

Disentangled contrastive learning for fair graph representations

Guixian Zhang^{a,b,c}, Guan Yuan^{a,b,c,*}, Debo Cheng^d, Lin Liu^d, Jiuyong Li^d and Shichao Zhang^{e,*}

^aSchool of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu, 221116, China

^bEngineering Research Center of Mine Digitalization, China University of Mining and Technology, Xuzhou, Jiangsu, 221116, China

^cArtificial Intelligence Research Institute, China University of Mining and Technology, Xuzhou, Jiangsu, 221116, China

^dUniSA STEM, University of South Australia, Adelaide, SA, 5095, Australia

^eGuangxi Key Lab of Multi-Source Information Mining & Security, Guangxi Normal University, Guilin, Guangxi, 541004, China

ARTICLE INFO

Keywords:

Fair representation learning

Graph neural network

Fairness

Causality-inspired machine learning

ABSTRACT

Graph Neural Networks (GNNs) play a key role in efficiently learning node representations of graph-structured data through message passing, but their predictions are often correlated with sensitive attributes and thus lead to potential discrimination against some groups. Given the increasingly widespread applications of GNNs, solutions are required urgently to prevent algorithmic discrimination associated with GNNs, to protect the rights of vulnerable groups and to build trustworthy artificial intelligence. To learn the fair node representations of graphs, we propose a novel framework, the **Fair Disentangled Graph Neural Network (FDGNN)**. With the proposed FDGNN framework, we enhance data diversity by generating instances that have identical sensitivity values but different adjacency matrices through data augmentation. Additionally, we design a counterfactual augmentation strategy for constructing instances with varying sensitive values while preserving the same adjacency matrices, thereby balancing the distribution of sensitive values across different groups. Subsequently, we employ a disentangled contrastive learning strategy to acquire disentangled representations of non-sensitive attributes such that sensitive information does not affect the prediction of node information. Finally, the learned fair representations of non-sensitive attributes are employed for building a fair predictive model. Extensive experiments on three real-world datasets demonstrate that FDGNN yields the best fairness predictions compared to the baseline methods. Additionally, the results demonstrate the potential of disentanglement in learning fair representations.


1. Introduction

In many real-world applications, data exists in the form of graph-structure within a non-Euclidean space, such as social networks (Fang et al., 2023), and citation networks (Wu et al., 2024). Graphs have varying numbers of nodes, and the nodes of a graph have different numbers of neighbors, making it difficult to handle the data by normal deep learning methods. Graph Neural Networks (GNNs) have emerged as a powerful approach for processing graph-structured data, gaining popularity in various areas, e.g., human resources management (Hang et al., 2022), smart education (Li et al., 2024), healthcare (Xu et al., 2024), and recommendation (Cai et al., 2023).

GNNs leverage both the topological structure of graphs and the node features to smoothly propagate information over graph edges, ultimately mapping nodes into a meaningful embedding space (Chen et al., 2022b; Liu et al., 2023; Wan et al., 2024). In a GNN, each node aggregates information from its neighbors, combining its features with the collected information to compute a new representation. This enables GNNs to learn node representations that capture both local neighborhood information and global graph topology information (Kipf and Welling, 2016a; Hamilton et al., 2017; Gasteiger et al., 2019; Guan et al., 2024; Wang et al., 2024). Despite the success of GNNs in processing graph-structured data and learning useful representations for downstream tasks, existing research on GNNs mostly ignores the discriminatory bias caused by sensitive attributes (Ma et al., 2022). Consequently, GNN models can lead to discriminatory results, thus limiting their deployment in human-centered real-world applications, where fairness has become a key part of trustworthy artificial intelligence (Kaur et al., 2022; Ling et al., 2024).

It is challenging to address discrimination caused by sensitive attributes in graph data (Dong et al., 2022; Wang et al., 2022). GNNs learn representations from historical data, which inadvertently incorporates past biases in the learned representations. In real-world scenarios, nodes with identical sensitive values often exhibit a tendency to cluster together (Wu et al., 2019; Nyhan et al., 2023). To obtain fair decision-making, sensitive attributes should not affect

*Corresponding authors.

 yuanguan@cumt.edu.cn (G. Yuan); zhangsc@mailbox.gxnu.edu.cn (S. Zhang)

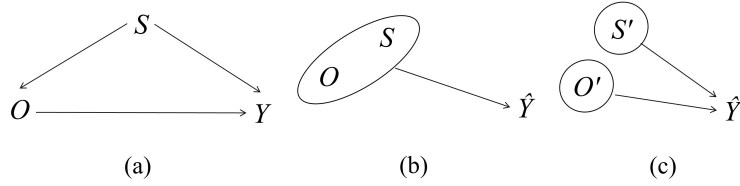


Figure 1: Three causal DAGs illustrating the problem studied in this work. ‘ \rightarrow ’ indicates a causal relation. O , S , and Y are a set of non-sensitive attributes, a set of sensitive attributes, and a target variable, respectively. Circled nodes indicate representations of the variables. O' (S') represents O (S) whose values have been augmented by data augmentation (counterfactual augmentation). (a) The DAG represents the data generation mechanism used in this work. (b) The DAG represents the representation learned from all attributes (O, S) by the current representation learning methods. A model built on the representation uses sensitive information. (c) The representations of O' and S' are independent, and our proposed method only uses O' for prediction \hat{Y} .

a model’s predictions. Fair representation learning aims to acquire representations that accurately model the target variable while remaining insensitive to sensitive attributes (Ma et al., 2023). For example, fair representation assists in predicting recidivism rates without regard to age or gender (Jordan and Freiberger, 2015).

However, owing to the influence of the neighbor aggregation mechanism of GNNs, these adjacent nodes with similar sensitive attributes are prone to be portrayed similarly by GNN models (Xhonneux et al., 2020; Agarwal et al., 2021). On the other hand, sensitive attributes may have already influenced the early stages of data collection and preprocessing, indicating that these attributes are inherently embedded in the data (Dong et al., 2022).

We use the three causal DAGs (directed acyclic graphs) (Pearl, 2009) in Fig. 1 to illustrate the difference between existing methods and our proposed disentangled contrastive learning. In Fig. 1 (a), the causal DAG represents the data generation mechanism of the problem studied in this work, where variables O , S , and Y are pairwise dependent. For example, researchers have found that African Americans and Hispanics or Latinos are more likely to test positive for COVID-19 (Martinez et al., 2020) and to be hospitalised or die (Price-Haywood et al., 2020) than non-Hispanic whites. In this case, it is critical to clarify whether this disparity stems from race or medical condition, which would imply a completely different approach to dealing with the epidemic. However, a major factor contributing to this disparity is the disproportionate access to medical care for patients of different races, which may be influenced by economic status (Chen et al., 2022a) and food insecurity (Wolfson and Leung, 2020). In Fig. 1 (a), this leads to the causal edge from $S \rightarrow O$.

There exist many excellent disentanglement algorithms (Li et al., 2021, 2022; Mo et al., 2023) proposed to improve the performance of GNN, but they do not consider fairness. Specifically, these methods align and fuse subspace representations separately and are not applicable to fair representation learning. In Fig. 1 (b), the causal DAG illustrates that these GNN models utilize the entire set of attributes (O, S) to learn a node representation, potentially incorporating sensitive attribute information (Creager et al., 2019; Park et al., 2021; Oh et al., 2022). Following previous work (Chen et al., 2022c), we can describe the generation process of \hat{Y} as $\hat{Y} := f_{GNN}(O, S)$, where f_{GNN} is a GNN method. Owing to the influence of the neighbor aggregation mechanism of GNNs, neighboring nodes with similar sensitive attributes are prone to be portrayed similarly by GNN models (Xhonneux et al., 2020; Agarwal et al., 2021). On the other hand, sensitive attributes may have already influenced the early stages of data collection and preprocessing, indicating that these attributes are inherently embedded in the data (Dong et al., 2022). The sensitive attributes S result in discrimination bias when (O, S) is used to predict the target Y in a GNN model. We aim to utilize disentangled contrastive learning to separate sensitive information from non-sensitive information, thereby ensuring fairness in node representation learning. Different from these works, our proposed method is based on causal DAG as shown in Fig. 1 (c).

To obtain a fair node representation, it is important to remove the effect of S on O instead of simply using O for prediction. To achieve this goal, we propose a novel method called **Fair Disentangled Graph Neural Network** (FDGNN) to achieve fair node representation for fair decision-making. In this work, we use the idea of matching in causal inference (Pearl, 2009; Hernán and Robins, 2010; Stuart, 2010), which is a well-known method in causal inference for removing bias, allowing different types of confounders to be balanced in distribution. First of all, FDGNN enriches the data and balances the distribution of sensitive attribute values across groups through data augmentation and counterfactual augmentation. Additionally, FDGNN incorporates node degree information into the node features

to capture the structure of the graph and distinguish between nodes. Subsequently, FDGNN utilises a disentangled contrastive learning method (i.e. borrowing the idea from the matching strategy used in causal inference (Pearl, 2009; Hernán and Robins, 2010)) to bring the non-sensitive information of the sample pairs closer together in the feature space, and pull the sensitive information of the nodes farther away from the non-sensitive information to achieve disentanglement. Specifically, the FDGNN controls for the variable O by matching sample pairs so that S does not exhibit systematic differences when predicting \hat{Y} . Our contributions are summarized as follows:

- We present a causal view of the influence of sensitive information, and form a causal graph for fair graph representation learning.
- We develop a novel augmentation strategy that includes both data augmentation and counterfactual augmentation to enrich the data and balance the distribution of non-sensitive features O with respect to the sensitive feature S .
- We propose a novel FDGNN method that utilizes contrastive learning for disentangling representations to achieve fair representations in GNNs. To the best of our knowledge, this is the first work which utilizes disentangled contrastive learning in GNNs for fair representations.
- We conduct extensive experiments on three real-world datasets to demonstrate the performance of FDGNN against state-of-the-art methods, highlighting the promise of disentangled contrastive learning for fair representation learning.

