Learning Fair Graph Representations via Probability of Necessity and Sufficiency

Chuxun Liu[1], Qingfeng Chen[2*], Debo Cheng[3], Jiangzhang Gan[3†], Jiuyong Li[4], Lin Liu[4]
[1] Guilin University of Electronic Technology
[2] Guangxi University
[3] Hainan University
[4] University of South Australia
chuxunliu@mails.guet.edu.cn, qingfeng@gxu.edu.cn, chengd@hainanu.edu.cn,
ganjz@hainanu.edu.cn, Jiuyong.Li@unisa.edu.au, Lin.Liu@unisa.edu.au

Abstract

Graph Neural Networks (GNNs) excel at modeling graph data but often amplify biases tied to sensitive attributes like gender and race. Existing causality-based methods use isolated interventions on graph topology or features but struggle to produce representations that balance predictive power with fairness. This leads to two issues: (1) weak predictive power, where representations miss critical task-relevant features, and (2) bias amplification, where representations encode sensitive attributes, causing unfair outcomes. To address these issues, we introduce the Probability of Necessity and Sufficiency (PNS), where necessity ensures representations capture only essential features for predictions, and sufficiency guarantees these features are adequate without relying on sensitive attributes. We propose FairSNR, a fairness-aware graph representation learning framework that introduces constraints based on the PNS. This leverages PNS to guide the learning of fair representations from graph data. In particular, FairSNR employs an encoder to learn node representations with high PNS for downstream tasks. To compute and optimize PNS, FairSNR introduces an intervenor to generate the most challenging counterfactual interventions on the representations, thereby enhancing the model's causal stability even under worst-case scenarios. Further, a discriminator is trained to detect and mitigate sensitive information leakage in the learned representations, effectively disentangling sensitive biases from task-relevant features. Experiments on real-world graph datasets demonstrate that FairSNR outperforms existing state-of-the-art (SOTA) methods in both fairness and utility.

Introduction

Graph Neural Networks (GNNs) have achieved remarkable success in modeling graph-structured data, driving breakthroughs in fields such as recommender systems, molecular property prediction, and bioinformatics (Ouyang et al. 2024; Li and Nabavi 2024; Lei et al. 2025). By leveraging the relational inductive biases inherent in graph data, GNNs effectively capture structural dependencies and contextual information (Job et al. 2025). However, as GNNs are increasingly deployed in high-stakes applications, growing concerns have emerged regarding their potential to amplify prediction biases associated with sensitive attributes (e.g., gender, race, or socioeconomic status) (Dai et al. 2024). Such biases can lead to unfair or even discriminatory outcomes, particularly in decision-critical contexts such as credit scoring (Trinh and Zhang 2024) and risk assessment (Li et al. 2025a).

Recently, fairness-aware graph representation learning has garnered significant research interest. With the growing adoption of GNNs, numerous methods have been proposed to improve fairness while preserving acceptable utility performance (Luo et al. 2025; Yang et al. 2024; Zhu et al. 2024b). Among these, causal inference has emerged as a theoretically grounded approach for fairness learning, aiming to fundamentally disentangle the influence of sensitive attributes. Causal methods mitigate spurious correlations and enable counterfactual reasoning by explicitly modeling the causal paths between sensitive attributes and prediction outcomes, thus offering a principled perspective on fairness. Typical strategies include edge pruning, feature masking, and counterfactual augmentation, which primarily reduce bias propagation through interventions on graph structures or node features (Li et al. 2024; Guo et al. 2023).

Existing fairness-aware GNN methods can be broadly classified into three categories (Dong et al. 2023). Structural intervention methods, such as FairDrop (Spinelli et al. 2021) and FairWalk (Rahman et al. 2019), aim to reduce bias propagation by adjusting edge retention probabilities or modifying path sampling strategies. Adversarial learning methods, such as FairGNN (Dai and Wang 2021) and FairVGNN (Wang et al. 2022), incorporate adversarial modules to explicitly disentangle sensitive attributes from the learned embeddings. Causal inference-based methods, such as NIFTY (Agarwal, Lakkaraju, and Zitnik 2021) and FairINV (Zhu et al. 2024a), attempt to model the causal effect of sensitive attributes on prediction outcomes, thereby mitigating discrimination through counterfactual reasoning or causal adjustment.

Current approaches however suffer from two critical limitations. Firstly, most methods rely on isolated interventions, such as modifying graph topology or balancing input features, which may lack semantic consistency and generalizability (Li et al. 2024). Secondly, and more fundamentally, these methods often overlook the necessity and sufficiency

*Co-corresponding authors

conditions of the learned representations, which are two crucial criteria for ensuring fair and robust prediction (Yang et al. 2023). Especially, a representation that is necessary but not sufficient may fail to provide predictive power, while one that is sufficient but not necessary may encode spurious dependencies with sensitive attributes, ultimately undermining fairness and generalizability.

To address these limitations, we propose FairSNR, a novel fairness-aware graph representation learning framework grounded in the causal principles of necessity and sufficiency. Our key insight is to treat fairness not merely as a post-hoc constraint or intervention, but as an intrinsic property of the learned fair representation. In practice, FairSNR employs PNS as a theoretical tool to regularize the representation learning process (Pearl 2009). By maximizing PNS with respect to the prediction task while minimizing its dependency on sensitive attributes, we ensure that the learned representation is both sufficiently informative for prediction and necessarily invariant to sensitive factors. In conjunction with adversarial learning, which decouples sensitive information from the learned representations, our framework effectively separates bias-inducing features from task-relevant information, thereby enhancing both fairness and utility in downstream tasks. The main contributions of this work are summarized as follows:

- We introduce the causal concept of PNS into the field of fair graph representation learning, where achieving fairness requires satisfying both sufficiency with respect to the prediction task and necessity invariance with respect to sensitive attributes. To the best of our knowledge, this is the first work in fairness that leverages causal PNS.
- We develop FairSNR, a novel fairness-aware graph representation learning model that explicitly incorporates PNS as a core regularization term. FairSNR maximizes the causal relevance of representations for the prediction task while minimizing their sensitivity to protected attributes via adversarial learning.
- Extensive empirical evaluations on multiple real-world benchmark datasets demonstrate that FairSNR consistently outperforms existing SOTA across a range of fairness and predictive performance metrics.

Preliminaries

In this section, we first introduce the notations used in this work, followed by the definition of PNS and the associated theoretical foundations.

