Advancing Graph Counterfactual Fairness Through Fair Representation Learning

Zichong Wang[1], Zhibo Chu[1], Ronald Blanco[1], Zhong Chen[2], Shu-Ching Chen[3], and Wenbin Zhang[1(✉)]

[1] Florida International University, Miami, USA
{ziwang,wenbin.zhang}@fiu.edu
[2] Southern Illinois University, Carbondale, USA
[3] University of Missouri-Kansas City, Kansas City, USA

Abstract. Graph neural networks (GNNs) have shown remarkable success in various domains. Nonetheless, studies have shown that GNNs may inherit and amplify societal bias, which critically hinders their application in high-stakes scenarios. Although efforts have been exerted to enhance the fairness of GNNs, most of them rely on the statistical fairness notion, which assumes that biases arise solely from sensitive attributes, neglecting the pervasive issue of labeling bias prevalent in real-world scenarios. To this end, recent works extend counterfactual fairness in graph data to address label bias, but they neglect the graph structure bias, where nodes sharing sensitive attributes tend to connect more closely. To bridge these gaps, we propose a novel GNN framework, Fair Disentangled GNN (FDGNN), designed to mitigate multi-sources biases to enhance the fairness of GNNs while preserving task-related information via fair node representation learning. Specifically, FDGNN initiates by mitigating graph structure bias by ensuring consistent representation of different subgroups. Subsequently, to achieve fair node representation, identified counterfactual instances are utilized as guides for disentangling a node's representation and eliminating sensitive attribute-related information via a de-identifiable sensitive attribute mechanism. Extensive experiments on multiple real-world graph datasets demonstrate the superiority of FDGNN in graph fairness compared to other state-of-the-art methods while achieving comparable utility performance.

Keywords: GNNs · Counterfactual fairness · Fair representation

1 Introduction

Graph neural networks (GNNs) have emerged as a powerful tool for learning node representation from graph-structured data, which are employed in various domains such as recommendation systems [10], social network analysis [13], and online advertisement [34]. Generally, GNNs adopt a message-passing mechanism (MP) [32], aggregating local neighborhood information for every node in each

A. Bifet et al. (Eds.): ECML PKDD 2024, LNAI 14947, pp. 40–58, 2024.
https://doi.org/10.1007/978-3-031-70368-3_3

layer. This aggregation process effectively renders the distinction between similar and dissimilar nodes while preserving node attributes and graph structure information, thereby enhancing the performance of downstream graph tasks [1]. Despite these successes, GNNs may make discriminatory predictions for subgroups defined by *sensitive attributes* (*e.g.,* gender or race) due to biases inherited from training data and further amplified by their message-passing mechanism. Such biased predictions give rise to ethical and societal concerns, which severely limits the adoption of GNNs in high-stake decision-making scenarios, such as job screening [22], healthcare [40] and criminal prediction [16]. For instance, a bank's loan decision-making process is influenced by the race information of the applicant and their close contacts, constituting a serious ethical problem [21,42,43].

To this end, many efforts have been taken towards fair GNNs [28]. Among them, most existing fairness work utilizes statistical fairness notions to evaluate and address bias in node representation learning on graphs, which highlights algorithmic decisions should equally treat subgroups or individuals, with these methods primarily focusing on *sensitive attributes* (*e.g.,* race or gender) as the only source of bias [22]. However, these strategies cannot quantify and mitigate labeling bias which arises when societal biases, prejudices, or discriminatory practices skew the data collection process [23]. This distortion introduces systemic biases into the training dataset, which GNNs may then learn and perpetuate, exacerbating the bias against the *deprived subgroups* (*e.g.,* female) [21].

To this end, recent research has incorporated counterfactual fairness into graph learning, aiming to address the model's bias from a causal perspective [27]. Typically, these approaches fall into two categories: generation of counterfactual instances based on real sample distributions or identification of potential counterfactual instances within the dataset. For example, GEAR [21] employs GraphVAE [24] to generate counterfactuals aimed at minimizing the disparity between original and counterfactual node representation to eliminate the impact of sensitive attributes. On the other hand, RFCGNN [27] aims to identify corresponding counterfactual instances directly from the representation space and learn disentangled representations, thereby removing sensitive attribute-related information to enhance fairness. A significant limitation of these approaches is neglecting the intricate interplay between sensitive attribute-related information and task-related information. Specifically, they aim to eliminate the sensitive attribute information to force GNNs to make decisions independent of the sensitive attribute, which inadvertently leads to the unintentional removal of the task-related information due to its correlations with the sensitive attribute.

Furthermore, these methods often overlook the graph structure bias present in the graph data, where nodes sharing the same sensitive attributes are likely to be connected [30]. Specifically, GNNs aggregate each node's neighboring node information and its own features to obtain a final node representation. However, the disparity in the distribution of neighboring nodes of the target node can lead to an over-association of node representation with sensitive attributes. This results in the obtained counterfactual instances being too tightly connected to

neighboring nodes with the same sensitive attributes, resulting in inaccurate counterfactual scenarios.

In this paper, we investigate counterfactual fairness to mitigate the root causes of bias, focusing on the potential causal interactions between each node and its neighboring nodes. While great progress has been made in the field, the application of counterfactual fairness to graphs faces distinctive challenges due to fundamental obstacles as follows. **1) Complexity of Counterfactual Graph Data Structures:** Unlike tabular data, graph-structured data contains node features and graph structure information. Thus, given the complexity of these relationships, in counterfactual scenarios, it is imperative to consider the implications of sensitive attribute flipping not only on the target node features but also on its connectivity with neighboring nodes. **2) Mitigating Bias in Node Representations:** To achieve fairness in GNNs, it is essential to mitigate bias while preserving model performance, which requires reasonably handling task-related information that is also associated with the sensitive attribute. This involves disentangling node representations to isolate sensitive attributes related information effectively, thereby ensuring the retention of valuable task-related information. **3) Obtaining Accurate Counterfactual Scenarios:** The essence of counterfactual fairness hinges on accurate counterfactual scenarios. Existing fairness works often overlook the graph structure bias, leading to the derivation of inaccurate counterfactual instances. An effective strategy is thus required to mitigate the association of learned representations with sensitive attributes while maintaining important information.