2. Related Work

2.1. Fairness in Graph Neural Networks

Conventional GNN models unintentionally perpetuate algorithmic bias and unfairness, particularly in subpopulations defined by sensitive attributes such as race, gender, and age. To tackle this concern, FairGNN (Dai and Wang, 2022) introduces the use of sensitive attribute estimators to augment sensitive attribute information. It incorporates adversarial debiasing and covariance constraints to regulate GNNs, ensuring fair node representations and predictions, respectively. NIFTY (Agarwal et al., 2021) proposes a novel objective function in conjunction with data augmentation, while EDITS (Dong et al., 2022) enhances model fairness by debiasing attribute and structural information. FVGNN (Wang et al., 2022) improves fairness by masking features to reduce the model’s ability to discriminate based on sensitive attributes. CAF (Guo et al., 2023) enriches the training sample by selecting other data to be used as counterfactuals during training. FairMILE (He et al., 2023) proposes a multi-level framework that fully integrates existing graph embedding methods. FairMigration (Hu et al., 2024) deals with demographic groups based on adversarial learning dynamics.

Prior methods for fair representation learning employ adversarial frameworks to remove sensitive attribute information. However, adversarial learning often leads to unstable training and convergence issues. Second, existing approaches do not fully account for biases that may emerge from dependencies between node attributes in the graph structure. In this paper, we propose a new contrastive learning method to obtain the disentangled representation and apply it to achieve fair predictions by GNN models.

2.2. Data Augmentation in Graph Neural Networks

Data augmentation generates new training instances by artificially introducing perturbations on the training graph, thus expanding the size of the training dataset and enhancing the model’s ability to generalize to different graph structures, including: 1) subgraph sampling (Sun et al., 2021; Zhang et al., 2024b), randomly sampling nodes and their neighbors from the original graph to induce subgraphs; 2) edge dropping (Yang et al., 2022; Zhang et al., 2023), randomly deleting edges from the graph with a certain probability; 3) edge adding (Kong et al., 2022; Zhang et al., 2024a), adding new edges with certain probabilities between pairs of nodes which do not have edges in the original graph; and 4) feature masking (Hou et al., 2022; Feng et al., 2022), partially masking the node features. These operations enhance the robustness of the model by perturbing features to constitute new training instances while maintaining the main topology and patterns of the original graph.

Some graph data augmentation methods have also emerged for fairness in graph neural networks. Ma et al. (Ma et al., 2022) improved fairness by minimizing the difference between the representation learned from the original graph and the counterfactual data augmentation. Zhang et al. (Zhang et al., 2024a) improved the fairness of the model

by mitigating the sensitive attribute leakage of high-degree nodes. Note that the augmentation operation should be tailored to the specific task to avoid potential negative effects. In this paper, we constructed two types of augmentation instances: 1) data augmentation instances with identical node attributes but different adjacencies, and 2) counterfactual augmentation instances with different sensitive attributes but shared non-sensitive attributes and graph structure.

3. Preliminary

3.1. Notations

We use $\mathcal{G} = (\mathbf{V}, \mathbf{E}, \mathbf{A}, \mathbf{X})$ to represent an attribute graph, where $\mathbf{V} = \{v_1, \dots, v_n\}$ is the set of nodes, \mathbf{E} the set of edges, $\mathbf{A} \in \mathbb{R}^{n \times n}$ the adjacency matrix, and $\mathbf{X} \in \mathbb{R}^{n \times d}$ the nodes-attributes matrix in which n is the number of nodes and d is the dimension of attributes. For node v_i , it has a label y and a sensitive attribute, denoted as $s_i \in \{0, 1\}$, and the node representation of v_i is \mathbf{z}_i . Note that \mathbf{z}_s is the node representation of the sensitive attributes, however \mathbf{z}_o initially includes the representation of all attributes of the node (i.e. the sensitive attributes), and removes the sensitive information after disentangled contrastive learning.

3.2. Objectives

Several different definitions of fairness exist in the discussion of fairness, in which group fairness and individual fairness are the two most prominent (Mehrabi et al., 2021). Group fairness advocates for equitable treatment of minority groups (Berk et al., 2021), while individual fairness emphasizes comparable outcomes for similar individuals (Mukherjee et al., 2020). This work focuses on group fairness and evaluates it using Δ_{SP} and Δ_{EO} defined based on statistical parity (Dwork et al., 2012) and equal opportunity (Hardt et al., 2016) respectively.

Definition 1 (Statistical Parity (Dwork et al., 2012)). *Statistical parity stipulates that the proportion of individuals receiving positive classifications is approximately equalized across demographic groups. Formally, statistical parity is defined as:*

$$P(\hat{y} | s = 0) = P(\hat{y} | s = 1), \quad (1)$$

where \hat{y} is the predicted output, and s is the sensitive attribute (e.g. race, gender).

Definition 2 (Equal Opportunity (Hardt et al., 2016)). *Equal opportunity stipulates that the true positive rate should be approximately equal across demographic groups. Formally, equal opportunity is defined as:*

$$P(\hat{y} = 1 | y = 1, s = 0) = P(\hat{y} = 1 | y = 1, s = 1), \quad (2)$$

where \hat{y} is the predicted output, y is the label, and s is the sensitive attribute (e.g. race, gender).

Based on Definition 1 and Definition 2, we can calculate the values of Δ_{SP} and Δ_{EO} as follows:

$$\Delta_{\text{SP}} = |P(\hat{y} = 1 | s = 0) - P(\hat{y} = 1 | s = 1)|, \quad (3)$$

$$\Delta_{\text{EO}} = |P(\hat{y} = 1 | y = 1, s = 0) - P(\hat{y} = 1 | y = 1, s = 1)|, \quad (4)$$

where Δ_{SP} measures statistical parity differences between two groups, while Δ_{EO} evaluates equalized opportunity differences. Lower values for both metrics indicate greater fairness in the model predictions across groups.

To achieve the fair node representation, we propose a novel disentangled contrastive learning method borrowing the concept from the matching strategy in causal inference (Stuart, 2010). Matching individuals from the control group with those in the treatment group who have identical or opposite covariates aims to balance the distribution of covariates across both groups. The matching strategy effectively neutralizes the bias of covariates on the outcome.

The FDGNN attains disentangled representations via contrastive learning that utilizes sample pairs. Specifically, it controls for O , ensuring that S does not exhibit systematic differences between the two sets of predictions, Y .

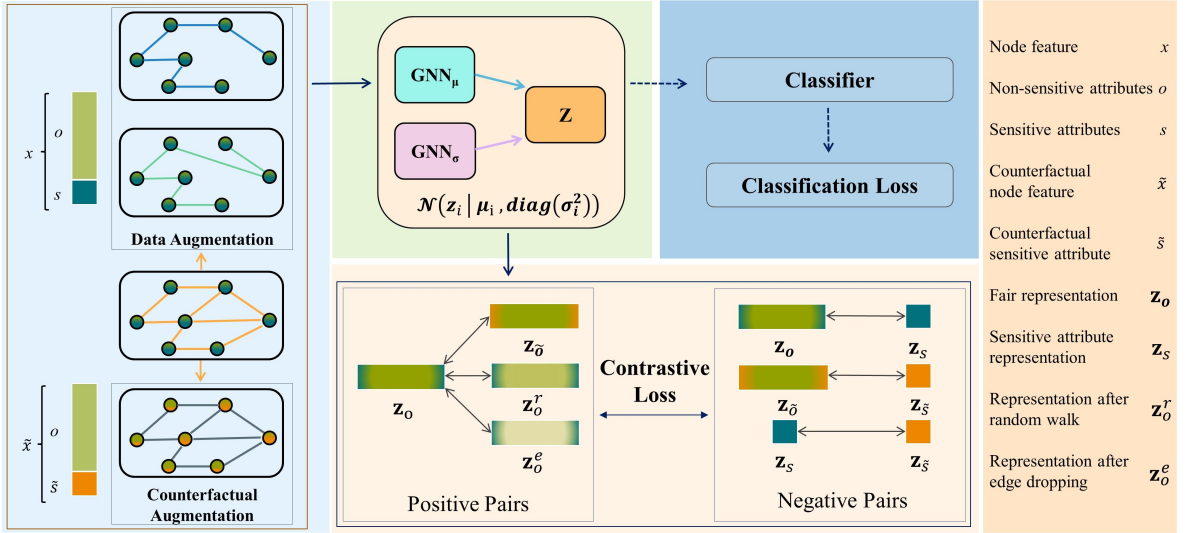


Figure 2: Architecture of the proposed FDGNN method. Solid arrows in the figure represent the pre-training process and dashed arrows represent the steps of the fine-tuning process. For data-augmented instance x , the node attributes consist of the non-sensitive attributes o and the sensitive attribute s . For counterfactual-augmented instance \tilde{x} , the node attributes consist of the sensitive attribute $\tilde{s} = 1 - s$ and the non-sensitive attributes o . Note that in data augmentation, \mathbf{z}_o and \mathbf{z}_s represent the fair node representation and sensitive attribute representation, respectively. In the context of counterfactual augmentation, we use $\mathbf{z}_{\tilde{o}}$ and $\mathbf{z}_{\tilde{s}}$ to indicate the node representation and sensitive attribute representation, respectively.

4. Methodology

Our proposed FDGNN framework aims to learn two distinct representations \mathbf{z}_o and \mathbf{z}_s such that \mathbf{z}_s and \mathbf{z}_o are independent. When the classification model is constructed using \mathbf{z}_o , the representation \mathbf{z}_s will not affect \hat{y} . Specifically, FDGNN attempts to establish a mapping that remains invariant to \mathbf{z}_s . By isolating sensitive aspects into an independent representation \mathbf{z}_s , the model can focus the primary representation, \mathbf{z}_o , on non-sensitive attributes relevant to the task. In this section, we will introduce the encoder and contrastive learning settings within the FDGNN framework, respectively.