Notations

Let $\mathcal{G} = (\mathcal{V}, \mathbf{A}, \mathbf{X})$ denote an unweighted, undirected graph, where $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ is the set of $n = |\mathcal{V}|$ nodes. The graph structure is represented by a symmetric adjacency matrix $\mathbf{A} \in \{0,1\}^{n \times n}$, where $\mathbf{A}_{ij} = 1$ if there is an edge between nodes v_i and v_j, and $\mathbf{A}_{ij} = 0$ otherwise. Each node v_i is associated with a feature vector $\mathbf{x}_i \in \mathbb{R}^d$, and we denote the node feature matrix as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{n \times d}$.

In fairness-aware settings, each node v_i is associated with a sensitive attribute $s_i \in S$ (e.g., gender, race) and a ground truth label $y_i \in \mathcal{Y}$, where S and $\mathcal{Y}$ denote the sets of possible sensitive attribute values and label categories, respectively. The downstream task considered in this work is node classification. The target label is denoted as $y \in \{0,1\}$, and the sensitive attribute is denoted as $s \in \{0,1\}$.

Definition of PNS on Graph-structured Data

We extend the concept of the PNS (Pearl 2009) to graph-structured data.

Definition 1 (Probability of Necessary and Sufficient) *Given a node v in the graph $\mathcal{G}$ with label y, the probability of a necessary and sufficient cause of a learned representation $\mathbf{c}$ for y is defined as:*

$$\begin{aligned} PNS(\mathbf{c}, \bar{\mathbf{c}}) := P\Big(Y_{do(\mathbf{C}=\mathbf{c})} = y \,\Big|\, \mathbf{C} = \bar{\mathbf{c}}, \\ Y \neq y\Big) \cdot P(\mathbf{C} = \bar{\mathbf{c}}, Y \neq y) \\ + P\Big(Y_{do(\mathbf{C}=\bar{\mathbf{c}})} \neq y \,\Big|\, \mathbf{C} = \mathbf{c}, \\ Y = y\Big) \cdot P(\mathbf{C} = \mathbf{c}, Y = y). \end{aligned} \tag{1}$$

Here, $\mathbf{C}$ and Y denote the cause (i.e., representation) and effect (i.e., label) of node v, respectively, where $Y = y$ represents the observed true label. $\mathbf{c}$ and $\bar{\mathbf{c}}$ refer to two different states of $\mathbf{C}$. The operator $do(\cdot)$ signifies an intervention, simulating an external manipulation of the causal mechanism (Pearl 2009). In the above definition, $P\big(Y_{do(\mathbf{C}=\mathbf{c})} = y \,\big|\, \mathbf{C} = \bar{\mathbf{c}}, Y \neq y\big)$ represents the probability that the outcome Y would occur if we intervened to set the cause $\mathbf{C}$ to $\mathbf{c}$, given that we actually observed $\mathbf{C} = \bar{\mathbf{c}}$ and $Y \neq y$.

In this context, the first term in the PNS formula corresponds to the probability of sufficiency, and the second term represents the probability of necessity. A higher value of PNS indicates that the variable $\mathbf{C}$ is more likely to be a necessary and sufficient cause of Y.

Computing such counterfactual-based probabilities in real-world systems, however, is often challenging or even infeasible due to the need for knowledge of the underlying structural causal model. Fortunately, PNS, although originally defined through counterfactual distributions, becomes identifiable from purely observational data under certain conditions, particularly when the assumptions of **exogeneity** (no unmeasured confounding) and **monotonicity** (no prevention) are satisfied (Tian and Pearl 2000; Pearl 2009).

Definition 2 (Exogeneity (Pearl 2009)) *A variable $\mathbf{C}$ is said to be* exogenous *with respect to Y if its interventional probability can be identified by the corresponding conditional probability.*

In the context of the graph domain, the node representation $\mathbf{C}$ is said to be exogenous with respect to its label Y in both the source graph domain $\mathcal{G}_s$ and the target graph domain $\mathcal{G}_t$ if and only if the following conditions hold:

$$\begin{aligned} P_{\mathcal{G}_s}\big(Y_{do(\mathbf{C}=\mathbf{c})} = y\big) &= P_{\mathcal{G}_s}(Y = y \mid \mathbf{C} = \mathbf{c}), \\ P_{\mathcal{G}_t}\big(Y_{do(\mathbf{C}=\mathbf{c})} = y\big) &= P_{\mathcal{G}_t}(Y = y \mid \mathbf{C} = \mathbf{c}). \end{aligned} \tag{2}$$

Eq. 2 implies that, when $\mathbf{C}$ is exogenous to Y, the discrepancy between the interventional and observational (i.e., conditional) distributions disappears. In other words, observing $\mathbf{C} = \mathbf{c}$ provides the same information about Y as actively intervening to set $\mathbf{C} = \mathbf{c}$, indicating the absence of confounding between the two variables.

Definition 3 (Monotonicity (Tian and Pearl 2000)) *The variable Y is said to be* monotonic *with respect to $\mathbf{C}$ if, for two possible states $\mathbf{c}$ and $\bar{\mathbf{c}}$ of $\mathbf{C}$, one of the following conditions holds:*

$$\begin{cases} P\big(Y_{do(\mathbf{C}=\mathbf{c})} = y,\, Y_{do(\mathbf{C}=\bar{\mathbf{c}})} \neq y\big) = 0, & or \\ P\big(Y_{do(\mathbf{C}=\mathbf{c})} \neq y,\, Y_{do(\mathbf{C}=\bar{\mathbf{c}})} = y\big) = 0 \end{cases} \tag{3}$$

This definition ensures that $\mathbf{C}$ has a monotonic causal effect on Y, meaning that changing $\mathbf{C}$ from one state to another does not simultaneously increase and decrease the likelihood of the outcome Y.

Lemma 1 ((Pearl 2009)) *If $\mathbf{C}$ is exogenous to Y and Y is monotonic with respect to $\mathbf{C}$, then the PNS can be computed as:*

$$\begin{aligned} PNS(\mathbf{c}, \bar{\mathbf{c}}) = & \underbrace{P_{\mathcal{G}_t}(Y = y \mid \mathbf{C} = \mathbf{c})}_{\textit{sufficiency}} \\ & - \underbrace{P_{\mathcal{G}_t}(Y = y \mid \mathbf{C} = \bar{\mathbf{c}})}_{\textit{necessity}}. \end{aligned} \tag{4}$$

According to Lemma 1, the computation of PNS becomes feasible using observational data under the joint assumptions of exogeneity and monotonicity. The original proof was presented by (Pearl 2009), further extended through probabilistic reasoning by (Yang et al. 2023), and subsequently generalized to subgraph learning by (Chen et al. 2025).

