Causality-Inspired Disentanglement for Fair Graph Neural Networks

Guixian Zhang[1,2], **Debo Cheng**[3], **Guan Yuan**[1,2], **Shang Liu**[1] and **Yanmei Zhang**[1]
[1]School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu, 221116, China
[2]Mine Digitization Engineering Research Center of the Ministry of Education, China University of Mining and Technology, Xuzhou, Jiangsu, 221116, China
[3]School of Computer Science and Technology, Hainan University, Haikou, Hainan, 570228, China
{guixian,yuanguan,shang,ymzhang}@cumt.edu.cn, chengdb2016@gmail.com

Abstract

Fair graph neural networks aim to eliminate discriminatory biases in predictions. Existing approaches often rely on adversarial learning to mitigate dependencies between sensitive attributes and labels but face challenges due to optimisation difficulties. A key limitation lies in neglecting intrinsic causality, which may lead to the entanglement of sensitive and causal factors, discarding causal factors or retaining sensitive factors in the final prediction, especially on unbalanced datasets. To address this issue, we propose a **C**ausality-inspired **D**isentangled framework for **F**air **G**raph neural networks (**CDFG**). In CDFG, node representations are conceptualised as a combination of causal and sensitive factors, enabling fair representation learning by only utilising the causal factors. We first use a counterfactual data generation mechanism to generate counterfactual data with similar causal factors but completely different sensitive factors. Then, we input real-world data and counterfactual data into the factor disentanglement module to achieve independence and disentanglement between the causal factors and sensitive factors. Finally, an adaptive mask module extracts the causal representation for fair and accurate graph-based predictions. Extensive experiments on three widely used datasets demonstrate that CDFG consistently outperforms existing methods, achieving competitive utility and significantly improved fairness.

1 Introduction

Graph Neural Networks (GNNs) have shown significant promise in handling structured data [Zhang *et al.*, ; Yan *et al.*, 2023], making them widely applicable in many fields [Guan *et al.*, 2024; Zhang *et al.*, 2025a; He *et al.*, 2025]. Despite their utility, recent research has highlighted a concerning issue: GNNs are prone to producing biased predictions [Dai and Wang, 2022; Wang *et al.*, 2022; Zhang *et al.*, 2024; Luo *et al.*, 2024; Zhang *et al.*, 2025b]. Particularly, in critical decision-making scenarios [Xu *et al.*, 2025; Yan *et al.*, 2025; Luo *et al.*, 2025; Li *et al.*, 2025], biased GNN predictions can result in unequal access to opportunities. For example, if historical data contain biases against students from certain regions, the model might incorrectly link academic success to geographic location. This could lead to the unfair rejection of applicants from less developed areas, even if they have outstanding academic achievements.

Recently, various GNN-based methods [Dai and Wang, 2022; Wang *et al.*, 2022; Dong *et al.*, 2022; Zhang *et al.*, 2024; Yang *et al.*, 2024; Luo *et al.*, 2024; Zhang *et al.*, 2025b] have been developed to enhance fairness without significantly sacrificing their performance. One common approach involves incorporating fairness constraints [Dai and Wang, 2022; Wang *et al.*, 2022; Yang *et al.*, 2024] to mitigate biases related to sensitive attributes (e.g., race or gender). For instance, FairGNN [Dai and Wang, 2022] performs adversarial training by setting up an additional discriminator that constrains the fairness of the representation, while FairSIN [Yang *et al.*, 2024] balances the distribution by aggregating neighbours with different sensitive attributes before performing adversarial training. These methods, however, often focus solely on statistical dependencies between data and labels, disregarding the underlying causal mechanisms that cause fairness issues.

Previous studies [Cheng *et al.*, 2024b; Meng *et al.*, 2025] have demonstrated that adversarial learning is difficult to optimise effectively. If adversarial training is performed directly on representations entangled with causal and sensitive factors, it can easily lead to losing some of the causal factors or retaining some of the sensitive factors in the prediction. Some studies [Dong *et al.*, 2022; Li *et al.*, 2024; Zhang *et al.*, 2025b] as shown in Fig. 1 (a) have attempted to use counterfactual inference to modify the data, but these methods tend to only focus on the node attributes [Dong *et al.*, 2022; Ma *et al.*, 2022; Li *et al.*, 2024; Zhang *et al.*, 2025b], ignoring the bias embedded in the adjacency matrix. Therefore, existing methods often suffer from overfitting of some groups, especially on unbalanced datasets.