4.1. Learning the Latent Representation by VGAE

In this work, our proposed FDGNN framework utilizes the Variational Graph Autoencoder (VGAE) (Kipf and Welling, 2016b) model as an encoder to learn disentangled representations of the graph structure. A key advantage of the VGAE for FDGNN is that its learned latent distribution encourages the separation of the underlying explanatory factors within the graph data. The variational autoencoder (VAE) framework imposes a Gaussian prior on the latent distribution. Regularizing the encoder to match this simple Gaussian, encourages independence between the latent representations. Specifically, the formula of VGAE is defined as follows:

$$q(\mathbf{z}_i^* | \mathbf{X}, \mathbf{A}) = \mathcal{N}(\mathbf{z}_i^* | \mu_i^*, \text{diag}((\sigma_i^*)^2)), \quad (5)$$

where $\mu^* = \text{GNN}_{\mu}^*(\mathbf{X}, \mathbf{A})$ and μ_i^* is the mean value of the node representation, $\log \sigma^* = \text{GNN}_{\sigma}^*(\mathbf{X}, \mathbf{A})$ and σ_i^* is the variance of the node representation, $\mathbf{z}_i^* = \mu^* + \sigma_i^* * \epsilon$ and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. When $*$ is r or d , it represents the representation after random walk operation or edge dropping operation respectively.

In our FDGNN framework, we incorporate each node's degree of the graph data as an additional attribute to enrich the node representation during the training process. The explicit degree feature enhances the difference between nodes, consequently augmenting the model's predictive capability for individual nodes. Formally, we define the node feature matrix as:

$$\mathbf{X} = [\bar{\mathbf{X}}, \bar{D}], \quad (6)$$

where $\bar{\mathbf{X}} \in \mathbb{R}^{n \times d}$ is the original feature matrix and $\bar{\mathbf{D}} \in \mathbb{R}^{n \times 1}$ is a vector of node degrees. n is the number of nodes and d is the dimension of attributes.

We can utilize the encoder to get the node representation \mathbf{Z}_O and sensitive attribute representation \mathbf{Z}_S :

$$f_{\Theta_e}(\mathbf{A}, \mathbf{X}) \rightarrow \mathbf{Z}_O, \quad (7)$$

$$f_{\Theta_e}(\mathbf{A}, \mathbf{S}) \rightarrow \mathbf{Z}_S, \quad (8)$$

where $\mathbf{S} \in \mathbb{R}^{n \times 1}$ represents the sensitive attribute of the nodes and f_{Θ_e} represents the encoder. Note that the two representations share an encoder to keep their representations in the same feature space (You et al., 2020). To clearly describe the construction of the loss function, we will use \mathbf{z}_o and \mathbf{z}_s instead of \mathbf{Z}_O and \mathbf{Z}_S to denote the node representation and the sensitive attribute representation of a given node x , respectively.

4.2. Learning Fair Disentangled Representations via Contrastive Learning

FDGNN comprises a self-supervised pre-training phase that employs contrastive learning, followed by fine-tuning through the utilization of a Multi-Layer Perceptron (MLP) as a classifier. To improve the model's ability to discriminate between non-sensitive and sensitive attributes, FDGNN constructs two types of augmentations: 1) data augmentation instances with identical node attributes but different adjacencies, and 2) counterfactual augmentation instances with different sensitive attributes but shared non-sensitive attributes and graph structure.

For a node with feature x , we obtain its counterfactual sample by flipping the value of the sensitive attribute s to \bar{s} and retaining the value of the non-sensitive attribute o (Ma et al., 2022, 2023). Similar to Eq. (7) and Eq. (8), we can obtain the node representation $\mathbf{z}_{\bar{o}}$ and sensitive attribute representation $\mathbf{z}_{\bar{s}}$ of the counterfactual augmented instances. To improve the model's understanding of the data, we generate different adjacency matrices via two data augmentation operations: random walk and droppedge. For random walk augmentation, we can get $\mathbf{A}'_r = \mathbf{A}[\mathbf{V}', \mathbf{V}']$, where \mathbf{V}' is the subset of nodes from truncated random walks of length l starting from sampled nodes. In this paper, the length l is set to 10. For edge dropping augmentation, $\mathbf{A}'_d = \mathbf{A} \odot \mathbf{B}$, where \mathbf{B} is a binary mask with $\mathbf{B}_{ij} \sim \text{Bernoulli}(1 - p)$ to randomly drop edges. The probability p is set to 0.5. By encoding the node features and the adjacency matrix of the augmented view, we can obtain the node representation \mathbf{z}'_o and sensitive attribute representation \mathbf{z}'_s after the random walk, and node representation \mathbf{z}^d_o and sensitive attribute representation \mathbf{z}^d_s after edge dropping.

Since the node features of the data augmentation instances are not changed, the node representations should be similar after different data augmentation operations. To learn accurate node representations for accuracy, FDGNN should minimize distances between node representations. So we can get positive pairs: $(\mathbf{z}_o, \mathbf{z}_{\bar{o}})$, $(\mathbf{z}_o, \mathbf{z}^d_o)$, and $(\mathbf{z}_o, \mathbf{z}'_o)$. To pursue fairness, node representations should be as independent of sensitive attributes as possible, i.e., the distance between node representations and sensitive attribute representations of each sample should be as maximized as possible. Meanwhile, to ensure the discriminative ability for sensitive information, FDGNN should accurately distinguish between sensitive attribute representation \mathbf{z}_s of the original graph and sensitive attribute representation $\mathbf{z}_{\bar{s}}$ of counterfactual augmentation sample, i.e., maximize the distance between them. So we can get the negative sample pairs: $(\mathbf{z}_o, \mathbf{z}_s)$, $(\mathbf{z}_{\bar{o}}, \mathbf{z}_{\bar{s}})$, and $(\mathbf{z}_s, \mathbf{z}_{\bar{s}})$.

To calculate the distance between different sample pairs, we first use an exponential transformation of the cosine similarity against the two instances to make the variation more pronounced, i.e. $M(z_1, z_2) = \exp(\frac{z_1^T z_2}{\|z_1\| \|z_2\|}) / \tau$, where τ is a temperature parameter. Subsequently we can obtain the distance of the sample pairs $\mathcal{L}(\mathbf{z}_i, \tilde{\mathbf{z}}_i) = -\log D$, in which D is defined as:

$$D(\mathbf{z}_i, \tilde{\mathbf{z}}_i) = \frac{M(\mathbf{z}_i, \tilde{\mathbf{z}}_i)}{M(\mathbf{z}_i, \tilde{\mathbf{z}}_i) + \sum_{j=1}^n \mathbb{1}_{[j \neq i]} M(\mathbf{z}_i, \tilde{\mathbf{z}}_j) + \sum_{i=1}^n \mathbb{1}_{[j \neq i]} M(\mathbf{z}_i, \mathbf{z}_j)}, \quad (9)$$

where $\mathbb{1}_{[j \neq i]}$ is the indicator function such that the similarity of the sample itself is excluded.

Thus, the complete loss function can be expressed as:

$$\min_{\Theta_e} \mathcal{L}_E = \{ \mathcal{L}(\mathbf{z}_o, \mathbf{z}_{\bar{o}}) + \mathcal{L}(\mathbf{z}_o, \mathbf{z}^d_o) + \mathcal{L}(\mathbf{z}_o, \mathbf{z}'_o) \}$$

$$- \left\{ \mathcal{L}(\mathbf{z}_o, \mathbf{z}_s) + \mathcal{L}(\mathbf{z}_{\bar{o}}, \mathbf{z}_{\bar{s}}) + \mathcal{L}(\mathbf{z}_s, \mathbf{z}_{\bar{s}}) \right\}. \quad (10)$$

Essentially, the contrastive loss proposed by FDGNN quantifies the similarity of the instances based on the relative distance between positive and negative pairs. The purpose of this is to separate sensitive information in the node representation to ensure fairness, and to maximize the retention of non-sensitive information to ensure accuracy.

In Theorem 1, we provide a theoretical analysis of the FDGNN, i.e., the fair node representations generated by the FDGNN can achieve statistical parity after the implementation of disentanglement for both non-sensitive and sensitive information.

Theorem 1. *Let \mathbf{z}_o represent a specific node representation learned by the FDGNN model and \mathbf{z}_s represent the representation of the sensitive attribute of this node. The node's sensitive attribute has a value of either 1 or 0. If \mathbf{z}_o and \mathbf{z}_s achieve disentanglement, then the FDGNN method will achieve statistical parity, i.e., $P(\hat{y} = 1 \mid s = 1) = P(\hat{y} = 1 \mid s = 0)$.*

Proof. When \mathbf{z}_o and \mathbf{z}_s achieve disentanglement, we can regard \mathbf{z}_o and \mathbf{z}_s as two independent representations. When s has a value of either 1 or 0, we can formally express this independence as $P(\mathbf{z}_o \mid s = 1) = P(\mathbf{z}_o \mid s = 0)$. Given this independence, we analyse the conditional probabilities of the predicted outcome \hat{y} given the sensitive attribute s :

$$\begin{aligned} P(\hat{y} = 1 \mid s = 1) &= \int_{\mathbf{z}_o} P(\hat{y} = 1 \mid \mathbf{z}_o) P(\mathbf{z}_o \mid s = 1) d\mathbf{z}_o \\ &= \int_{\mathbf{z}_o} P(\hat{y} = 1 \mid \mathbf{z}_o) P(\mathbf{z}_o \mid s = 0) d\mathbf{z}_o \\ &= P(\hat{y} = 1 \mid s = 0). \end{aligned} \quad (11)$$

Therefore, we conclude that if \mathbf{z}_o and \mathbf{z}_s achieve disentanglement, the FDGNN method will achieve statistical parity, ensuring fairness in the node representation. \square

Following self-supervised pre-training, we utilize an MLP classifier applied to the encoder output for final prediction. More formally, let $f(x)$ represent the pre-trained encoder network mapping the input to embeddings. An MLP classifier $g(f(x))$ is then trained to predict targets y from embeddings $f(x)$ by minimizing the loss:

$$\min_{\Theta_c} \mathcal{L}_C = -\mathbb{E}_{v_i \sim \mathcal{V}} (y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)), \quad (12)$$

where y_i is the label of the node v_i and \hat{y}_i is the prediction of the node by the classifier.