Methodology

This section presents the methodology of the proposed FairSNR framework.

Causal Analysis

Recent research has increasingly incorporated causal modeling into fairness studies to uncover intrinsic biases associated with sensitive attributes (Li et al. 2024; Agarwal, Lakkaraju, and Zitnik 2021). In this work, we propose a novel causal perspective that unifies the graph data generation process and the GNN prediction mechanism within a Structural Causal Model (SCM) (Pearl 2009), as illustrated in Figure 1.

- $S \rightarrow \mathbf{X} \rightarrow \mathbf{C} \rightarrow Y$ and $S \rightarrow \mathbf{A} \rightarrow \mathbf{C} \rightarrow Y$: Achieving fairness in GNNs presents a unique challenge compared to other data modalities like images or text. Unlike conventional models, GNNs make predictions based on a node's entire contextual subgraph, allowing the sensitive attribute S to influence the prediction Y through two critical causal pathways. $S \rightarrow \mathbf{X} \rightarrow \mathbf{C} \rightarrow Y$: This pathway captures how the sensitive attribute S influences a node's initial features $\mathbf{X}$. GNNs encoder then processes these biased features into its representation $\mathbf{C}$, which ultimately leads to a discriminatory outcome Y. The causal pathway $S \rightarrow \mathbf{A} \rightarrow \mathbf{C} \rightarrow Y$ illustrates how sensitive attributes S influence the prediction outcome Y through the graph topology $\mathbf{A}$ and node representations $\mathbf{C}$. More precisely, S affects $\mathbf{A}$ by inducing structural bias. for example, social homophily may lead to stronger connectivity between nodes with similar attributes. The message-passing mechanism of GNNs then aggregates information based on this biased adjacency matrix $\mathbf{A}$, encoding the structural bias into the learned node representations $\mathbf{C}$, which ultimately propagates to the prediction Y.

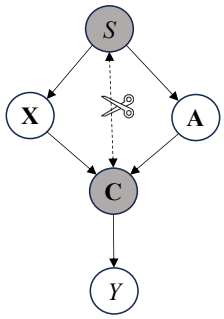


Figure 1: The SCM illustrates the GNNs prediction process. Fairness issues in graph learning can arise through three causal pathways: $S \rightarrow \mathbf{X} \rightarrow \mathbf{C} \rightarrow Y$, $S \rightarrow \mathbf{A} \rightarrow \mathbf{C} \rightarrow Y$, and $S \rightarrow \mathbf{C} \rightarrow Y$. The FairSNR framework mitigates these issues by simultaneously blocking all three pathways that transmit sensitive information S to the prediction target Y.

- $S \leftrightarrow \mathbf{C} \rightarrow Y$: This causal pathway represents the direct influence of the sensitive attribute S on the node representation $\mathbf{C}$, which in turn affects the prediction outcome Y. To mitigate this bias, we employ adversarial decorrelation to explicitly weaken sensitive information from the learned representations. Specifically, an adversarial loss is introduced to discourage the model from encoding any predictive signal of S in $\mathbf{C}$. This weakens the dependency between between S and $\mathbf{C}$, effectively blocking the causal link and ensuring that the final prediction Y is made based on fair and unbiased representations.

In summary, discriminatory decisions in GNNs arise from the three causal pathways discussed above. To eliminate the influence of the sensitive attribute S on the prediction Y through the paths $S \rightarrow \mathbf{X} \rightarrow \mathbf{C} \rightarrow Y$ and $S \rightarrow \mathbf{A} \rightarrow \mathbf{C} \rightarrow Y$, we propose FairSNR, which learns a representation $\mathbf{C}$ that is both sufficient and necessary, while also being independent of S. Sufficiency ensures that $\mathbf{C}$ contains all information in $\mathbf{X}$ and $\mathbf{A}$ that is relevant for predicting Y, making Y conditionally independent of $\mathbf{X}$ and $\mathbf{A}$ given $\mathbf{C}$. Necessity guarantees that $\mathbf{C}$ retains only the essential causal factors of Y, excluding redundant information. Furthermore, by enforcing $\mathbf{C} \perp S$, FairSNR explicitly removes the dependence of $\mathbf{C}$ on S, thereby blocking the causal influence of S on Y via $\mathbf{X}$ or $\mathbf{A}$. This severs the unfair pathways and fundamentally eliminates bias arising from sensitive attributes.

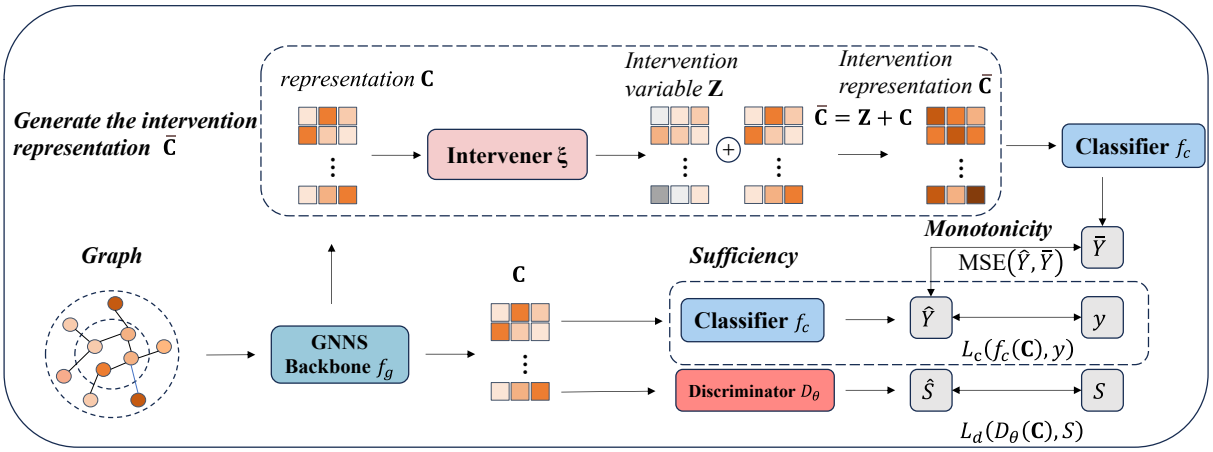


Figure 2: Overview of the FairSNR framework. The GNN backbone f_g extracts node representations $\mathbf{C}$ from the input graph. The Intervener ξ generates an intervention variable $\mathbf{Z}$, which is used to construct the counterfactual representation $\bar{\mathbf{C}} = \mathbf{C} + \mathbf{Z}$. The classifier f_c ensures sufficiency by predicting the label $\hat{Y}$, while the discriminator D_θ mitigates sensitive attribute leakage by minimizing the dependence on sensitive information S. Monotonicity is enforced by minimizing the mean squared error (MSE) between the original prediction $\hat{Y}$ and the intercention-based prediction $\bar{Y}$.