In order to address all the above-mentioned challenges, this paper proposes a novel framework named *Fair Disentangled Graph neural networks* (FDGNN), which aims to learn fair node representation while preserving task-related information. *To the best of our knowledge, this is the first work that utilizes authentic counterfactual samples to learn disentangled node representation to mitigate the multi-source biases from sensitive attributes, graph structure, and the labeling process collectively.* Specifically, we conduct a comprehensive causal analysis of both original and counterfactual instances, establishing a set of constraints that foster the learning of disentangled representations. This strategy effectively diminishes the associations between sensitive attributes and unrelated representation dimensions. Moreover, by imposing fairness constraints on components associated with sensitive attributes, FDGNN minimizes the influence of the sensitive attribute-related information on other representation channels. This approach prevents unnecessary task-related information loss, leading to a more balanced and effective model. The main contributions are as follows:

- **A novel graph causal model.** We introduce a novel causal formulation that paves the way for understanding the generation process of graph structures and the fair learning task of node representation.
- **A novel framework for mitigating graph-structured data bias via counterfactual instance.** We propose FDGNN, a fair graph representation learning framework that utilizes accurate counterfactual instances to mitigate multi-source biases, including sensitive attributes, graph structure, and

the labeling process. In addition, FDGNN preserves task-relevant information associated with sensitive attributes by effectively disentangling sensitive attributes. This approach enables our model to enhance fairness without compromising performance.
- **Extensive experiments are conducted to evaluate our proposed approach.** We conduct extensive experiments on three real-world datasets and five evaluative metrics, the results show that FDGNN acquires superior performance and significantly enhances fairness compared with baselines.

The organization of this paper is as follows: An overview of the relevant literature is provided in Sect. 2. Notations are presented in Sect. 3. Our proposed method is detailed in Sect. 4. Section 5 describes the experimental framework and discusses the experiment results. Lastly, Sect. 6 concludes the paper.

2 Related Work

2.1 Graph Neural Networks

Graph Neural Networks have shown great ability in representation learning on graph-structured data and have been used in a variety of tasks such as node classification [18], graph classification [25], and link prediction [32]. Their notable success across these diverse tasks has propelled GNNs into the forefront of research and application, extending their utility into critical decision-making systems [27]. For instance, financial institutions increasingly rely on GNNs to evaluate credit card applications or make loan approval decisions [29]. However, the application in critical decision-making systems places higher demand for GNNs to not only be effective but also fair and interpretable [38]. In this context, there is a trend for the research community to design fairer GNNs to mitigate biases and ensure fair outcomes in graph-based tasks [28].

2.2 Fairness in Graph

Fairness in the graph has received intensive attention [4,5,8,31,37,39]. Most existing fair graph learning works are based on statistical fairness notation, including individual fairness [17,26,41] and group fairness [6,7,40], aiming to ensure fair GNN predictions. While these approaches have achieved notable success, their focus on correlation metrics often renders them ineffective at addressing biases introduced by statistical anomalies. To address this limitation, counterfactual fairness [19] leverages the causal perspective to measure and eliminate the root bias. For example, NIFTY [1] generates counterfactual instances by directly flipping the sensitive attributes of nodes to enhance the consistency between original and counterfactual representations. Similarly, GEAR [21] employs GraphVAE [24] to generate counterfactuals, focusing on minimizing the difference between representations derived from the original and counterfactuals. Furthermore, RFCGNN [27] identifies counterfactuals within the existing representation space to learn fair representation.

Our work is distinct from existing works in that it: i) employs disentangled representation learning to preserve essential task-relevant information, minimizing performance loss; and ii) pays attention to fundamental yet neglected graph structural bias.

3 Notations

Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathbf{X}\}$ denote an undirected attributed graph, comprised of a set of $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ nodes and a set of $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ edges. $\mathbf{X} \in \mathbb{R}^{n \times d}$ represents the node feature matrix with the i-th row of $\mathbf{X}$, *i.e.,* $\mathbf{X}_{i,:}$ as node feature of v_i with d being the dimension of node features. The adjacency matrix $\mathbf{A} \in \{0,1\}^{n \times n}$ encapsulates the graph structure information, where $\mathbf{A}_{i,j} = 1$ indicates that there exists edge $e_{ij} \in \mathcal{E}$ between the node v_i and v_j, and $\mathbf{A}_{i,j} = 0$ otherwise. Meanwhile, in this work, we focus on binary sensitive attributes and binary node classification tasks. Each node v_i has a sensitive attribute $s_i \in \{0, 1\}$, where $s_i = 0$ indicates that node v_i belongs to the deprived group $S_0 = \{\forall\ v_i : v_i \in \mathcal{V} \wedge s_i = 0\}$; if $s_i = 1$, v_i belongs to the *favored group* $S_1 = \{\forall\ v_i : v_i \in \mathcal{V} \wedge s_i = 1\}$. It is important to note that the sensitive attribute s_i is incorporated within the feature vector $\mathbf{X}_{i,:}$ of each node. In addition, we let $\mathcal{L}$ denote the set of labeled vertices, and let Y denote the corresponding set of ground-truth labels.

4 Methodology

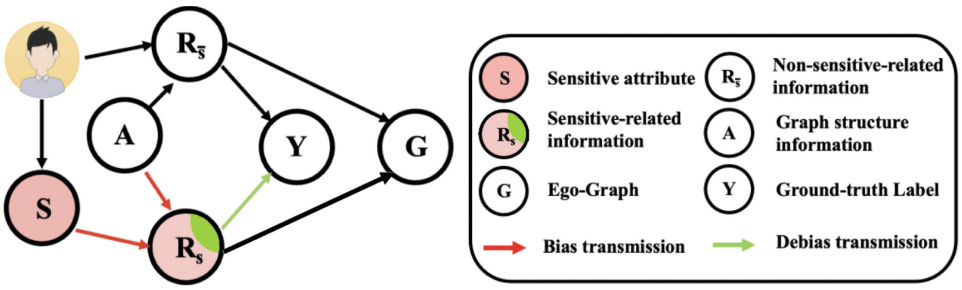

Fig. 1. The causal model of FDGNN with the red color denoting sensitive related information and white color representing non-sensitive related information, while green color is task-related information that is also related to the sensitive attribute.