5. Experiments

This section presents the results of comprehensive experiments conducted on three real-world datasets. We aim to evaluate the performance of FDGNN, and compare it with state-of-the-art models. Before discussing the experimental results, we will detail the datasets, comparison algorithms used and implementation details.

5.1. Experiment Setup

5.1.1. The Details of Datasets

We employed three well-known datasets, namely the Recidivism, the Credit, and the German datasets (Agarwal et al., 2021; Dai and Wang, 2022; Wang et al., 2022; Dong et al., 2022). The details of these datasets are summarized in Table 1. The **Credit** dataset includes 13 attributes such as marital status, age, and maximum payment amount (Yeh and Lien, 2009), and we use age as a sensitive attribute in our experiments. The **German** dataset includes 27 attributes such as employment status, gender, and income (Dua and Graff, 2017), and we use gender as a sensitive attribute in our experiments. The **Recidivism** dataset includes 18 attributes such as type of case, race, and case duration (Jordan and Freiburger, 2015), and we use race as a sensitive attribute in our experiments. To better analyze the differences between the datasets, we define density as the ratio of the number of edges to the number of nodes to characterize the datasets.

Table 1

A summary of the datasets.

Dataset	Credit	German	Recidivism
#of nodes	30,000	1,000	18,876
#of node attributes	13	27	18
#of edges	1,436,858	22,242	321,308
Sensitive attribute	Age	Gender	Race
Average node degree	95.79	44.48	34.04
Graph density	47.90	22.24	17.02

5.1.2. Baseline

Our experiments compare the proposed FDGNN model against several state-of-the-art fair GNN algorithms. Prior works on improving fairness in graph neural networks can be categorized into two main approaches: adversarial learning-based methods such as FairGNN (Dai and Wang, 2022) and FVGNN (Wang et al., 2022) and augmentation-based methods such as NIFTY (Agarwal et al., 2021), EDITS (Dong et al., 2022) and FairMILE (He et al., 2023). We brief these comparison methods as follows.

- FairGNN (Dai and Wang, 2022) is designed to address bias and discrimination in GNN predictions by leveraging limited sensitive attributes and graph structures.
- NIFTY (Agarwal et al., 2021) is a novel framework that establishes a connection between counterfactual fairness and stability in GNNs, enabling the learning of fair and stable representations.
- EDITS (Dong et al., 2022) is a novel framework that aims to achieve fairer GNNs from both feature and structural perspectives by mitigating biases in the input graph itself.
- FVGNN (Wang et al., 2022) is a novel approach that targets discriminatory bias by effectively addressing the variation in feature correlations during feature propagation through feature masking strategies.
- FairMILE (He et al., 2023) is a novel multi-level framework that learns fair graph representations with fairness constraints.

5.1.3. Evaluation Metrics

The goal of FDGNN is to obtain a fair node representation \mathbf{z}_o . To evaluate the fairness and utility of the learned \mathbf{z}_o , we use Δ_{SP} and Δ_{EO} as the fairness metrics. The lower Δ_{SP} and Δ_{EO} values represent better fairness of the model. Besides, we use accuracy (ACC, %) and F_1 score to evaluate the performance of the prediction model over the learned representation \mathbf{z}_o .

5.1.4. Implementation Details

To ensure the validity of the experimental results, we employ a five-fold cross-validation methodology for each experiment. As different GNN backbones may result in varying levels of sensitive attribute leakage and fairness performance, we evaluate the above algorithms using three GNN backbone models: GCN (Kipf and Welling, 2016a), GraphSAGE (Hamilton et al., 2017), and APPNP (Gasteiger et al., 2019). In this paper, the foundational architecture (i.e., the backbone) of all proposed methods consists of two layers.

In our proposed FDGNN model, we use two linear layers to act as a classifier. For the German and Credit datasets, the learning rate is set to 0.001, and the hidden layer size is 64. We set the number of pre-training and fine-tuning epochs for all model architectures to 10. For the Recidivism dataset, the learning rate is adjusted to 0.03 with a hidden layer size of 128. The epoch settings for different model architectures are as follows: for GCN, 50 pre-training and 30 fine-tuning epochs; for APPNP, 30 pre-training and 10 fine-tuning epochs; and GraphSAGE, 45 pre-training and 15 fine-tuning epochs. The temperature parameter, denoted by τ , is fixed at 0.5.

Table 2

Results of prediction and fairness performance on the Credit dataset. The best results are **bold-faced** and the runner-up results are underlined. Note that FDGNN with SAGE and APPNP obtains the best and runner-up fairness results, respectively, and has a similar F_1 score as that of the best predictive model.

Dataset	Credit			
Method	ACC	F_1	Δ_{SP}	Δ_{EO}
GCN	73.62±0.06	81.88±0.06	12.93±0.26	10.65±0.18
FairGCN	73.27±1.16	81.73±1.15	7.89±3.32	6.78±4.06
NI-GCN	72.15±2.44	81.54±0.04	5.94±0.77	9.47±0.08
FV-GCN	78.44±0.41	87.57±0.09	3.37±2.57	1.59±1.30
ED-GCN	<u>81.12±4.28</u>	<u>77.73±2.35</u>	8.64±2.57	6.30±2.32
MI-GCN	80.32±0.26	87.24±0.16	1.03±0.51	0.79±0.53
FD-GCN	77.71±0.13	87.45±0.09	0.44±0.17	0.56±0.21
APPNP	74.73±0.30	82.83±0.20	15.00±0.55	12.85±0.60
FairAPPNP	73.16±1.02	81.58±0.78	14.43±3.17	11.88±3.27
NI-APPNP	72.91±0.97	83.76±0.34	8.65±3.97	9.14±2.19
FV-APPNP	78.44±0.49	87.69±0.12	2.40±2.54	1.24±1.33
ED-APPNP	79.61±2.78	77.71±2.39	8.75±2.75	5.81±2.76
MI-APPNP	80.27±0.30	87.22±0.18	1.41±0.58	0.94±0.21
FD-APPNP	77.79±0.08	87.50±0.06	<u>0.40±0.22</u>	<u>0.34±0.15</u>
SAGE	74.20±0.60	82.45±0.52	16.35±2.36	14.12±2.64
FairSAGE	75.44±3.28	81.35±1.83	10.46±5.69	9.47±6.10
NI-SAGE	73.80±4.75	81.21±0.59	8.09±2.77	7.41±1.54
FV-SAGE	76.06±4.37	84.43±4.23	6.06±3.63	3.90±3.54
ED-SAGE	83.73±0.73	76.93±0.89	7.28±0.49	5.09±0.78
MI-SAGE	80.18±0.27	87.16±0.17	1.21±0.39	0.84±0.14
FD-SAGE	77.87±0.02	87.56±0.01	0.24±0.14	0.30±0.18

5.2. Evaluation of the FDGNN

5.2.1. Results on Effectiveness

The results in Table 2, Table 3 and Table 4 demonstrate the competitive performance of FDGNN on both node classification and fairness metrics, validating its effectiveness in learning fair representations. The results of the comparative evaluation show that our proposed method obtains the most optimum metric values. Specifically, FDGNN achieves optimal values on the fairness metrics Δ_{SP} and Δ_{EO} for the German and Credit datasets, outperforming other recent methods.

However, we observe that FDGNN does not perform as well on the Recidivism dataset, which has the lowest graph density defined as the ratio of edges to nodes. We posit that the lower connectivity impedes the VGAE encoder in FDGNN from adequately capturing node relationships in the input features. This likely impacts the latent space distribution learning accuracy. Moreover, with fewer edges, the graph data augmentations fail to provide sufficient differentiation or useful positive sample information. The sparser connectivity and feature interactions of Recidivism pose challenges for our current data augmentation and sampling approach. Future work should explore more advanced data augmentation and sampling techniques tailored to sparse graph structures to further enhance fairness across varying graph densities.

5.2.2. Results on Efficiency

We benchmark FDGNN on the training speeds against the state-of-the-art methods with GCN serving as the backbone on a Tesla V100 GPU. The results reported in Table 5 show that FDGNN achieves significantly faster training speeds compared to the methods, particularly for dense graphs.

The time complexity of FDGNN is analyzed as follows: Consider a graph with $|V|$ nodes and $|E|$ edges. Let t denote the number of epochs, p and q denote the keep probabilities of two data augmentation steps. Since we generate two independent sub-graphs per epoch, given the fact that the number of non-zero elements in the adjacency matrices of the full training graph and the two sub-graphs are $2 * |E|$, $2 * p * |E|$ and $2 * q * |E|$ respectively. For the counterfactual augmentation, it is performed only once for all nodes at the beginning of the model, i.e., it is computed

Table 3

Results of prediction and fairness performance on the German dataset. The best results are **bold-faced** and the runner-up results are underlined. Note that FDGNN with APPNP and GCN obtains the best and runner-up fairness results, respectively, and has a similar ACC score as that of the best predictive model.