GNNs in FairSNR

GNNs leverage a message-passing framework, enabling each node to iteratively aggregate information from its neighbors to refine its latent representations. At the k-th layer, a node v_i aggregates information from its neighborhood $\mathcal{N}(i)$ and updates its representation accordingly. This process is typically formulated as:

$$\mathbf{H}^{(k)} = \sigma\left(\tilde{\mathbf{A}}\,\mathbf{H}^{(k-1)}\,\mathbf{W}^{(k-1)}\right), \tag{5}$$

where $\mathbf{H}^{(k)}$ is the node embedding matrix at layer k (with $\mathbf{H}^{(0)} = \mathbf{X}$), $\tilde{\mathbf{A}}$ is a normalized or modified adjacency matrix (e.g., $\tilde{\mathbf{A}} = \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\,\tilde{\mathbf{D}}^{-1/2}$), $\mathbf{W}^{(k-1)}$ is a learnable weight matrix, and $\sigma(\cdot)$ is a non-linear activation function.

For node classification tasks, a GNN model f typically consists of a GNN-based encoder f_g and a task-specific classifier f_c. The f_g maps the input graph structure and node features to a low-dimensional representation:

$$\mathbf{C} = f_g(\mathbf{A}, \mathbf{X}), \quad \mathbf{C} \in \mathbb{R}^{n \times d'} \tag{6}$$

where d' is the embedding dimension. The classifier in FairSNR then predicts labels based on this learned representation:

$$\hat{Y} = f_c(\mathbf{C}). \tag{7}$$

Necessary and Sufficient Graph Representation

As illustrated in the Figure 2, our objective is to learn a node representation $\mathbf{C}$ that is both necessary and sufficient for predicting the node's label y, while remaining independent of the node's sensitive attribute S.

We generalize the PNS risk proposed in (Yang et al. 2023) to the graph setting at the node level. For a node v with label y, the PNS risk is designed to evaluate the necessity and sufficiency of the representation $\mathbf{C}$ learned by the GNN encoder f_g. We define the empirical PNS risk on the graph as $\hat{R}_t(f_c, f_g, \xi)$:

$$\hat{R}_t(f_c, f_g, \xi) := \mathbb{E}_{(v,y)\sim\mathcal{T}_G}\Big[\mathbb{I}\left[\operatorname{sign}(f_c(\mathbf{c})) \neq y\right] + \mathbb{E}_{\bar{\mathbf{c}}\sim P_t(\bar{\mathbf{C}}|\mathbf{x},\mathbf{A})}\mathbb{I}\left[\operatorname{sign}(f_c(\bar{\mathbf{c}})) = y\right]\Big], \tag{8}$$

where $\mathcal{T}_G$ denotes the empirical distribution of node-label pairs on the graph, $\mathbf{x}$ represents the input features of node v, and $\mathbf{A}$ is the adjacency matrix encoding the graph structure. $\mathbf{C} \sim P_t(\mathbf{C} \mid \mathbf{x}, \mathbf{A})$ denotes a sufficient representation sampled from the learned distribution, while $\bar{\mathbf{C}} \sim P_t(\bar{\mathbf{C}} \mid \mathbf{x}, \mathbf{A})$ denotes a necessary representation obtained via intervention. The indicator function $\mathbb{I}[\cdot]$ evaluates the correctness of the prediction.

Directly optimizing the PNS risk is intractable. Following the approach in (Yang et al. 2023), we instead decompose the risk and employ a more tractable upper bound, which consists of the sufficiency risk and a monotonicity measurement. The empirical sufficiency risk $\hat{\text{SF}}^s$ corresponds directly to the classification error in the node classification task. It is defined as:

$$\hat{\text{SF}}^s(f_c, f_g) := \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\mathbb{E}_{\mathbf{c}_v\sim\hat{P}^{f_g}(\mathbf{C}|v,\mathcal{G})}\mathbb{I}\left[\operatorname{sign}(f_c(\mathbf{c}_v)) \neq y\right], \tag{9}$$

where $\mathbf{c}_v$ is a representation sampled from the encoder-induced distribution $\hat{P}^{f_g}(\mathbf{C} \mid v, \mathcal{G})$, and y is the ground-truth label of node v. The indicator function $\mathbb{I}[\cdot]$ returns 1 if the predicted label $\operatorname{sign}(f_c(\mathbf{c}))$ differs from the true label y, and 0 otherwise.

The empirical monotonicity measurement $\hat{M}^s$ evaluates the consistency between the predictions made on the original representation $\mathbf{c}$ and the intervention-based representation $\bar{\mathbf{c}}$:

$$\hat{M}^s(f_c, f_g, \xi) := \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\mathbb{E}_{\mathbf{c}\sim\hat{P}^{f_g},\,\bar{\mathbf{c}}\sim\hat{P}^{\xi}}\mathbb{I}[\operatorname{sign}(f_c(\mathbf{c})) = \operatorname{sign}(f_c(\bar{\mathbf{c}}))]. \tag{10}$$

Based on Eqs. 9 and 10, the empirical PNS risk is upper bounded by the combination of monotonicity measurement and sufficiency risk:

$$\hat{R}^s \leq \hat{M}^s(f_c, f_g, \xi) + 2 \cdot \hat{\text{SF}}^s(f_c, f_g). \tag{11}$$

Therefore, our optimization objective is to minimize this upper bound, encouraging the learned representation to be both necessary and sufficient for prediction while maintaining a monotonic behavior under intervention.