4.1 Causal Model

This section introduces the proposed causal model, which is pivotal for examining counterfactual scenarios, *i.e.,* querying outcomes in a counterfactual world under certain conditions were altered. To address multi-source biases, a scenario

that exposes the limitations of the fairness notions solely based on statistics, a Structural Causal Model (SCM) is constructed from the observed graph, as depicted in Fig. 1. Specifically, SCM encapsulates causal relationships among five key variables: sensitive attribute (S), ground-truth label (Y), the graph structure (A), ego-graph (G), and information-related (R_S) or unrelated ($R_{\overline{S}}$) to sensitive attribute (S). In SCM, every connection denotes a deterministic causal link between variables, with the reasoning and explanations outlined as follows:

- $S \rightarrow R_S$: This link denotes that the node representation learned by the GNNs is influenced by sensitive attribute (S), thereby introducing bias into the final node representation. To enhance model fairness, we need to accurately identify R_S in the node representation, thus paving the way for mitigating bias in subsequent processes.
- $R_S \leftarrow A \rightarrow R_{\overline{S}}$: A impacts R_S and $R_{\overline{S}}$. For example, the connection between two nodes might stem from sensitive attribute-influenced interactions (*e.g.,* two individuals sharing the same neighborhood) or from non-sensitive factors (*e.g.,* common interests in activities such as soccer).
- $R_S \perp\!\!\!\perp R_{\overline{S}}$: To effectively address biases while minimizing the impact on performance, it's imperative to disentangle and isolate R_S from $R_{\overline{S}}$. This separation ensures that only R_S is adjusted to ensure fairness without unnecessarily compromising the information crucial for predicting Y.
- $R_S \rightarrow Y \leftarrow R_{\overline{S}}$: This model structure guarantees that both R_S and $R_{\overline{S}}$ influence the prediction of Y. The objective is to carefully modulate the impact of R_S to mitigate bias, *i.e.,* minimizing the sensitive information represented by red in R_S, while concurrently maintaining task-related information encapsulated within both R_S (*e.g.,* represented by green semicircle) and $R_{\overline{S}}$.
- $R_S \rightarrow G \leftarrow R_{\overline{S}}$: Same as the above substructure, but from a graph structure perspective, both R_S and $R_{\overline{S}}$ have direct causal effects on G, ensuring that it an accurate reconstruction of the ego-graph.

4.2 Framework Overview

Building upon the proposed causal model, a novel framework is designed to enhance the fairness of GNNs. This framework initially identifies accurate counterfactual instances from the existing samples. Subsequently, it utilizes de-identifiable sensitive attribute mechanisms to preserve task-relevant information while eliminating biased information from node representation. Figure 2 presents the overview of FDGNN, which incorporates three major phases. First, the Fair Ego-graph Generation Module aims to generate a subgraph for each node that contains important neighboring nodes while ensuring a fair and consistent representation of different subgroups. Second, the Counterfactual Data Augmentation Module finds accurate counterfactual instances to facilitate subsequent disentanglement learning. Last, the Fair Disentangled Representation Learning Module aims to perform sensitive information decomposition in node representation to keep task-relevant information while removing biased information through de-identifiable sensitive attributes. Each of these components will be introduced in the following sections.

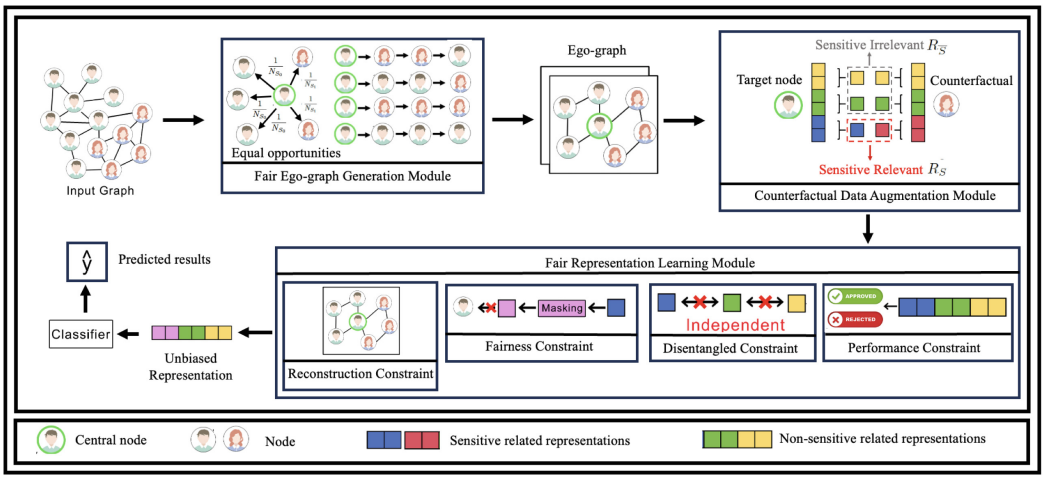

Fig. 2. Overview of the proposed FDGNN framework.

4.3 Fair Ego-Graph Generation Module

The inherent complexity of graph data poses a computational challenge to directly constructing causal models, especially for large-scale networks such as social networks. To this end, most existing methods aim to extract an ego graph for each node. This strategy is based on the local dependency assumption, *i.e.*, a node is primarily influenced by its nearest neighbors [15]. Despite the efficiency, it overlooks the critical aspect of local fairness within each node's ego graph. Specifically, the existing work may result in a biased ego graph, where an ego graph disproportionately consists of nodes sharing the same sensitive attribute. Such disparity in neighbor node distribution can result in over-association of the learned representation with sensitive attributes. In response, a Fair Ego-graph Generation Module is introduced to foster equitable representation across different subgroups within each ego-graph ($\mathcal{G}_{v_i}$) while avoiding limiting the distance of neighboring nodes, which can lead to the loss of important neighboring nodes. To achieve this, the concept of a *Related Score* (RS) for each node pair is introduced, inspired by PageRank [14], to quantify the relevance of node v_j to node v_i. Mathematically, it is represented as:

$$RS = \xi(I - (1 - \xi)\widetilde{A}) \tag{1}$$

where $\xi \in [0, 1]$ represents a parameter that controls the probability of a random walk restarting at the central node, while I is the identity matrix, and $\widetilde{A} = \mathbf{A}D^{-1}$ represents the transfer probability with D being the diagonal matrix where $D_{i,j} = \sum_j A_{i,j}$. Each entry of this matrix, denoted as $IS_{i,j}$, measures the relevance of node v_j to node v_i. Moreover, $IS_{i,:}$ denotes the vector of importance scores for node v_i.

However, PageRank, designed to assign uniform transfer probabilities to each neighboring node, can lead to biases, particularly in networks where nodes sharing the same sensitive attribute tend to form stronger connections, thereby skew-

ing transitions toward these neighbors. To address this, a fairness constraint is introduced to adjust these probabilities, promoting equitable representation among nodes from different subgroups. This fairness constraint categorizes neighbors based on their sensitive attributes and then adjusts the selection probabilities to balance the representation of each group during node transitions. As illustrated in Fig. 2, this adjustment grants male and female nodes probabilities of $\frac{1}{N_{S_0}}$ and $\frac{1}{N_{S_1}}$, respectively, ensuring an even representation of both subgroups in the sampling outcome. Mathematically, it is represented as:

$$\sum(P_{v_j}|\overline{A}_{i,j} = 1, s_j \in S_1) = \sum(P_{v_u}|\overline{A}_{i,j} = 1, s_u \in S_0) \tag{2}$$

where P_{v_j} and P_{v_u} represent the transition probabilities to neighboring nodes belonging to the deprived and favored groups, respectively.