Dataset	German			
Method	ACC	F_1	Δ_{SP}	Δ_{EO}
GCN	72.45±0.75	81.73±2.31	20.36±5.27	19.71±5.19
FairGCN	68.31±5.17	69.19±5.52	9.75±4.31	7.67±3.69
NI-GCN	67.94±3.16	82.36±0.86	10.15±3.46	7.17±2.28
FV-GCN	70.08±0.85	81.45±0.55	4.71±4.37	4.29±2.78
ED-GCN	69.84±0.31	80.67±1.41	4.13±3.41	5.28±3.29
MI-GCN	70.56±0.14	81.16±0.89	2.04±1.01	1.11±0.59
FD-GCN	69.52±0.25	81.82±0.57	<u>0.69±0.85</u>	<u>0.78±0.59</u>
APPNP	71.59±2.34	79.89±1.45	13.37±5.56	14.32±4.84
FairAPPNP	67.58±5.16	75.16±4.57	9.23±4.13	8.69±4.15
NI-APPNP	68.32±7.39	81.81±1.25	8.31±5.51	6.72±3.19
FV-APPNP	69.28±0.47	81.80±0.38	2.17±2.18	2.71±2.34
ED-APPNP	70.56±0.90	79.54±1.73	9.05±5.76	6.56±4.90
MI-APPNP	70.08±1.32	80.89±0.93	1.89±0.87	1.13±0.84
FD-APPNP	70.00±0.25	<u>82.29±0.15</u>	0.18±0.10	0.66±0.46
SAGE	<u>71.63±1.35</u>	81.08±1.04	14.33±5.11	12.53±7.56
FairSAGE	<u>70.83±1.66</u>	79.57±2.61	6.21±2.34	5.36±2.07
NI-SAGE	66.24±4.12	78.27±1.25	8.03±7.19	4.40±4.18
FV-SAGE	69.60±1.13	81.33±0.55	2.50±3.01	1.26±1.07
ED-SAGE	65.60±6.81	77.89±6.06	4.35±4.29	4.41±3.81
MI-SAGE	70.08±1.48	80.87±0.94	1.40±0.99	0.78±0.61
FD-SAGE	70.00±0.51	82.29±0.31	<u>0.68±0.99</u>	<u>0.67±0.97</u>

$|V|$ times. Let d represent the hidden layer dimension. Considering only the inner product in our analysis, the total complexity of training per epoch is $O(|E|d|V|)$. Therefore, the time complexity of our FDGNN model during the whole training phase is $O((|E|d|V|)t)$.

In disentangled contrastive learning, the concept of intuitive matching offers ample discriminative capability, effectively separating latent representations. This enables FDGNN to learn high-quality graph embeddings more efficiently, often achieved in just a few iterations while other methods may require hundreds to accomplish. We argue that the combination of richer structural information and more complex feature interactions in dense graphs creates an effective optimization environment for FDGNN. Consequently, FDGNN is capable of achieving optimal results on dense graphs in as few as 10 epochs.

5.3. Sensitivity Analysis on Hyperparameters

As part of evaluating our proposed FDGNN model, we conducted a sensitivity analysis to assess the impact of key hyperparameters - learning rate, training epochs, and temperature parameter τ . Specifically, while keeping the other hyperparameters constant, we initially set the temperature parameter τ to 0.5.

For the Recidivism dataset, we varied the learning rate from 0.005 to 0.025 in Table 6. For the German dataset and Credit dataset, we varied the learning rate, which ranged from 0.0005 to 0.0025 in Table 7. We increased the number of training epochs with different backbones on different dataset in Fig 4, and we use Δ_{SP} as a measure for selecting the number of epochs. After identifying the optimal learning rate and epoch number, we then performed a sensitivity analysis on the temperature parameter τ in Table 8.

Our results demonstrate that minor fluctuations in the learning rate have a relatively small impact on model performance, with learning rates of 0.001 to 0.002 producing robust and consistent results on the German and Credit datasets. However, the Recidivism dataset benefits from a higher learning rate of 0.005 to 0.01 due to its lower density of edges. As the learning rate increases, the model’s predictive performance becomes unstable, but there’s a proportional increase in the fairness index. This suggests that there is room for further optimization of learning for sparse graphs. Peak performance for dense graphs is observed at 5-10 epochs, while the sparse Recidivism dataset continues to improve up to 200 epochs before overfitting becomes a concern.

Table 4

Results of prediction and fairness performance on the Recidivism dataset. The best results are **bold-faced** and the runner-up results are underlined. Note that FDGNN obtains the best and runner-up fairness results.

Dataset	Recidivism			
Method	ACC	F_1	Δ_{SP}	Δ_{EO}
GCN	82.49±0.82	77.52±1.35	9.31±2.12	8.59±1.13
FairGCN	83.40±1.19	78.73±1.31	8.09±0.84	5.77±1.10
NI-GCN	79.36±5.36	77.94±2.95	5.29±1.34	5.76±1.17
FV-GCN	85.47±0.47	79.38±0.25	5.66±0.68	4.29±1.15
ED-GCN	85.66±0.17	79.90±0.26	5.74±0.19	3.59±0.26
MI-GCN	85.99±0.31	80.17±0.51	3.85±0.38	2.35±0.94
FD-GCN	80.37±5.53	79.02±3.19	2.89±1.36	2.15±1.33
APPNP	81.17±0.77	78.69±1.13	10.27±1.57	9.69±1.92
FairAPPNP	82.55±0.82	76.57±0.93	6.81±1.51	5.37±1.35
NI-APPNP	81.74±2.25	70.19±3.15	6.01±1.45	4.23±1.31
FV-APPNP	81.47±1.06	76.83±2.15	3.48±0.41	3.64±1.55
ED-APPNP	82.82±1.17	78.04±1.18	5.92±0.86	3.26±1.63
MI-APPNP	86.89±0.26	81.63±0.47	3.35±0.28	2.61±0.44
FD-APPNP	83.19±6.89	79.45±5.18	3.27±1.25	1.55±0.89
SAGE	87.44±1.34	81.57±1.19	8.14±1.08	7.43±1.75
FairSAGE	83.56±2.70	78.37±1.99	6.88±1.41	5.77±1.48
NI-SAGE	80.11±5.39	79.85±3.16	5.96±2.13	5.57±1.69
FV-SAGE	87.61±1.30	82.67±0.87	3.49±1.74	2.42±1.29
ED-SAGE	83.15±2.96	80.42±2.53	6.57±1.35	5.61±1.73
MI-SAGE	87.48±0.28	82.52±0.50	3.17±0.21	1.72±0.56
FD-SAGE	<u>80.82±4.02</u>	<u>77.88±7.27</u>	1.57±1.09	0.68±0.43

Table 5

Results on runtime (in seconds). The shortest time in each dataset is **bold-faced**.

Method	Credit	German	Recidivism
FairGNN	100.39±8.35	15.11±0.58	48.70±9.53
NIFTYY	49.25±0.33	32.14±0.84	42.02±0.75
FVGNN	183.46±0.91	116.57±0.39	485.32±4.07
EDITS	318.47±7.30	37.58±0.43	154.38±12.14
FairMILE	129.80±2.33	8.21±0.27	78.86±3.71
FDGNN	3.57±0.29	0.92±0.29	18.94±0.25

We conducted a sensitivity analysis on the temperature parameter τ within our contrastive loss function, spanning values from 0.1 to 0.9. The results in Table 8 demonstrate that $\tau = 0.5$ yields optimal performance across all metrics. However, FDGNN exhibits robustness to τ , achieving strong results within the range of $0.3 \leq \tau \leq 0.7$.

5.4. Ablation Studies

5.4.1. Ablation Studies on the Data Augmentation Strategies

To better understand the contribution of each component in FDGNN, we perform ablation experiments by removing individual modules from each of the three different backbones. Specifically, we validate the significance of the data enhancement operation and degree information. We also use the feature masking operation instead of the edge-dropping operation to explore the role of different types of data enhancement operations. We visualize the ablation experiments on different datasets in Fig. 3 for GraphSAGE-based FDGNN, w.r.t. data-augmentation.

From Fig. 3, each module in our framework plays a positive role. Random walk simulates the incomplete graph data that may be obtained in the actual process, which reflects the local structure of the original graph and can provide more examples to enhance the model’s generalization ability. Edge dropping removes edges in the original graph randomly with a certain probability and can obtain a new graph with a slightly changed structure, which simulates the