To learn a representation $\mathbf{C}$ that is both necessary and sufficient for the downstream task while remaining unbiased, we formulate the learning process as a min-max optimization problem. Simultaneously, we incorporate an adversarial loss to mitigate the influence of the sensitive attribute S.

The overall optimization objective is defined as:

$$\begin{aligned} \min_{f_g, f_c} \max_{\xi} \; & \mathcal{L}_{\text{PNS}}(f_c, f_g, \xi) + \lambda_{\text{k}} \cdot \mathcal{L}_{\text{KL}}(f_g, \xi) \\ & + \lambda_{\text{d}} \cdot \mathcal{L}_{\text{d}}(f_g, \theta). \end{aligned} \tag{12}$$

The three losses are the PNS loss ($\mathcal{L}_{\text{PNS}}$), the KL-divergence regularization ($\mathcal{L}_{\text{KL}}$), and the disentanglement loss ($\mathcal{L}_{\text{d}}$). where $\mathcal{L}_{\text{PNS}}$ encourages the learned representations $\mathbf{C}$ to be sufficient for the task and necessary under interventions; $\mathcal{L}_{\text{d}}$ aims to disentangle the sensitive attribute S from the representation $\mathbf{C}$ through adversarial learning; λ_{k} and λ_{d} are non-negative hyperparameters that balance the contributions of the KL regularization and adversarial loss, respectively.

Our objective is to minimize the upper bound of the PNS risk while incorporating a semantic separability constraint to ensure that the representation $\mathbf{c}$ and the intervention-based representation $\bar{\mathbf{c}}$ remain at least a distance δ apart. Formally, the loss $\mathcal{L}_{\text{PNS}}(f_g, f_c, \xi)$ is defined as:

$$\begin{aligned} \min_{f_g, f_c} \max_{\xi} \quad & \hat{M}^s(f_c, f_g, \xi) + \hat{\text{SF}}^s(f_c, f_g), \\ \text{s.t.} \quad & \|\mathbf{c} - \bar{\mathbf{c}}\|_2 > \delta. \end{aligned} \tag{13}$$

where the constraint $\|\mathbf{c} - \bar{\mathbf{c}}\|_2 > \delta$ is imposed based on the Semantic Separability assumption to ensure sufficient perturbation between the original and intervention-based representations. The term $\mathcal{L}_{\text{KL}}$ serves as a regularizer for the representation distributions, ensuring that the empirical risk approximates the expected risk effectively:

$$\begin{aligned} \mathcal{L}_{\text{KL}}(f_g, \xi) = \mathbb{E}_{v \in \mathcal{V}} \Big[& KL\big(\hat{P}^{f_g}(\mathbf{C} \mid \mathbf{X}, \mathbf{A}) \,\|\, \pi_{\mathbf{C}}\big) \\ & + KL\big(\hat{P}^{\xi}(\bar{\mathbf{C}} \mid \mathbf{X}, \mathbf{A}) \,\|\, \pi_{\bar{\mathbf{C}}}\big) \Big], \end{aligned} \tag{14}$$

where $\pi_{\mathbf{C}}$ and $\pi_{\bar{\mathbf{C}}}$ are predefined prior distributions, typically chosen as standard Gaussian distributions.

To remove the influence of the sensitive attribute S, we introduce an adversarial loss. A discriminator D_θ is trained to predict the sensitive attribute S from the representation $\mathbf{c}$, while the GNN encoder f_g aims to generate representations that prevent D_θ from correctly identifying S, i.e., effectively minimizing the dependency between $\mathbf{C}$ and S.

$$\mathcal{L}_{\text{d}}(f_g, \theta) = \mathbb{E}_{v \in \mathcal{V}} \left[\log D_\theta \left(S \mid \mathbf{C} = f_g(\mathbf{X}, \mathbf{A}) \right) \right]. \tag{15}$$

Algorithm 1: Training Algorithm of FairSNR

Input: Graph $\mathcal{G} = (\mathcal{V}, \mathbf{A}, \mathbf{X})$, sensitive attribute S; hyperparameters α, β, T.
Output: Trained encoder f_g, classifier f_c, and intervenor ξ
Initialize f_g, f_c, ξ, variational encoder f_{mv}, discriminator D
Set optimizers for all modules accordingly
for $t = 1$ to T **do**
  Reset parameters of all modules
  for $i = 1$ to $|\mathcal{V}_{train}|$ **do**
    // Step A: Train Discriminator D
    Freeze f_g; compute $\mathbf{C} \leftarrow f_g(\mathbf{X}, \mathbf{A})$
    $\hat{\mathbf{S}} \leftarrow D(\mathbf{C})$
    $\mathcal{L}_d \leftarrow$ Calculate loss function according to Eq 15
    Update D by gradient descent
    // Step B: Update f_g, f_c, f_{mv}, ξ
    Enable training for f_g, f_c, f_{mv}, ξ
    $\mathbf{C} \leftarrow f_g(\mathbf{X}, \mathcal{E})$, $(\boldsymbol{\mu}, \boldsymbol{\sigma}) \leftarrow f_{mv}(\mathbf{X}, \mathbf{A})$
    $\hat{\mathbf{S}} \leftarrow D(\text{GRL}(\mathbf{C}))$
    $\mathcal{L}_d \leftarrow$ Calculate loss function according to Eq 15
    $\mathbf{Z} \leftarrow \xi(\mathbf{C})$, $\bar{\mathbf{C}} \leftarrow \mathbf{C} + \mathbf{Z}$
    $\hat{Y}, \overline{Y} \leftarrow f_c(\mathbf{C}), f_c(\bar{\mathbf{C}})$
    $\mathcal{L}_{PNS} \leftarrow$ Calculate loss function according to Eq 13
    $\mathcal{L}_{KL} \leftarrow$ Calculate loss function according to Eq 14
    $\mathcal{L} \leftarrow \mathcal{L}_{PNS} + \alpha \mathcal{L}_{KL} + \beta \mathcal{L}_d$
    Update f_g, f_c, f_{mv} by gradient descent
  end for
end for
return f_g, f_c, ξ

In practice, the discriminator D_θ is trained by minimizing $\mathcal{L}_{\text{d}}$, while the encoder f_g maximizes it using a Gradient Reversal Layer (GRL) (Ganin et al. 2016).