4.4 Counterfactual Data Augmentation Module

With the learned fair ego-graph, central to the proposed causal framework SCM (*c.f.,* Sect. 4.1) is the distinction between sensitive attribute-related node representation (R_S, illustrated as red or blue squares within the red box in this module in Fig. 2) and sensitive information irrelevant node representation ($R_{\overline{S}}$, depicted as green or yellow squares within the gray box in this module in Fig. 2). To ensure accurate dissociation of these representations, identifying accurate counterfactual instances is essential. Specifically, consider a node v_i characterized by a factual sensitive attribute s_i and a corresponding label y_i. When flipping its sensitive attribute to $1 - s_i$, the representation independent of sensitive attributes, $R_{\overline{S}}$, should remain consistent, while the representation associated with sensitive attributes, R_S, should adapt to reflect this change. This forms the counterfactual subgraph $\mathcal{G}_{v_i}^C$, expressed mathematically as:

$$\mathcal{G}_{v_i}^C = \min \sum_{m=1}^{M} \left(d(\mathcal{G}_{v_i}, \mathcal{G}_{v_j}^m) | y_i = y_j, s_i \neq s_j \right) \tag{3}$$

where $\mathcal{G} = \{\mathcal{G}_{v_i} | v_i \in \mathcal{V}\}$, v_j denotes the corresponding counterfactuals of v_i, and $d(\cdot)$ measures the distance between pairs of ego-graphs.

Existing methods for generating graph counterfactual samples, as discussed in Sect. 1, may obtain inaccurate counterfactual samples. Therefore, we aim to find potential candidate counterfactual instances with the observed factual graphs. This strategy avoids making assumptions about how graphs that include sensitive attributes are generated while obviating the necessity for additional supervised signals to select counterfactuals. However, computing pairwise distances between ego-graphs becomes highly inefficient and impractical Given the complexity of graph structures and the vast search space of graph data. To address this issue, we aim to measure distances in the representation space, leveraging the captured graph structure and node attribute information to enhance computational efficiency. The task in Eq. 3 in thus reformulated as:

$$\mathcal{G}_{v_i}^{C} = \min_{h_j \in H} \sum_{m=1}^{M} \left(\|h_i - h_j\|_2^2 | y_i = y_j, s_i \neq s_j \right) \tag{4}$$

where $H = \{h_i | v_i \in \mathcal{V}\}$ is learned representation matrix and the L2 distance is employed to calculate the distance between h_i and h_j. Note that for each v_i, a set of counterfactual samples is obtained instead of one sample. Consequently, the counterfactual $\mathcal{G}_{v_i}^{c_i}$ can naturally extend to a set of counterfactuals consisting of M samples $\{\mathcal{G}_{v_i}^{C_i} | i = 1, \ldots, M\}$, where M is a constant number.

4.5 Fair Disentangled Representation Learning Module

FDGNN is now prepared to disentangle $R_{\overline{S}}$ and R_S within the node representation space, guided by the identified counterfactual instances. Besides, given that both $R_{\overline{S}}$ and R_S contain critical information for downstream tasks, our strategy aims to obtain informative yet sensitive-irrelevant node representation. To this end, we aim to segregate sensitive related information into a distinct component of the node representation and subsequently dissociate the sensitive related information within that component. This methodology prevents unnecessary performance degradation linked to enforcing fairness constraints on sensitive-relevant components, thereby minimizing performance loss while enhancing fairness. To effectively implement this disentanglement, the following four specific constraints are introduced:

1) Disentangled Constraint ($\mathcal{L}_D$). This constraint ensures the independence of $R_{\overline{S}}$ and R_S, preventing information leakage between them. To achieve this, we disentanglement the node representation into c distinct channels, with each channel influenced by a unique latent factor K, ensuring that they operate independently. Notably, only one of these factors, K_i, is associated with the sensitive attributes S, thus effectively segregating sensitive attribute-related information from the overall node representation. To assess the impact of different node neighbors on these partitioned representations, we employ an adaptive encoder configured as a multilayer perceptron (MLP). Specifically, for any pair of nodes v_i and v_j, their attributes x_i and x_j are input into the adaptive encoder ($\boldsymbol{\rho}_{v_i,v_j} = F_\rho([x_i, x_j])$) to evaluate the relevance of connection e_{ij} across c latent factors, where $\boldsymbol{\rho}_{v_i,v_j}$ is the vector of score indicates importance for e_{ij}, with $\rho_{v_i,v_j}^c \in \boldsymbol{\rho}_{v_i,v_j}$ representing scores for each latent factor c, and $F_\rho(\cdot)$ denoting the adaptive encoder operation. This score is normalized via a Softmax function to derive connection weights ω_{v_i,v_j}^c, as follows:

$$\omega_{v_i,v_j}^c = \text{Softmax}(\rho_{v_i,v_j}^c) \tag{5}$$

where ω_{v_i,v_j}^c represents the weight from node v_i to node v_j for channel c, indicating the likelihood that the connection is influenced by a latent factor c, with N_c reflecting the total number of channels.

Building on this, we further employ disentangled layers for graph convolution across multiple channels. Each disentangled layer comprises c channels of

graph convolution, all sharing the same network architecture, with each channel dedicated to a specific latent factor. Initially, we reduce the dimensionality of the original node attributes by projecting these attributes into different subspaces, each corresponding to a latent factor. For any given node representation R_{v_i}, a linear layer is employed for dimensionality reduction, transforming the representation from R_{v_i}-dimensional to N_c-dimensional space. This reduction operation $F_R(\cdot)$ is independently applied N_c times to generate N_c reduced node attributes, corresponding to the different latent factors K_i. Consequently, the disentangled node representation of a node v_i at the l^{th} layer, denoted by $h^l_{v_i}$, is formed by concatenating the reduced representations across all channels: $h^l_{v_i} = [r^l_{v_i,1}, r^l_{v_i,2}, \ldots, r^l_{v_i,N_c}]$. Extending this to all nodes, R^c represents the disentangled representations of all nodes within the c^{th} channel. Thus, h^l symbolizes the aggregated disentangled representations across all channels at layer l: $h^l = [R^l_1, R^l_2, \ldots, R^l_{N_c}]$.