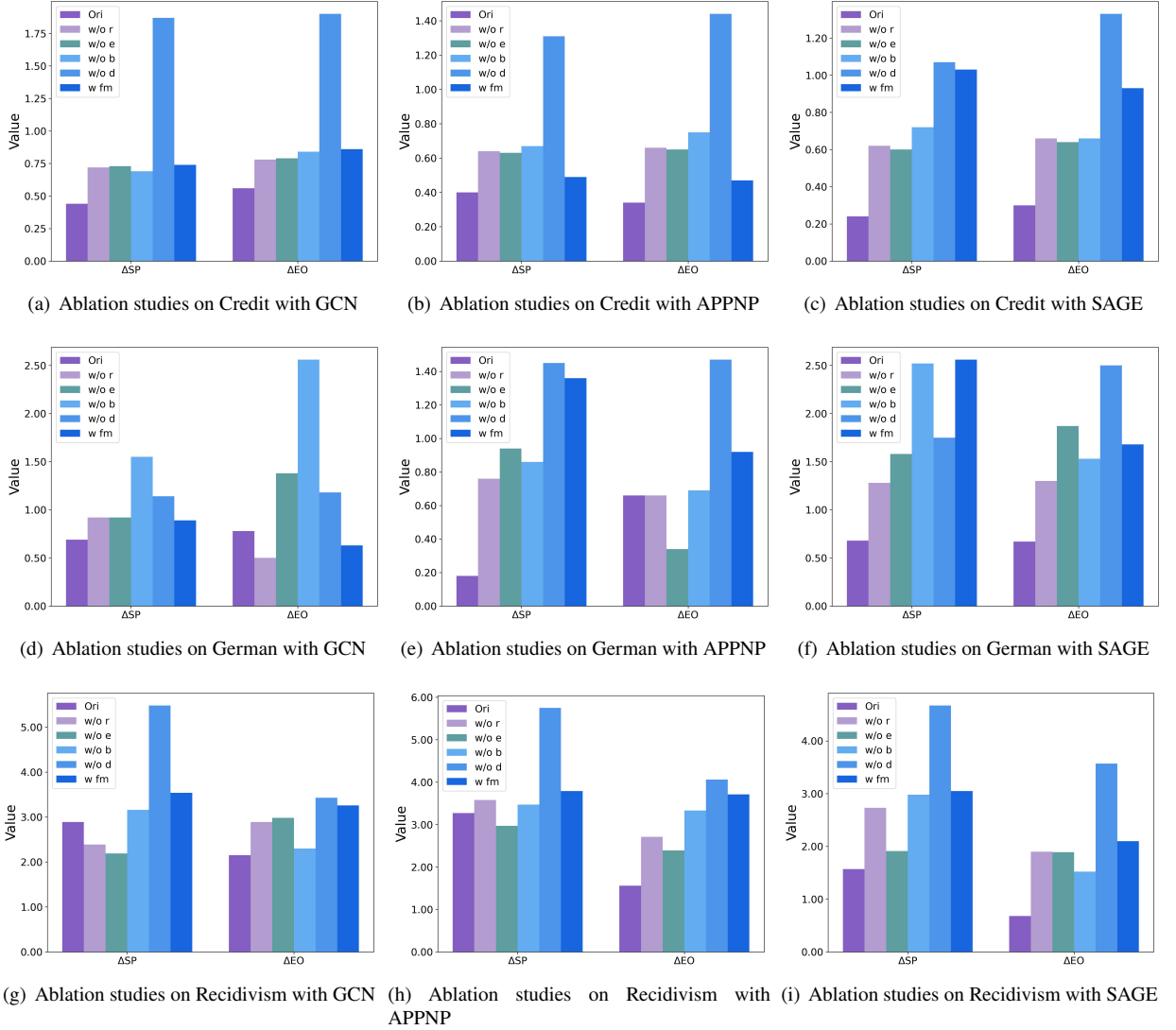
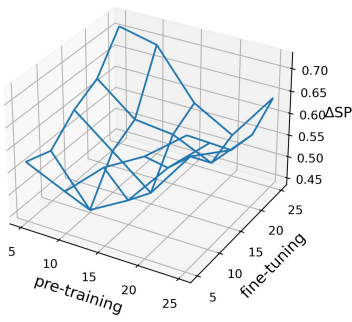


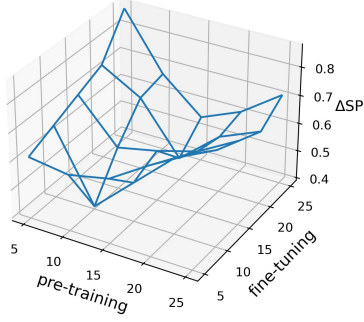
Figure 3: Ablation studies on different datasets. We denote the original FDGNN by ‘Ori’, the removal of the random walk operation alone by ‘w/o r’, the removal of the edge dropping operation alone by ‘w/o e’, the removal of both operations by ‘w/o b’, the removal of the extra added degree information alone by ‘w/o d’, and the use of feature masking operation instead of edge dropping operation by ‘w fm’.

possible changes in the graph topology, enhances the robustness of the model to structural changes, and helps to prevent over-fitting. These two methods improve FDGNN’s ability to discriminate the non-sensitive attributes by changing the adjacency matrix, which preserves the commonalities between the positive sample pairs.

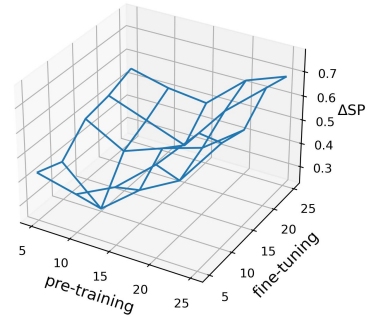
On the other hand, feature masking randomly masks the original input feature dimensions of the nodes, forcing the model to learn features from different subspaces. This can motivate the model to build representations without over-relying on some original features, thus improving the generalization ability. However, feature masking in FDGNN does not have a positive effect in many settings, and a negative effect occurs in some settings. This is because the goal of FDGNN is to learn the disentangled representations, and randomly masking features will make the model unable to differentiate between non-sensitive and sensitive attributes effectively, thus reducing the fairness of the final representation.



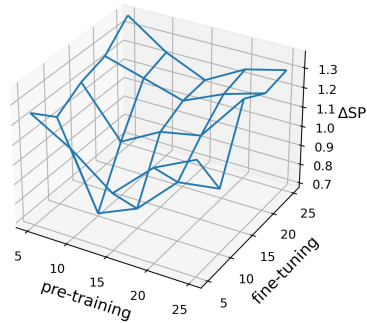
(a) Sensitivity analysis on Credit with GCN



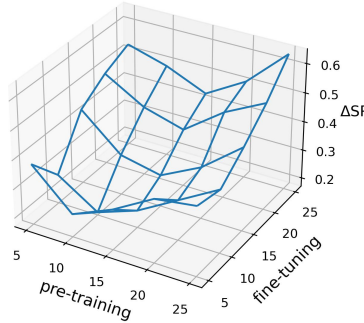
(b) Sensitivity analysis on Credit with APPNP



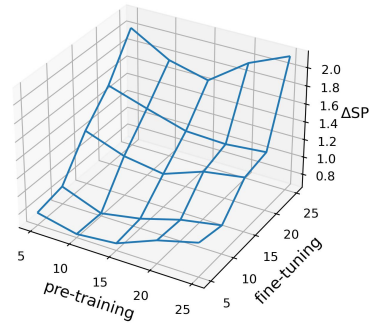
(c) Sensitivity analysis on Credit with SAGE



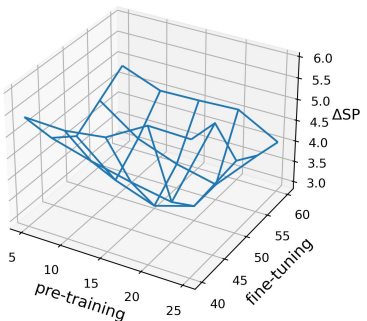
(d) Sensitivity analysis on German with GCN



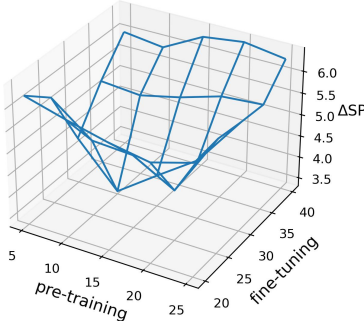
(e) Sensitivity analysis on German with APPNP



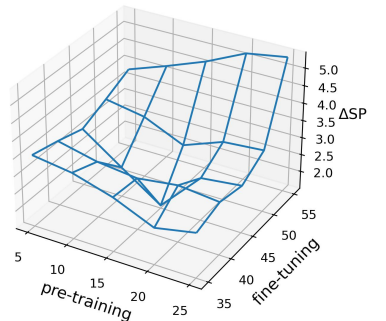
(f) Sensitivity analysis on German with SAGE



(g) Sensitivity analysis on Recidivism with GCN



(h) Sensitivity analysis on Recidivism with APPNP



(i) Sensitivity analysis on Recidivism with SAGE

Figure 4: Sensitivity analysis of epochs on different datasets.

5.4.2. Ablation Studies on Counterfactual Augmentation Strategy

To evaluate the efficacy of the counterfactual augmentation strategy in FDGNN, we conduct ablation studies by removing the counterfactual augmentation strategy from FDGNN. For positive pair construction, we no longer utilize original features paired with counterfactuals, instead pairing real samples with each enhancement separately.

Table 6

Sensitivity analysis of learning rate on Recidivism dataset.

Dataset	Recidivism			
LR	ACC	F_1	Δ_{SP}	Δ_{EO}
0.02	71.67 \pm 2.64	73.29 \pm 1.47	4.22 \pm 0.97	3.47 \pm 1.10
0.025	80.56 \pm 6.59	75.32 \pm 4.65	2.82 \pm 0.26	1.17 \pm 1.06
0.03	80.82 \pm 4.02	77.88 \pm 7.27	1.57\pm1.09	0.68\pm0.43
0.035	85.62\pm3.72	82.94\pm3.02	3.15 \pm 1.87	2.32 \pm 0.76
0.04	84.67 \pm 2.46	80.29 \pm 1.49	6.22 \pm 4.33	2.37 \pm 2.10

Table 7

Sensitivity analysis of learning rate on Credit dataset and German dataset.

Dataset	Credit				German			
LR	ACC	F_1	Δ_{SP}	Δ_{EO}	ACC	F_1	Δ_{SP}	Δ_{EO}
0.0005	77.87\pm0.02	87.56 \pm 0.11	0.33 \pm 0.19	0.34 \pm 0.32	69.92 \pm 0.16	82.28 \pm 0.11	1.06 \pm 0.27	1.53 \pm 0.71
0.001	77.87\pm0.02	87.56\pm0.01	0.24\pm0.14	0.30\pm0.18	70.00\pm0.51	82.29\pm0.31	0.68\pm0.99	0.67\pm0.97
0.0015	77.87\pm0.02	87.56 \pm 0.08	0.34 \pm 0.35	0.31 \pm 0.30	69.76 \pm 1.03	81.95 \pm 0.66	0.99 \pm 0.62	1.28 \pm 0.97
0.002	77.83 \pm 0.05	87.18 \pm 0.06	0.41 \pm 0.41	0.49 \pm 0.29	69.28 \pm 1.65	80.40 \pm 1.90	1.25 \pm 1.61	2.56 \pm 1.92
0.0025	77.47 \pm 0.65	87.09 \pm 0.46	0.74 \pm 0.71	0.40 \pm 0.49	66.80 \pm 2.75	78.25 \pm 3.13	2.58 \pm 1.73	3.35 \pm 1.38

Table 8Sensitivity analysis of temperature parameter τ .