Through this adversarial training process, the encoder f_g is guided to learn a representation $\mathbf{C}$ that not only enables accurate prediction of y by minimizing the sufficiency risk $\hat{\text{SF}}^s$, but also satisfies the causal constraints of necessity and sufficiency by minimizing the monotonicity measurement $\hat{M}^s$. Simultaneously, it obfuscates sensitive information by maximizing the adversarial loss $\mathcal{L}_{\text{d}}$, effectively fooling the sensitive attribute discriminator. In this way, FairSNR achieves the goal of learning a graph representation that is sufficient, necessary, and fair.

Experiments

In this section, we conduct comprehensive experiments to examine the performance of the proposed FairSNR approach, following the experimental protocol established in (Li et al. 2024; Yang et al. 2024). Both fairness and utility metrics are considered across multiple datasets. Our experiments are designed to address the following research questions (RQs):

RQ1: Can FairSNR outperform existing baseline methods in terms of both utility and fairness? **RQ2:** Do necessary

Dataset	Metric	Vanilla GCN	EDITS	NIFTY	FairGNN	FairVGNN	FairINV	FairGB	Fprompt	FairSNR
German	AUC (↑)	**73.49±2.15**	69.41±2.33	68.78±2.69	67.35±2.13	72.12±1.10	69.18±1.62	59.77±7.59	70.25±1.53	66.17±0.59
	F1 (↑)	80.76±2.35	81.55±0.59	81.40±0.50	82.01±0.26	82.14±0.42	82.25±0.42	**82.46±0.23**	81.31±1.33	82.38±0.09
	DP (↓)	33.75±12.34	4.25±3.28	5.73±5.25	3.49±2.15	1.71±1.68	1.15±1.32	1.68±3.30	2.39±1.65	**0.64±0.85**
	EO (↓)	25.73±8.36	3.87±2.23	5.08±4.29	3.40±2.15	1.21±2.11	0.90±1.81	1.08±1.80	2.41±1.83	**0.36±0.71**
Credit	AUC (↑)	**73.80±0.23**	73.01±0.11	71.96±0.19	71.95±1.43	67.34±0.45	69.76±3.24	72.02±1.53	70.21±1.33	68.34±5.84
	F1 (↑)	82.63±0.21	81.81±0.28	81.72±0.05	81.84±1.19	81.08±0.74	80.42±2.64	85.43±3.34	80.35±1.92	**85.78±2.44**
	DP (↓)	12.53±0.25	10.90±1.22	11.68±0.07	12.64±2.11	5.02±5.22	6.05±4.37	2.30±3.00	4.22±1.65	**1.71±1.72**
	EO (↓)	10.63±0.02	8.75±1.21	9.39±0.07	10.41±2.03	3.60±4.31	4.00±3.87	1.75±2.07	3.45±2.19	**0.73±0.67**
Pokec-z	AUC (↑)	75.42±0.33	OOM	71.59±0.17	76.12±0.12	76.02±0.16	75.79±0.08	OOM	76.03±1.77	**76.15±0.68**
	F1 (↑)	70.32±0.20	OOM	67.13±1.66	69.75±2.65	70.45±0.57	70.78±0.50	OOM	70.89±0.53	**71.17±0.54**
	DP (↓)	7.02±1.38	OOM	3.06±1.85	2.73±2.23	2.90±0.77	2.70±0.96	OOM	1.47±0.54	**0.93±0.46**
	EO (↓)	7.60±1.24	OOM	3.86±1.65	2.17±1.85	3.09±0.97	2.23±0.66	OOM	1.38±0.68	**0.98±0.94**
Pokec-n	AUC (↑)	74.87±0.18	OOM	69.23±0.56	73.49±0.28	73.73±0.92	73.55±0.16	OOM	75.83±1.32	**77.27±0.50**
	F1 (↑)	65.35±0.54	OOM	61.75±1.05	64.80±0.89	63.35±1.64	65.19±0.62	OOM	67.39±0.52	**68.27±0.96**
	DP (↓)	7.17±1.46	OOM	6.96±1.80	2.26±1.19	4.28±1.33	1.24±0.64	OOM	1.23±1.09	**1.12±1.25**
	EO (↓)	5.66±0.43	OOM	7.75±1.43	3.21±2.28	5.34±1.27	2.80±0.78	OOM	**1.03±1.29**	1.69±1.97

Table 1: Model performance on the German, Credit, Pokec-z, and Pokec-n datasets in terms of utility and fairness. Bold indicates the best results for each metric, while underline denotes the runner-up results. Arrows (↑/↓) indicate whether higher or lower values are preferable. Each result is averaged over five independent runs. All models use GCN as the backbone encoder.

and sufficient representations enhance model performance? **RQ3:** How do hyperparameters influence the performance of FairSNR?

Datasets and Implementation Details

We evaluate our FairSNR on four widely used real-world datasets: **German** (Asuncion and Newman 2007), **Credit** (Yeh and Lien 2009), as well as two variants of the Pokec dataset, **Pokec_n** and **Pokec_z** (Takac and Zabovsky 2012). The key statistics of these datasets are summarized in the appendix.

Baselines

We compared the performance of FairSNR with seven baseline methods across three backbone architectures. FairGNN (Dai and Wang 2021), EDITS (Dong et al. 2022), NIFTY (Agarwal, Lakkaraju, and Zitnik 2021), FairVGNN (Wang et al. 2022), FairINV (Zhu et al. 2024a), FairGB (Li et al. 2024), FPrompt (Li et al. 2025b). The details of the baselines can be found in the appendix.