However, the above process primarily addresses disentanglement at the sample level, neglecting the independence among latent factors, especially mutual independence across different channels. As depicted in Fig. 2, the goal is to achieve zero correlation between distinct channel representations, such as the blue, green, and yellow squares, to truly enhance the disentanglement process. To achieve this, we propose the Independence Constraint, mathematically formulated as:

$$\mathcal{L}_I = \sum_{c_1=1}^{N_c} \sum_{c_2 \neq c_1}^{N_c} \frac{D(K_{c_1}, K_{c_2})}{Norm(K_{c_1}, K_{c_2})} \tag{6}$$

where $D(\cdot)$ denotes the distance covariance, and $Norm(\cdot)$ represents the normalization function.

With fully disentangled channels, the next step is to pinpoint the latent factors that correlate with sensitive information. Counterfactual instances serve as a pivotal guide in this process. Specifically, for a given counterfactual (CI_i), its non-sensitive representation ($R^{CI_i}_{\overline{S}}$) is similar to the target sample, while its sensitive representation ($R^{CI_i}_{S}$) is distinct. By leveraging counterfactuals, facilitates a strategy aimed at minimizing the similarity between channels unrelated to sensitive attributes and maximizing it between channels that are related to sensitive attributes. Consequently, the Sensitive Identification Constraints are proposed:

$$\mathcal{L}_C = \frac{1}{|\mathcal{V}| \times M} \sum_{v_i \in \mathcal{V}} \sum_{m=1}^{M} \left[d(R^{v_i}_{\overline{S}}, R^{CI_m}_{\overline{S}}) - d(R^{v_i}_{S}, R^{CI_m}_{S}) \right] \tag{7}$$

where $d(\cdot)$ is a distance metric, with M indicating the size of the counterfactual sample set. By amalgamating $\mathcal{L}_I$ and $\mathcal{L}_S$, the Disentangled Constraint is formally defined as:

$$\mathcal{L}_D = \mathcal{L}_I + \mathcal{L}_C \tag{8}$$

2) Fairness Constraint ($\mathcal{L}_M$). The objective of this constraint is to disassociate the component associated with sensitive attributes. As demonstrated in this module in Fig. 2, it ensures that the gender information of the node from the purple square cannot be inferred, thereby preventing bias from impacting downstream tasks. To achieve this, we employ a de-identifiable sensitive attribute technique, which utilizes a learnable vector ($\mathbf{W}$) to remove the identifiability of sensitive attributes from node representations. This process transforms R_S into an unbiased representation, denoted as $\not{R}_S = R_S \odot \mathbf{W}$, making it indiscernible whether a node belongs to any specific subgroup. This unbiased representation $\not{R}_S$ and $R_{\overline{S}}$ are subsequently used for predicting instance labels. Further, we use covariance as a constraint to ensure the effective removal of sensitive information, aiming to minimize the absolute covariance between the sensitive attribute and the label predictions. Mathematically, this is expressed as:

$$\mathcal{L}_F = \sum_{i=1}^{d_{R_{v_i}}} Abs(\mathbb{E}\left[(S_{v_i} - \mathbb{E}(S_{v_i}))(\overline{R}_{v_i} - \mathbb{E}(\overline{R}_{v_i}))\right]), \tag{9}$$

where $\overline{R}_{v_i}$ denotes the unbiased node representation for node v_i. In addition, we let $\mathbb{E}(\cdot)$ indicate the expectation operation, and $\mathrm{Abs}(\cdot)$ is the absolute value function. This constraint ensures that the predictions are unbiased by sensitive attributes, thereby enhancing model fairness.

3) Performance Constraint ($\mathcal{L}_P$). Ensuring that the representations $R_{\overline{S}}$ and $\not{R}_S$ for each node v_i incorporate vital node attributes and neighborhood information is essential to uphold their utility for downstream tasks, thereby aiding accurate label predictions. Thus, the Performance Constraint is established to enforce alignment between the prediction $\hat{y}_i$ and the ground truth y_i:

$$\mathcal{L}_P = \frac{1}{|\mathcal{V}_L|} \sum_{v_i \in \mathcal{V}_L} -(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)) \tag{10}$$

where classifier takes $R_{\overline{S}}$ and $\not{R}_S$ as input and $\hat{y}_i$ is prediction results for node v_i.

4) Reconstruction Constraint ($\mathcal{L}_R$). For each node v_i, the learned representations $R_{\overline{S}}$ and $\not{R}_S$ should be sufficient to reconstruct the observed ego-graph G_{v_i}, transforming into an adjacency matrix reconstruction task. The effectiveness of node representation is thus evaluated by the discrepancies between the reconstructed adjacency matrices and the original graph structure. In addition, considering the sparsity of positive edges, FDGNN also incorporates negative sampling to address the distribution disparity between existent (positive) and non-existent (negative) edges. Specifically, for each positive edge $\{A(v_i, v_j) = 1 \; \forall \; i, j\}$, we counterpart this with a randomly selected non-existent edge $\{A(v_i, v_k) = 0 \; \forall \; i, k\}$, thereby forming a set of negative samples, M^-. Lastly, the Reconstruction Constraint, $\mathcal{L}_R$, is defined mathematically as:

$$\mathcal{L}_R = \sum_{A(v_i,v_j) \in M^+, A(v_i,v_k) \in M^-} \|\hat{A}(v_i, v_j) - A(v_i, v_j)\|_F^2 + \|\hat{A}(v_i, v_k) - A(v_i, v_k)\|_F^2 \tag{11}$$

where $\hat{A}$ and A are the predicted and observed adjacency matrices of input graph $\mathcal{G}$.

4.6 Final Optimization Objectives

The final objective function of FDGNN, as presented in Eq. 12, brings together the above three modules. Specifically, this function consists of four parts and is governed by the tunable hyperparameters α, β, and γ to balance the contributions of various elements: i) $\mathcal{L}_P$ aims to minimize the prediction loss, ii) $\mathcal{L}_I$ encourages the decomposition of learned representations into different independent channels and distinguishes between sensitive relevant and irrelevant representations, iii) $\mathcal{L}_F$ aims to mitigate sensitive-related information in node representation thereby improving the fairness of the model, and iv) $\mathcal{L}_R$ works to minimize the reconstruction loss for the node representations.

$$\min \mathcal{L}_{total} = \mathcal{L}_P + \alpha\mathcal{L}_D + \beta\mathcal{L}_F + \gamma\mathcal{L}_R \tag{12}$$

5 Experiment

5.1 Datasets

Experiments are conducted on three real-world graph datasets: i) The **German** dataset [2] contains credit information from clients at a German bank. Each node in this dataset represents a client, with edges reflecting the similarity between clients' credit profiles. The sensitive attribute is the clients' gender, and the classification task focuses on distinguishing clients into good versus bad credit risks. ii) The **Credit** dataset [36] consists of default payment records for individuals, where each node denotes an individual and edges indicate similarities in their expenditure and payment behaviors. The age of the individuals serves as the sensitive attribute, and the predictive task aims to determine whether an individual is likely to default on their credit card payments. iii) The **Bail** dataset [1] presents data related to defendants granted bail in U.S. state courts. Nodes represent defendants, and edges between nodes denote similarities in criminal records and demographic information. The sensitive attribute in this dataset is the race of the defendants, with the classification objective being to identify defendants as either suitable or unsuitable for bail (Table 1).

