(s, u).$$

We consider the same causal structure in Fig. 6 but neglect the network causal effect. To compute the possibility of output given certain intervention on the sensitive attribute $P(y|do(s))$:

$$\begin{aligned} P(y \mid do(s)) &= \sum_{z,x} P(z)P(x \mid s, z)P(y \mid x) \\ &= \sum_{z,x} P(z \mid s)\frac{P(s)}{P(s \mid z)}P(x \mid s, z)P(y \mid x) \\ &= \sum_{z,x} P(z, x \mid s)\frac{P(s)}{P(s \mid z)}P(y \mid x) \\ &= \mathbb{E}_{z,x\sim P(z,x|s)}\left[\frac{P(s)}{P(s \mid z)}P(y|x)\right] \end{aligned} \tag{6}$$

Then, we design the regularization term as $P(y \mid do(s^+)) - P(y \mid do(s^-))$.

C.3 Computational Cost

To compare the computational cost of our method and the baselines, we measured the runtime and memory usage for our method and the baselines on the Credit dataset. As shown in Tab. 5, MPVA achieves the lowest memory usage (1365 MB) and offers the second fastest running time (29.2 seconds), compared to the fastest runtime (17.0893 seconds) of NIFTY, demonstrating its computational efficiency. The superior memory efficiency of MPVA makes it particularly suitable for large-scale graph applications where memory constraints are critical. Notably, GEAR exceeded the 3600-second time limit and could not complete the experiment, highlighting the scalability challenges of existing methods.

Table 5: Runtime and memory usage comparison of different methods.

Method	**Time (s)**	**Memory (MB)**
FairGNN	81.7968	3639
NIFTY	17.0893	2419
GEAR	> 3600	–
FairINV	63.7923	2103
MPVA	29.2358	1365