Evaluation Metrics

To evaluate the performance of the downstream classification task, we adopt AUC and F1 score as the primary utility metrics. These measures provide a comprehensive view of model effectiveness, particularly in imbalanced classification scenarios. For fairness evaluation, we employ two widely used group fairness criteria: Demographic Parity (DP) (Dwork et al. 2012) and Equal Opportunity (EO) (Hardt, Price, and Srebro 2016)

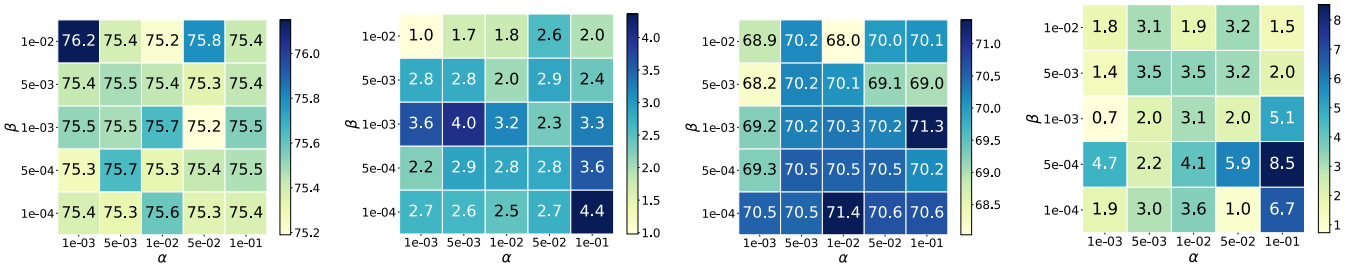
GNN Backbones

In our experimental setup, we utilize three popular GNN models as the foundation of our encoder: GCN (Kipf and Welling 2017), GIN(Xu et al. 2019), and GraphSAGE (Hamilton, Ying, and Leskovec 2017). These architectures are broadly recognized in the research community and have shown robust effectiveness across a range of graph-based learning tasks. The details of the GIN and GraphSAGE baseline can be found in the appendix.

Comparison Results (RQ1)

We conduct a comprehensive comparison between our proposed FairSNR model and several state-of-the-art fair GNN methods on four publicly available benchmark datasets. As shown in Table 1, across all four datasets, FairSNR consistently achieves the lowest fairness disparities. For instance, on the **German** dataset, FairSNR reduces the DP and EO by **44.35%** and **60%**, respectively, compared to the best-performing baseline method in terms of fairness. We attribute this performance gain to the core idea of our model, which is to learn graph representations that are both necessary and sufficient. This learning paradigm explicitly encourages the model to identify and retain only those features that have a direct causal relationship with the prediction task while actively discarding all spurious correlations. In fairness-sensitive scenarios, the influence of sensitive attributes on prediction outcomes is often mediated through spurious correlations. By discovering "necessary" representations, FairSNR fundamentally severs this connection, thereby producing predictions that are invariant to sensitive attributes and inherently fair.

On larger and structurally more complex datasets such as **Pokec-z** and **Pokec-n**, FairSNR not only leads in fairness metrics but also consistently outperforms all baseline models in predictive utility (AUC and F1). For instance, on the Pokec-n dataset, FairSNR achieves a relative improvement of approximately 1.90% in AUC and 1.31% in F1 score compared to the best-performing baseline. This result indicates that, as the graph structure becomes more complex, models that indiscriminately learn all correlations may become entangled in a multitude of spurious associations. In contrast, FairSNR leverages its causality-aware learning objective to more precisely extract true predictive signals, thus



AUC (Pokec-z) ΔDP (Pokec-z) AUC (Credit) ΔDP (Credit)

Figure 3: Hyper-parameter analysis on **Pokec-z** and **Credit**.

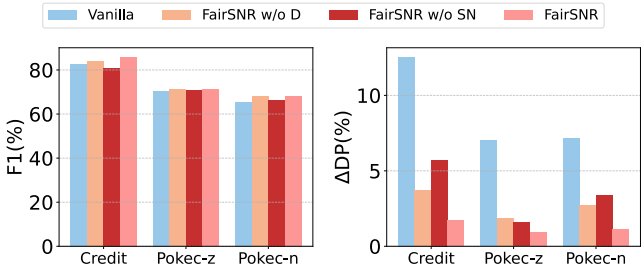


Figure 4: Ablation study of FairSNR showing the trade-off between F1 score and DP.

minimizing overfitting to noise and bias. On relatively small datasets, FairSNR achieves slightly lower AUC scores than some baselines. This can be attributed to the trade-off between accuracy and fairness, as the model strictly enforces the necessity constraint. In such datasets, features correlated with sensitive attributes may also be strongly linked to the target label, despite lacking genuine causal relevance.

Ablation Study (RQ2)

To answer the research question **RQ2** and evaluate the effectiveness of our proposed FairSNR, we construct two ablated variants of the model. **w/o D** denotes the version without the discriminator, while **w/o SN** refers to the version without the necessity and sufficiency constraints. We observe that both variants perform worse than the complete FairSNR model in balancing utility and fairness, which demonstrates the effectiveness of each individual component and the soundness of their integration. The details of the ablation study can be found in the appendix. In the **w/o SN** setting, although adversarial learning can partially mitigate bias, it leads to a noticeable drop in predictive performance. On the other hand, the **w/o D** variant generally achieves better fairness than **w/o SN**, but still falls short of the full FairSNR. This indicates that learning necessary and sufficient representations serves as an effective fairness intervention. It weakens major spurious correlation pathways while retaining essential predictive information, thus laying a strong foundation for achieving both fairness and utility, as shown in Figure 4. Although the two components may exhibit varying strengths across different tasks, we observe that FairSNR consistently benefits from their complementary effects, demonstrating a synergistic gain when the two modules are combined.

Hyper-parameter Analysis (RQ3)

In this study, we investigate the impact of two key hyperparameters in FairSNR on model performance and fairness: the KL regularization strength α and the adversarial learning strength β. More precisely, α controls the alignment between the representation space and the prior distribution, thereby affecting the degree of compression and generalization of the learned representations. In contrast, β determines the extent to which sensitive attribute information is suppressed during training. To systematically assess the effects of these two hyperparameters, we conduct a grid search on the **Credit** and **Pokec-z** datasets with $\alpha \in \{0.0001, 0.0005, 0.001, 0.005, 0.01\}$ and $\beta \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$. The results are summarized in Figures 3. Overall, a too-small α fails to adequately regularize the representation space, while an excessively large α suppresses the expressiveness needed for accurate prediction. Meanwhile, a moderate adversarial strength (e.g., $\beta = 0.01$) effectively disrupts the influence paths of sensitive attributes, thereby enhancing fairness. Based on these findings, we adopt $\alpha = 0.001$ and $\beta = 0.01$ as default settings in all subsequent experiments to ensure a good trade-off between utility and fairness.

Conclusion

In this paper, we propose a novel approach, FairSNR, to address the problem of unfairness in GNN representation learning. FairSNR consists of two synergistic modules that jointly learn necessary and sufficient fair representations in graphs. Guided by principles from causal theory, FairSNR mitigates bias by enforcing both necessity and sufficiency constraints on node representations. Additionally, it incorporates adversarial disentanglement to eliminate residual sensitive attribute information embedded in the learned representations. Experimental results on four real-world benchmark datasets demonstrate that FairSNR achieves SOTA performance in balancing predictive utility and group fairness. In future work, greater attention could be devoted to the fundamental challenge of fair representation learning, with particular emphasis on leveraging graph structural properties to design more effective bias mitigation strategies.