5.2 Evaluation Metrics

To effectively evaluate our proposed model, we measured our model performance from two perspectives: classification performance and fairness. For classification performance, we adopt Accuracy, F1-Score, and AUROC to evaluate the performance on node classification tasks. All three performance metrics close to 1 indicate better classification performance. To evaluate fairness, we use two commonly used fairness metrics, *i.e.,* Statistical Parity Difference (SPD) [20] and Equal Opportunity Differences (EOD) [12]. For both fairness metrics, values closer to 0 are indicative of greater model fairness.

Table 1. Summary of the datasets used in the experiments.

Dataset	German	Credit	Bail
Vertices	1,000	30,000	18,876
Edges	21,742	137,377	311,870
Feature dimension	27	13	18
Sensitive Attribute	Gender	Age	Race

5.3 Baselines

The proposed FDGNN is compared against seven state-of-the-art methods across three categories to evaluate its effectiveness. These include vanilla models like GCN [18], which leverages spatial graph convolutions for neighbor representation aggregation; GraphSAGE [11], which addresses GCN's scalability by training on node mini-batches; and GIN [33] enhancing node representation learning through MLP. Additionally, the fair node classification method FairGNN [7], which employs adversarial training to achieve group fairness by obscuring deprived group identities from discriminators, is also considered, in addition to graph counterfactual fairness methods like NIFTY [1], GEAR [21], and RFCGNN [27], detailed in Sect. 2.

5.4 Experiment Results

For thorough evaluation, the following research questions are addressed:

RQ1: How well does FDGNN performance compared to the state-of-the-art bias mitigation algorithms?

To answer RQ1, we experiment on three datasets with the comparison to the baselines on the node classification task. Each experiment is conducted 10 times, the results are shown in Table 2. As we can see, FDGNN outperforms all baseline methods across all evaluation metrics in most cases. Specifically, FDGNN demonstrates superior fairness performance, as evidenced by the significant margin overall baseline methods across all datasets. The enhancement of fairness is attributed to FDGNN accurately identifying sensitive attribute-related information via counterfactual instances and disentangling it into an independent component. It then mitigates its influence on prediction outcomes via a de-identifiable sensitive attribute mechanism. Simultaneously, FDGNN showcases commendable utility performance, surpassing other methods in most cases, which is indicative of FDGNN's capability to maintain important task-relevant information. This is because FDGNN avoids directly enforcing the fairness constraints by disentangling node representation, which facilitates the retention of task information related to sensitive attributes. Overall, the experimental results demonstrate the effectiveness of FDGNN in improving fairness while achieving comparable performance.

Table 2. Results on performance and fairness for FDGNN and baselines. The darkest cells indicate the top rank, while lighter cells represent the second rank.

Dataset	Methods Metrics	SPD (↓)	EOD (↓)	Accuracy (↑)	F1-Score (↑)	AUROC (↑)
German	GCN	0.364	0.312	0.684	0.786	0.654
	GraphSAGE	0.231	0.157	0.746	0.817	0.781
	GIN	0.148	0.091	0.720	0.812	0.734
	FairGNN	0.086	0.054	0.653	0.817	0.671
	NIFTY	0.077	0.049	0.674	0.792	0.736
	GEAR	0.085	0.046	0.681	0.780	0.722
	RFCGNN	0.067	0.041	0.721	0.823	0.747
	FDGNN	0.058	0.024	0.727	0.837	0.781
Credit	GCN	0.108	0.096	0.689	0.835	0.707
	GraphSAGE	0.113	0.124	0.739	0.859	0.767
	GIN	0.132	0.127	0.724	0.823	0.729
	FairGNN	0.126	0.104	0.674	0.812	0.711
	NIFTY	0.094	0.113	0.703	0.806	0.727
	GEAR	0.097	0.084	0.734	0.817	0.738
	RFCGNN	0.074	0.064	0.735	0.849	0.743
	FDGNN	0.056	0.047	0.736	0.861	0.747
Bail	GCN	0.093	0.044	0.828	0.784	0.871
	GraphSAGE	0.086	0.041	0.847	0.793	0.894
	GIN	0.072	0.043	0.728	0.658	0.768
	FairGNN	0.067	0.044	0.815	0.776	0.872
	NIFTY	0.035	0.028	0.753	0.671	0.796
	GEAR	0.047	0.024	0.823	0.783	0.786
	RFCGNN	0.031	0.024	0.861	0.802	0.747
	FDGNN	0.025	0.020	0.854	0.785	0.896

RQ2: What is the impact on FDGNN's performance when individual components are ablated?

To answer RQ2, we conduct ablation studies to gain insights into the effect of each module of FDGNN on improving fairness. Initially, our first analysis examined the significance of the fair ego-graph generation module. By substituting this module with the FDGNN-NFG variant, which employs an extractor to capture 2-hop neighboring nodes as ego-graphs for each node. Figure 3 presents ablation results on German, Credit, and Bail datasets. We observe that the fairness of FDGNN-NFG noticeably decreases. This reduction in fairness is ascribed to the FDGNN-NFG variant's inability to equitably represent diverse

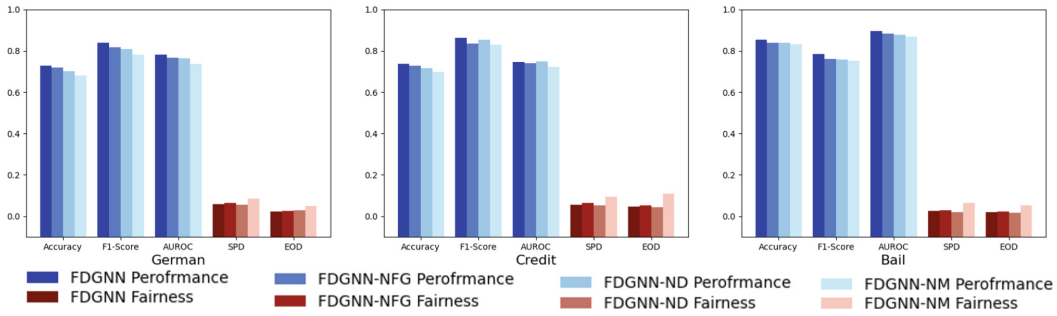

Fig. 3. Ablation study results for FDGNN, FDGNN-NFG, FDGNN-ND, and FDGNN-NM.