Dataset	Credit				German				Recidivism			
τ	ACC	F_1	Δ_{SP}	Δ_{EO}	ACC	F_1	Δ_{SP}	Δ_{EO}	ACC	F_1	Δ_{SP}	Δ_{EO}
0.1	77.87 \pm 0.53	87.56 \pm 0.33	0.93 \pm 0.58	0.37 \pm 0.74	70.48\pm0.37	80.52 \pm 0.41	1.79 \pm 1.98	0.52 \pm 0.43	83.59\pm2.22	76.97\pm3.02	4.68 \pm 1.20	3.92 \pm 1.71
0.3	77.81 \pm 0.12	87.51 \pm 0.08	0.39 \pm 0.21	0.27\pm0.19	70.16 \pm 0.32	80.40 \pm 0.18	0.77 \pm 0.49	0.61 \pm 0.41	81.63 \pm 2.34	78.03 \pm 3.34	2.40 \pm 1.41	2.43 \pm 1.32
0.5	77.87\pm0.02	87.56\pm0.01	0.24\pm0.14	0.30 \pm 0.18	70.00 \pm 0.51	82.29 \pm 0.31	0.68\pm0.99	0.67 \pm 0.97	80.82 \pm 4.02	77.88 \pm 3.27	1.57\pm1.09	0.68\pm0.43
0.7	77.85 \pm 0.04	87.54 \pm 0.03	0.24 \pm 0.15	0.56 \pm 0.69	69.92 \pm 0.30	82.26 \pm 0.18	0.69 \pm 0.25	0.50\pm0.41	79.79 \pm 4.96	77.85 \pm 3.64	3.48 \pm 1.63	3.49 \pm 0.89
0.9	77.83 \pm 0.07	87.53 \pm 0.04	0.35 \pm 0.47	0.74 \pm 0.12	70.48 \pm 0.81	82.36\pm0.16	0.83 \pm 0.58	0.61 \pm 0.41	82.73 \pm 1.26	75.51 \pm 2.58	4.98 \pm 1.38	3.10 \pm 1.48

Table 9Ablation studies on counterfactual augmentation strategy. The best results are **bold-faced** and the runner-up results are underlined. For simplicity and clarity in the table, we abbreviate the counterfactual augmentation strategy as 'counter'.

Dataset Method		Credit		German		Recidivism	
		Δ_{SP}	Δ_{EO}	Δ_{SP}	Δ_{EO}	Δ_{SP}	Δ_{EO}
GCN	FD-GCN	0.44 \pm 0.17	0.56 \pm 0.21	0.69 \pm 0.85	0.78 \pm 0.59	2.89 \pm 1.36	2.15 \pm 1.33
	w/o counter	1.60 \pm 0.90	1.84 \pm 0.91	0.81 \pm 0.63	0.69 \pm 0.95	4.53 \pm 1.34	2.91 \pm 1.50
APPNP	FD-APPNP	<u>0.40\pm0.22</u>	<u>0.34\pm0.15</u>	0.18\pm0.10	0.66\pm0.46	3.37 \pm 1.27	<u>1.56\pm0.76</u>
	w/o counter	0.52 \pm 0.13	0.60 \pm 0.14	0.76 \pm 0.31	0.68 \pm 0.42	4.86 \pm 1.16	3.97 \pm 0.40
SAGE	FD-SAGE	0.24\pm0.14	0.30\pm0.18	<u>0.68\pm0.99</u>	<u>0.67\pm0.97</u>	1.57\pm1.09	0.68\pm0.43
	w/o counter	1.58 \pm 1.02	1.76 \pm 1.18	1.36 \pm 1.15	1.22 \pm 0.33	3.02 \pm 1.59	2.28 \pm 0.91

For negative pairs, we solely use a node's sensitive and non-sensitive representations, excluding the counterfactual counterparts.

We report the partial results in Table 9. From Table 9, we know that FDGNN without counterfactual augmentation strategy reduces fairness after ablating counterfactual data, validating their importance for effectively obtaining fair node representation. In other words, the constructed counterfactual instances balance the distribution of sensitive values across different groups by utilizing the matching strategy in causal inference (Pearl, 2009; Stuart, 2010). Thus, the ablation experiments confirm the role of the counterfactual augmentation strategy in our FDGNN model.

Table 10

Ablation studies on the encoder and pre-training approach. The best results are **bold-faced** and the runner-up results are underlined. For clarity of the table, we have added 'E' to the beginning of the end-to-end FDGNN, 'G' to the beginning of the pre-trained fine-tuned FDGNN that does not use VGAE, and 'EG' to the beginning of the end-to-end FDGNN that does not use VGAE.

Dataset Method		Credit		German		Recidivism	
		Δ_{SP}	Δ_{EO}	Δ_{SP}	Δ_{EO}	Δ_{SP}	Δ_{EO}
VGAE	FD-GCN	0.44±0.17	0.56±0.21	0.69±0.85	0.78±0.59	2.89±1.36	2.15±1.33
	E-FD-GCN	1.68±0.77	1.86±0.94	1.33±1.64	0.92±1.05	3.81±0.56	2.30±2.28
w/o VGAE	G-FD-GCN	0.60±0.28	0.70±0.28	0.90±1.19	0.82±0.53	3.14±1.88	2.75±1.57
	EG-FD-GCN	2.09±2.74	2.14±2.80	1.71±1.73	1.13±0.94	4.79±1.80	3.95±1.47
VGAE	FD-APPNP	0.40±0.22	0.34±0.15	0.18±0.10	0.66±0.46	3.27±1.25	1.55±0.89
	E-FD-APPNP	0.62±0.42	0.64±0.47	1.85±1.45	2.65±3.36	3.87±0.84	2.45±0.53
w/o VGAE	G-FD-APPNP	0.76±0.75	0.68±0.78	<u>0.51±0.64</u>	0.67±1.34	4.10±1.67	1.33±1.16
	EG-FD-APPNP	2.20±0.48	2.10±0.47	0.85±0.56	1.01±0.34	4.46±1.60	2.87±0.51
VGAE	FD-SAGE	0.24±0.14	0.30±0.18	0.68±0.99	<u>0.67±0.97</u>	1.57±1.09	0.68±0.43
	E-FD-SAGE	0.58±0.45	0.59±0.46	0.94±1.15	0.71±0.87	3.01±2.04	1.22±0.59
w/o VGAE	G-FD-SAGE	0.55±0.24	0.51±0.25	2.21±1.44	1.58±0.89	<u>2.58±3.53</u>	<u>0.93±0.49</u>
	EG-FD-SAGE	0.56±0.33	0.61±0.35	3.81±1.93	1.85±1.11	4.02±1.58	2.19±0.41

5.4.3. Ablation Studies on the Encoder and Pre-training Approach

We conduct ablation experiments to examine the contribution of the VGAE encoder and the pre-training to the performance of FDGNN. In detail, we compare the performance of FDGNN with the following three variants: 1) G-FDGNN, which replaces VGAE in FDGNN with the backbone model serving as encoder, 2) E-FDGNN, which removes pre-training in FDGNN and uses end-to-end training only, and 3) EG-FDGNN, which combines the above two changes, i.e. replaces VAGE in FDGNN with the backbone encoder and uses end-to-end training without pre-training. We report the partial results in Table 10.

From Table 10, we observe that VGAE’s resampling introduces greater diversity into the inputs for the encoder. This increased randomness helps correct for biases in the data distribution. Consequently, it mitigates the influence of inherent data biases on the disentanglement effect, enabling the encoder to learn a more robust disentangled representation. Furthermore, pre-training has been found to provide more robust initial representations and reduce the difficulty of the contrastive learning task. Starting from scratch with end-to-end training has made it more challenging to simultaneously optimize the disentanglement and contrastive objectives. The experiments provide empirical evidence supporting the design choices made in our FDGNN framework.

6. Conclusion

Eliminating the effect of sensitive attributes from data representation constitutes the paramount objective of fair representation learning. Disentanglement techniques present an efficacious way to achieve this goal. The FDGNN framework proposed in this paper, pioneers the integration of disentanglement and contrastive learning to obtain the fairness of GNN models. In detail, FDGNN establishes positive-negative sample pairs through two varieties of edge-based data augmentation and counterfactual augmentation. Subsequently, it employs contrastive learning to uncover the commonalities within non-sensitive attributes while effectively segregating sensitive attributes from the representations, thus obtaining a fair disentangled representation. Extensive experiments across three real-world datasets validate the efficacy of our proposed FDGNN in acquiring fair representations, indicating that FDGNN offers a promising avenue for facilitating fair decision-making in graph-based machine-learning tasks.

CRedit authorship contribution statement

Guixian Zhang: Conceptualisation, Methodology, Validation, Investigation, Writing - original draft & editing, Visualisation. **Guan Yuan:** Writing - review & editing, Supervision, Funding acquisition. **Debo Cheng:** Conceptualisation, Writing - review & editing, Formal analysis. **Lin Liu:** Conceptualisation, Writing - review & editing. **Jiuyong**

Li: Conceptualisation, Writing - review & editing. **Shichao Zhang:** Conceptualisation, Writing - review & editing, Supervision, Funding acquisition.

Data availability

Data will be made available on request.

Declarations of competing interest

The authors declare that they have no conflict of interest that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 71774159 and 62277046, the Project of Guangxi Science and Technology GuiKeAB23026040, China Postdoctoral Science Foundation under Grants 2021T140707, the Jiangsu Postdoctoral Science Foundation under grant number 2021K565C, Xuzhou K&D Program under grant KC23296, the Australian Research Council Discovery Project 230101122, the Graduate Innovation Program of China University of Mining and Technology 2024WLKXJ183, the Fundamental Research Funds for the Central Universities 2024-10949, the Postgraduate Research & Practice Innovation Program of Jiangsu Province KYCX24_2781.