Acknowledgements

This work was partially supported by the Specific Research Project of Guangxi for Research Bases and Talents(GuiKe AD24010011) and the Key Research & Development Program Project of Guangxi (GuiKe AB25069095) . We also wish to acknowledge the support from the Australian Research Council (under grant DP230101122).

References

Agarwal, C.; Lakkaraju, H.; and Zitnik, M. 2021. Towards a unified framework for fair and stable graph representation learning. In *Uncertainty in artificial intelligence*, 2114–2124. PMLR.

Asuncion, A.; and Newman, D. 2007. UCI machine learning repository.

Chen, X.; Cai, R.; Zheng, K.; Jiang, Z.; Huang, Z.; Hao, Z.; and Li, Z. 2025. Unifying invariant and variant features for graph out-of-distribution via probability of necessity and sufficiency. *Neural Networks*, 107044.

Dai, E.; and Wang, S. 2021. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *Proceedings of the 14th ACM international conference on web search and data mining*, 680–688.

Dai, E.; Zhao, T.; Zhu, H.; Xu, J.; Guo, Z.; Liu, H.; Tang, J.; and Wang, S. 2024. A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability. *Machine Intelligence Research*, 1011–1061.

Dong, Y.; Liu, N.; Jalaian, B.; and Li, J. 2022. Edits: Modeling and mitigating data bias for graph neural networks. In *Proceedings of the ACM web conference*, 1259–1269.

Dong, Y.; Ma, J.; Wang, S.; Chen, C.; and Li, J. 2023. Fairness in graph mining: A survey. *IEEE Transactions on Knowledge and Data Engineering*, (10): 10583–10602.

Dwork, C.; Hardt, M.; Pitassi, T.; et al. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, (59): 1–35.

Guo, Z.; Li, J.; Xiao, T.; Ma, Y.; and Wang, S. 2023. Towards fair graph neural networks via graph counterfactual. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, 669–678.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*.

Job, S.; Tao, X.; Cai, T.; Li, L.; Xie, H.; Xu, C.; and Yong, J. 2025. HebCGNN: Hebbian-enabled causal classification integrating dynamic impact valuing. *Knowledge-Based Systems*, 113094.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR, Toulon, France, April 24-26*.

Lei, S.; Chang, X.; Yu, Z.; He, D.; Huo, C.; Wang, J.; and Jin, D. 2025. Feature-Structure Adaptive Completion Graph Neural Network for Cold-start Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 12022–12030.

Li, B.; and Nabavi, S. 2024. A multimodal graph neural network framework for cancer molecular subtype classification. *BMC bioinformatics*, 25(1): 27.

Li, X.; Li, W.; Yu, X.; Han, Z.; and Jin, Q. 2025a. Financial risk assessment of imbalanced data based on nonlinear causal time-series network. *Information Processing & Management*, 62(3): 104025.

Li, Z.; Dong, Y.; Liu, Q.; and Yu, J. X. 2024. Rethinking fair graph neural networks from re-balancing. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1736–1745.

Li, Z.; Lin, M.; Wang, J.; and Wang, S. 2025b. Fairness-aware prompt tuning for graph neural networks. In *Proceedings of the ACM on Web Conference 2025*, 3586–3597.

Luo, R.; Huang, H.; Lee, I.; Xu, C.; Qi, J.; and Xia, F. 2025. Fairgp: A scalable and fair graph transformer using graph partitioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 12319–12327.

Ouyang, Z.; Zhang, C.; Hou, S.; Zhang, C.; and Ye, Y. 2024. How to improve representation alignment and uniformity in graph-based collaborative filtering? In *Proceedings of the International AAAI Conference on Web and Social Media*, 1148–1159.

Pearl, J. 2009. *Causality*. Cambridge university press.

Rahman, T. A.; Surma, B.; Backes, M.; and Zhang, Y. 2019. Fairwalk: Towards Fair Graph Embedding. In Kraus, S., ed., *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI, Macao, China, August 10-16*, 3289–3295.

Spinelli, I.; Scardapane, S.; Hussain, A.; and Uncini, A. 2021. Fairdrop: Biased edge dropout for enhancing fairness in graph representation learning. *IEEE Transactions on Artificial Intelligence*, (3): 344–354.

Takac, L.; and Zabovsky, M. 2012. Data analysis in public social networks. In *International scientific conference and international workshop present day trends of innovations*, 6.

Tian, J.; and Pearl, J. 2000. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1): 287–313.

Trinh, T. K.; and Zhang, D. 2024. Algorithmic Fairness in Financial Decision-Making: Detection and Mitigation of Bias in Credit Scoring Applications. *Journal of Advanced Computing Systems*, 36–49.

Wang, Y.; Zhao, Y.; Dong, Y.; Chen, H.; Li, J.; and Derr, T. 2022. Improving fairness in graph neural networks via mitigating sensitive attribute leakage. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 1938–1948.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2019. How Powerful are Graph Neural Networks? In *7th International Conference on Learning Representations, ICLR New Orleans, LA, USA, May 6-9,*.

Yang, C.; Liu, J.; Yan, Y.; and Shi, C. 2024. Fairsin: Achieving fairness in graph neural networks through sensitive information neutralization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 9241–9249.

Yang, M.; Fang, Z.; Zhang, Y.; Du, Y.; Liu, F.; Ton, J.-F.; Wang, J.; and Wang, J. 2023. Invariant learning via probability of sufficient and necessary causes. *Advances in Neural Information Processing Systems*, 79832–79857.

Yeh, I.-C.; and Lien, C.-h. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, (2): 2473–2480.

Zhu, Y.; Li, J.; Bian, Y.; Zheng, Z.; and Chen, L. 2024a. One Fits All: Learning Fair Graph Neural Networks for Various Sensitive Attributes. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4688–4699.

Zhu, Y.; Li, J.; Zheng, Z.; and Chen, L. 2024b. Fair Graph Representation Learning via Sensitive Attribute Disentanglement. In *Proceedings of the ACM Web Conference 2024*, 1182–1192.