subgroups within subgraphs, leading to oversight of information from neighboring nodes with different sensitive attributes and, consequently, introducing graph structural bias. Next, we assessed the impact of the disentangled constraint by introducing the FDGNN-ND variant, which eschews this constraint by setting $N_c = 1$ and excluding $\mathcal{L}_D$. The results, depicted in Fig. 3, revealed a decline in both fairness and overall performance. This downturn can be attributed to the direct application of fairness constraints across the entire representation space in the absence of disentanglement, inevitably removing some task-related information. Lastly, we evaluate the effectiveness of our fairness constraint by creating the FDGNN-NF variant, removing $\mathcal{L}_F$. Compared to the FDGNN, there is a marked degradation in fairness, demonstrating the critical role of the fairness constraint in removing sensitive attribute information from node representation. To sum up, experimental results demonstrate the indispensability and efficacy of each component within the FDGNN framework.

RQ3: What the effect of Different N_c Values on Fairness and Predictive Performance?

To answer RQ3, we conducted experiments with a variety of values for N_c as $\{1, 2, 4, 8, 16\}$, keeping all other training factors the same. We compare FDGNN's predictive performance and fairness under different settings. We observe that (Fig. 4): i) As N_c increases, the FDGNN achieves better fairness, demonstrating better disentanglement of information related to sensitive attributes. ii) When N_c is a modest value, the model fairness is hardly affected or even increases. The FDGNN mostly strikes the right balance between maintaining model utility and fostering fairness with proper choices of N_c in here. iii) When N_c is significant, , a noticeable decline in fairness is observed. This is attributed to the model's inability to isolate sensitive attribute information within a singular channel, resulting in cross-channel correlations that retain sensitive attribute data within node representation. In essence, increasing N_c allows for finer disentanglement and recognition of sensitive attribute-related information up to a point. Beyond this threshold, however, the decomposition into an excessive number of chan-

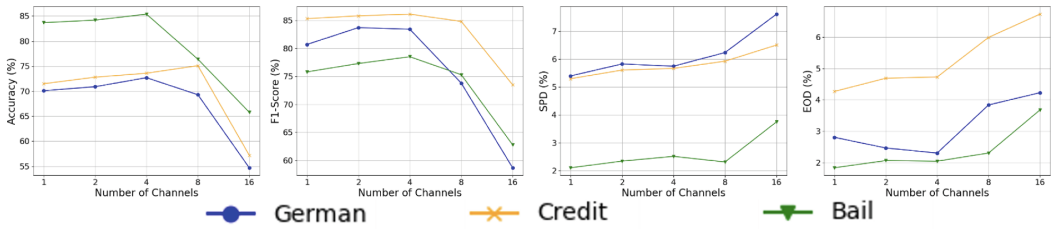

Fig. 4. Study the choice of N_c-value on German, Credit and Bail datasets.

nels introduces interference among them. This complexity hampers the model's ability to ensure channel independence, resulting in a decrease in model fairness.

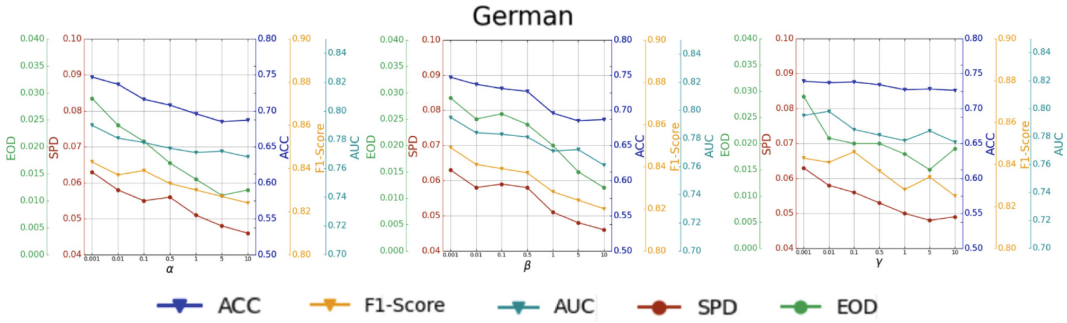

Fig. 5. Exploring hyperparameters study results in the German dataset.

RQ4: How do hyperparameters affect the performance and fairness of FDGNN?

To answer RQ4, we delve into the effects of three critical hyperparameters, *i.e.*, α, β, and γ, which respectively modulate the influence of disentanglement, decorrelation, and the model's reconstruction performance within FDGNN. For this analysis, we individually varied each hyperparameter through a range from 0.001 to 10, keeping all other training factors the same. Figure 5 presents the relevant findings from the German dataset. Specifically, an increase in α and β will increase model fairness but at the cost of some predictive performance degradation. This phenomenon occurs as the increased weightage of these parameters strengthens the model's ability to disentangle node representation and mitigate the correlation with sensitive attributes. Consequently, this diminishes the influence of sensitive attribute information on node representation, thereby advancing model fairness. As for γ, its increment initially bolsters fairness up to a certain point, beyond which fairness begins to decrease, though without significantly affecting performance. This is because a higher γ value improves the model's fidelity in reconstructing the graph structure, thereby avoiding the introduction of noise into the node representation and improving the model's ability to capture the underlying factors behind the data.

6 Conclusion

In this work, we study the problem of learning fair graph representation within GNNs. Inspired by causal theory, we introduce the Fair Disentangled Graph Neural Network (FDGNN) framework, which aims to achieve counterfactual fairness in graph-based representations while preserving important task-related information. FDGNN conducts a causal analysis of both original and counterfactual samples, effectively disentangling sensitive attributes into distinct components and subsequently mitigating their undue influence on the learned representations. This strategy allows FDGNN to enhance fairness without compromising the utility of the node representation for downstream tasks. Empirical evaluations on three real-world datasets validate the effectiveness of our framework with respect to both prediction performance and fairness.

Acknowledgement. This work was supported in part by the National Science Foundation (NSF) under Grant No. 2245895.