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*, pages 2114–2124. PMLR.
- Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44.
- Cai, R., Wu, F., Li, Z., Qiao, J., Chen, W., Hao, Y., and Gu, H. (2023). Rest: Debaised social recommendation via reconstructing exposure strategies. *ACM Transactions on Knowledge Discovery from Data*, 18(2):1–24.
- Chen, C. Y.-C., Byrne, E., and Vélez, T. (2022a). Impact of the 2020 pandemic of covid-19 on families with school-aged children in the united states: Roles of income level and race. *Journal of Family Issues*, 43(3):719–740.
- Chen, Y., Yang, H., Zhang, Y., KAILI, M., Liu, T., Han, B., and Cheng, J. (2022b). Understanding and improving graph injection attack by promoting unnoticeability. In *International Conference on Learning Representations*.
- Chen, Y., Zhang, Y., Bian, Y., Yang, H., Kaili, M., Xie, B., Liu, T., Han, B., and Cheng, J. (2022c). Learning causally invariant representations for out-of-distribution generalization on graphs. In *Advances in Neural Information Processing Systems*, volume 35, pages 22131–22148.
- Creager, E., Madras, D., Jacobsen, J.-H., Weis, M., Swersky, K., Pitassi, T., and Zemel, R. (2019). Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning*, pages 1436–1445. PMLR.
- Dai, E. and Wang, S. (2022). Learning fair graph neural networks with limited and private sensitive attribute information. *IEEE Transactions on Knowledge & Data Engineering*, pages 1–14.
- 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 2022*, pages 1259–1269.
- Dua, D. and Graff, C. (2017). UCI machine learning repository.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226.
- Fang, J., Meng, X., and Qi, X. (2023). A top-k poi recommendation approach based on lbn and multi-graph fusion. *Neurocomputing*, 518:219–230.
- Feng, S., Jing, B., Zhu, Y., and Tong, H. (2022). Adversarial graph contrastive learning with information regularization. In *Proceedings of the ACM Web Conference 2022*, pages 1362–1371.
- Gasteiger, J., Bojchevski, A., and Günnemann, S. (2019). Predict then propagate: Graph neural networks meet personalized pagerank. In *International Conference on Learning Representations*, pages 1–15.
- Guan, R., Li, Z., Tu, W., Wang, J., Liu, Y., Li, X., Tang, C., and Feng, R. (2024). Contrastive multiview subspace clustering of hyperspectral images based on graph convolutional networks. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14.
- 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*, pages 669–678.
- Hamilton, W. L., Ying, R., and Leskovec, J. (2017). Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1025–1035.
- Hang, J., Dong, Z., Zhao, H., Song, X., Wang, P., and Zhu, H. (2022). Outside in: Market-aware heterogeneous graph neural network for employee turnover prediction. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 353–362.
- Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3323–3331.

- He, Y., Gurukar, S., and Parthasarathy, S. (2023). Fairmile: Towards an efficient framework for fair graph representation learning. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–10.
- Hernán, M. A. and Robins, J. M. (2010). Causal inference.
- Hou, Z., Liu, X., Cen, Y., Dong, Y., Yang, H., Wang, C., and Tang, J. (2022). Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 594–604.
- Hu, Y., Liao, T., Chen, J., Bian, J., Zheng, Z., and Chen, C. (2024). Migrate demographic group for fair graph neural networks. *Neural Networks*, page 106264.
- Jordan, K. L. and Freiburger, T. L. (2015). The effect of race/ethnicity on sentencing: Examining sentence type, jail length, and prison length. *Journal of Ethnicity in Criminal Justice*, 13(3):179–196.
- Kaur, D., Uslu, S., Rittichier, K. J., and Duresi, A. (2022). Trustworthy artificial intelligence: a review. *ACM Computing Surveys (CSUR)*, 55(2):1–38.
- Kipf, T. N. and Welling, M. (2016a). Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, pages 1–14.
- Kipf, T. N. and Welling, M. (2016b). Variational graph auto-encoders. *NIPS Workshop on Bayesian Deep Learning*, pages 1–3.
- Kong, K., Li, G., Ding, M., Wu, Z., Zhu, C., Ghanem, B., Taylor, G., and Goldstein, T. (2022). Robust optimization as data augmentation for large-scale graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 60–69.
- Li, H., Wang, X., Zhang, Z., Yuan, Z., Li, H., and Zhu, W. (2021). Disentangled contrastive learning on graphs. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pages 21872–21884.
- Li, H., Zhang, Z., Wang, X., and Zhu, W. (2022). Learning invariant graph representations for out-of-distribution generalization. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 11828–11841.
- Li, M., Zhuang, X., Bai, L., and Ding, W. (2024). Multimodal graph learning based on 3d haar semi-tight framelet for student engagement prediction. *Information Fusion*, 105:102224.
- Ling, Z., Xu, E., Zhou, P., Du, L., Yu, K., and Wu, X. (2024). Fair feature selection: A causal perspective. *ACM Transactions on Knowledge Discovery from Data*.
- Liu, C., Zhan, Y., Yu, B., Liu, L., Du, B., Hu, W., and Liu, T. (2023). On exploring node-feature and graph-structure diversities for node drop graph pooling. *Neural Networks*, 167:559–571.
- Ma, J., Guo, R., Wan, M., Yang, L., Zhang, A., and Li, J. (2022). Learning fair node representations with graph counterfactual fairness. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 695–703.
- Ma, J., Guo, R., Zhang, A., and Li, J. (2023). Learning for counterfactual fairness from observational data. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1–11.
- Martinez, D. A., Hinson, J. S., Klein, E. Y., Irvin, N. A., Saheed, M., Page, K. R., and Levin, S. R. (2020). Sars-cov-2 positivity rate for latinos in the baltimore–washington, dc region. *Jama*, 324(4):392–395.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.
- Mo, Y., Lei, Y., Shen, J., Shi, X., Shen, H. T., and Zhu, X. (2023). Disentangled multiplex graph representation learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 24983–25005.
- Mukherjee, D., Yurochkin, M., Banerjee, M., and Sun, Y. (2020). Two simple ways to learn individual fairness metrics from data. In *International Conference on Machine Learning*, pages 7097–7107. PMLR.
- Nyhan, B., Settle, J., Thorson, E., Wojcieszak, M., Barberá, P., Chen, A. Y., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., et al. (2023). Like-minded sources on facebook are prevalent but not polarizing. *Nature*, pages 1–8.
- Oh, C., Won, H., So, J., Kim, T., Kim, Y., Choi, H., and Song, K. (2022). Learning fair representation via distributional contrastive disentanglement. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1295–1305.
- Park, S., Hwang, S., Kim, D., and Byun, H. (2021). Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2403–2411.
- Pearl, J. (2009). *Causality*. Cambridge university press.
- Price-Haywood, E. G., Burton, J., Fort, D., and Seoane, L. (2020). Hospitalization and mortality among black patients and white patients with covid-19. *New England Journal of Medicine*, 382(26):2534–2543.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1.
- Sun, Q., Li, J., Peng, H., Wu, J., Ning, Y., Yu, P. S., and He, L. (2021). Sugar: Subgraph neural network with reinforcement pooling and self-supervised mutual information mechanism. In *Proceedings of the Web Conference 2021*, pages 2081–2091.
- Wan, G., Huang, W., and Ye, M. (2024). Federated graph learning under domain shift with generalizable prototypes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15429–15437.
- 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*, pages 1938–1948.
- Wang, Z., Yang, P., Fan, X., Yan, X., Wu, Z., Pan, S., Chen, L., Zang, Y., Wang, C., and Yu, R. (2024). Contig: Continuous representation learning on temporal interaction graphs. *Neural Networks*, 172:106151.
- Wolfson, J. A. and Leung, C. W. (2020). Food insecurity and covid-19: disparities in early effects for us adults. *Nutrients*, 12(6):1648.
- Wu, Q., Zhang, H., Gao, X., He, P., Weng, P., Gao, H., and Chen, G. (2019). Dual graph attention networks for deep latent representation of multifaceted social effects in recommender systems. In *The world wide web conference*, pages 2091–2102.
- Wu, Z., Mo, Y., Zhou, P., Yuan, S., and Zhu, X. (2024). Self-training based few-shot node classification by knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15988–15995.

- Xhonneux, L.-P., Qu, M., and Tang, J. (2020). Continuous graph neural networks. In *International Conference on Machine Learning*, pages 10432–10441. PMLR.
- Xu, J., Yuan, C., Ma, X., Shang, H., Shi, X., and Zhu, X. (2024). Interpretable medical deep framework by logits-constraint attention guiding graph-based multi-scale fusion for alzheimer’s disease analysis. *Pattern Recognition*, page 110450.
- Yang, C., Zou, J., Wu, J., Xu, H., and Fan, S. (2022). Supervised contrastive learning for recommendation. *Knowledge-Based Systems*, 258:109973.
- 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*, 36(2):2473–2480.
- You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., and Shen, Y. (2020). Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33:5812–5823.
- Zhang, G., Cheng, D., Yuan, G., and Zhang, S. (2024a). Learning fair representations via rebalancing graph structure. *Information Processing & Management*, 61(1):103570.
- Zhang, G., Sheng, J., Wang, S., and Liu, T. (2024b). Noise-disentangled graph contrastive learning via low-rank and sparse subspace decomposition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5880–5884. IEEE.
- Zhang, S., Chen, X., Shen, X., Ren, B., Yu, Z., Yang, H., Jiang, X., Shen, D., Zhou, Y., and Zhang, X.-Y. (2023). A-gcl: Adversarial graph contrastive learning for fmri analysis to diagnose neurodevelopmental disorders. *Medical Image Analysis*, 90:102932.