References

1. Agarwal, C., Lakkaraju, H., Zitnik, M.: Towards a unified framework for fair and stable graph representation learning. In: Uncertainty in Artificial Intelligence, pp. 2114–2124. PMLR (2021)
2. Asuncion, A., Newman, D.: UCI machine learning repository (2007)
3. Chinta, S.V., et al.: Optimization and improvement of fake news detection using voting technique for societal benefit. In: 2023 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 1565–1574. IEEE (2023)
4. Chu, Z., et al.: History, development, and principles of large language models-an introductory survey. arXiv preprint arXiv:2402.06853 (2024)
5. Chu, Z., Wang, Z., Zhang, W.: Fairness in large language models: a taxonomic survey. arXiv preprint arXiv:2404.01349 (2024)
6. Creager, E., et al.: Flexibly fair representation learning by disentanglement. In: International Conference on Machine Learning, pp. 1436–1445. PMLR (2019)
7. Dai, E., Wang, S.: 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, pp. 680–688 (2021)
8. Doan, T.V., Chu, Z., Wang, Z., Zhang, W.: Fairness definitions in language models explained (2024)
9. Dzuong, J., Wang, Z., Zhang, W.: Uncertain boundaries: multidisciplinary approaches to copyright issues in generative AI. arXiv preprint arXiv:2404.08221 (2024)
10. Gao, C., Wang, X., He, X., Li, Y.: Graph neural networks for recommender system. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, pp. 1623–1625 (2022)
11. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
12. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems, vol. 29 (2016)

13. He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., Wang, M.: LightGCN: simplifying and powering graph convolution network for recommendation. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 639–648 (2020)
14. Jeh, G., Widom, J.: Scaling personalized web search. In: Proceedings of the 12th International Conference on World Wide Web, pp. 271–279 (2003)
15. Jiao, Y., Xiong, Y., Zhang, J., Zhang, Y., Zhang, T., Zhu, Y.: Sub-graph contrast for scalable self-supervised graph representation learning. In: 2020 IEEE International Conference on Data Mining (ICDM), pp. 222–231. IEEE (2020)
16. Jin, G., Wang, Q., Zhu, C., Feng, Y., Huang, J., Zhou, J.: Addressing crime situation forecasting task with temporal graph convolutional neural network approach. In: 2020 12th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), pp. 474–478. IEEE (2020)
17. Kang, J., He, J., Maciejewski, R., Tong, H.: Inform: Individual fairness on graph mining. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 379–389 (2020)
18. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
19. Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
20. Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., Ntoutsi, E.: A survey on datasets for fairness-aware machine learning. Wiley Interdisc. Rev. Data Min. Knowl. Discov. **12**(3), e1452 (2022)
21. Ma, J., Guo, R., Wan, M., Yang, L., Zhang, A., Li, J.: Learning fair node representations with graph counterfactual fairness. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, pp. 695–703 (2022)
22. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. (CSUR) **54**(6), 1–35 (2021)
23. Olteanu, A., Castillo, C., Diaz, F., Kıcıman, E.: Social data: Biases, methodological pitfalls, and ethical boundaries. Front. Big Data **2**, 13 (2019)
24. Simonovsky, M., Komodakis, N.: GraphVAE: towards generation of small graphs using variational autoencoders. In: Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I. (eds.) ICANN 2018. LNCS, vol. 11139, pp. 412–422. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01418-6_41
25. Sui, Y., Wang, X., Wu, J., Lin, M., He, X., Chua, T.S.: Causal attention for interpretable and generalizable graph classification. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 1696–1705 (2022)
26. Wang, Z., et al.: Individual fairness with group awareness under uncertainty. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer Nature Switzerland (2024)
27. Wang, Z., Narasimhan, G., Yao, X., Zhang, W.: Mitigating multisource biases in graph neural networks via real counterfactual samples. In: 2023 IEEE International Conference on Data Mining (ICDM), pp. 638–647. IEEE (2023)
28. Wang, Z., Qiu, M., Chen, M., Salem, M.B., Yao, X., Zhang, W.: Towards fair graph neural networks via real counterfactual samples. Knowl. Inf. Syst. (2024). https://doi.org/10.1007/s10115-024-02161-z
29. Wang, Z., Saxena, N., et al.: Preventing discriminatory decision-making in evolving data streams. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT) (2023)

30. Wang, Z., Wallace, C., Bifet, A., Yao, X., Zhang, W.: FG^2AN: fairness-aware graph generative adversarial networks. In: Koutra, D., Plant, C., Gomez Rodriguez, M., Baralis, E., Bonchi, F. (eds.) Machine Learning and Knowledge Discovery in Databases: Research Track, ECML PKDD 2023, LNCS, vol. 14170, pp. 259–275. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43415-0_16
31. Wang, Z., et al.: Towards fair machine learning software: Understanding and addressing model bias through counterfactual thinking. arXiv preprint arXiv:2302.08018 (2023)
32. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Philip, S.Y.: A comprehensive survey on graph neural networks. IEEE Trans. Neural Netw. Learn. Syst. **32**(1), 4–24 (2020)
33. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? arXiv preprint arXiv:1810.00826 (2018)
34. Yang, Z., Pei, W., Chen, M., Yue, C.: Wtagraph: web tracking and advertising detection using graph neural networks. In: 2022 IEEE Symposium on Security and Privacy (SP), pp. 1540–1557. IEEE (2022)
35. Yazdani, S., Saxena, N., Wang, Z., Wu, Y., Zhang, W.: A comprehensive survey of image and video generative AI: recent advances, variants, and applications (2024)
36. Yeh, I.C., Lien, C.H.: The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Syst. Appl. **36**(2), 2473–2480 (2009)
37. Yin, Z., Wang, Z., Zhang, W.: Improving fairness in machine learning software via counterfactual fairness thinking. In: Proceedings of the 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings, pp. 420–421 (2024)
38. Yuan, H., Yu, H., Gui, S., Ji, S.: Explainability in graph neural networks: a taxonomic survey. IEEE Trans. Pattern Anal. Mach. Intell. **45**(5), 5782–5799 (2022)
39. Zhang, W., Wang, Z., Kim, J., Cheng, C., Oommen, T., Ravikumar, P., Weiss, J.: Individual fairness under uncertainty. In: 26th European Conference on Artificial Intelligence, pp. 3042–3049 (2023)
40. Zhang, W., Weiss, J.C.: Fair decision-making under uncertainty. In: 2021 IEEE International Conference on Data Mining (ICDM), pp. 886–895. IEEE (2021)
41. Zhang, W., Weiss, J.C.: Longitudinal fairness with censorship. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 12235–12243 (2022)
42. Zhang, W., Weiss, J.C., Zhou, S., Walsh, T.: Fairness amidst non-iid graph data: a literature review. arXiv preprint arXiv:2202.07170 (2022)
43. Zhang, W., Zhang, L., Pfoser, D., Zhao, L.: Disentangled dynamic graph deep generation. In: Proceedings of the 2021 SIAM International Conference on Data Mining (SDM), pp. 738–746. SIAM (2021)