Inspired by Reichenbach's Common Cause Principle [Reichenbach, 1991] and Independent Causal Mechanisms Principle [Parascandolo *et al.*, 2018], we assume that graph data consist of independent causal and sensitive factors. From a fairness perspective, we introduce a causal graph for GNNs [Pearl, 2009], as shown in Fig. 1 (b). In this causal graph, causal factors C and sensitive factors S are inter-

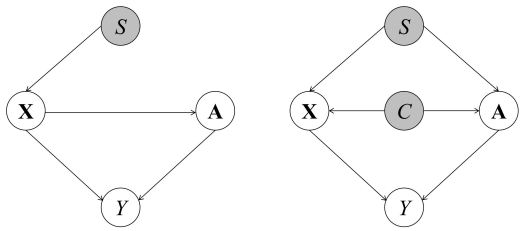


Figure 1: (a) Prior studies have tended to only consider the effect of sensitive factors on attributes and constrain the representation through adversarial learning. (b) The causal graph of GNN from a fairness perspective is presented in this paper. Both the adjacency matrix and node attributes of the graph data are affected by causal factors C and sensitive factors S.

twined, influencing both node attributes $\mathbf{X}$ and adjacency matrix $\mathbf{A}$. To achieve fair node representations, it is crucial to ensure that (1) causal factors C are separated from sensitive factors S, achieving disentanglement, and (2) causal factors C are sufficient for accurately predicting the label Y. However, in graph data, causal and sensitive factors are not directly observed and can interact in intricate ways. Additionally, GNNs introduce fairness issues in three main ways: (1) the original features of the nodes are correlated with sensitive attributes, leading to the unintentional inclusion of sensitive information in the learned representations; (2) social networks exhibit homophily [McPherson *et al.*, 2001], where similar nodes tend to connect, creating structural biases that are challenging to eliminate; (3) the message passing mechanism [Xu *et al.*, 2019] of GNN can make the sensitive information of a node leak to the neighbouring nodes, making the causal and sensitive factors from the adjacency matrix and node attributes further entangled.

To address these challenges, we propose a **C**ausality-inspired **D**isentangled framework for **F**air **G**raph neural networks (**CDFG**), which is designed to extract causal factors. Specifically, for the input graph data, we first utilise the counterfactual data generation module to generate counterfactual data with completely different sensitive factors S. Compared with the real-world data, the counterfactual data has similar causal factors C but different sensitive factors S. Subsequently, we constrain each dimension to be independent of each other through the factor disentanglement module and draw the distance between the counterfactual representation and the real-world representation to ensure that a portion of the dimensions of the real-world representation represents the causal factors C. Finally, the causal representation representing the causal factors C was identified through the adaptive mask module and used for prediction, achieving the best performance in both fairness metrics and the F1 metric. Our contributions are summarised as follows:

- We analyse the inherent causal mechanisms of graph representation learning from a fairness perspective and propose a new counterfactual generation mechanism.
- We propose a causality-inspired disentangled framework for fair GNNs, achieving causal representations and enabling fair graph representation learning.
- Comprehensive experiments on three widely used datasets demonstrate the effectiveness and fairness of our approach.

2 Realted Works

Graph Neural Networks (GNNs) have demonstrated a strong ability to learn representations of graph-structured data and have been used for a variety of tasks, such as node classification [Chen *et al.*, 2024b; Chen *et al.*, 2024a], graph classification [Wu *et al.*, 2024]. Their remarkable success in these different tasks has pushed GNNs to the forefront of research and applications [Fu *et al.*, 2023; Fu *et al.*, 2024; Guan *et al.*, 2024; Liu *et al.*, 2024; Liu *et al.*, 2022], extending their utility to critical decision-making systems. Therefore, the fairness of GNNs has received close attention.

Most of the fairness studies in GNN are based on statistical fairness [Kang *et al.*, 2020; Dong *et al.*, 2022; Zhang *et al.*, 2025b; Li *et al.*, 2024; Ma *et al.*, 2022], where group fairness [Berk *et al.*, 2021] is one of the most popular concepts, which aims to provide equal opportunities for each group. Most of the past work is based on adversarial learning to remove the effects brought by sensitive factors [Dai and Wang, 2022; Wang *et al.*, 2022; Dong *et al.*, 2022; Zhang *et al.*, 2025b; Li *et al.*, 2024; Ma *et al.*, 2022], but this may not be effective in removing sensitive factors due to the instability of adversarial learning [Cheng *et al.*, 2024b] or inadvertently lead to the removal of some causal factors [Cheng *et al.*, 2024a]. To address this limitation, some researchers have proposed a series of methods from the counterfactual perspective [Dong *et al.*, 2022; Zhang *et al.*, 2025b; Li *et al.*, 2024; Ma *et al.*, 2022], however, these methods do not take into account the inherent structural bias [Dong *et al.*, 2022] that exists in the original adjacency matrix. As a result, it cannot effectively address the overall sensitivity factors in graph data. Both node attributes and adjacency matrices are used as inputs to the GNN, and both are affected by the causal factor C and the sensitivity factor S. Therefore, in this paper, we propose a counterfactual method that considers both node attributes and adjacency matrices, effectively disentangles C and S, and achieves predictions with both fairness and utility.

3 Preliminaries

In the node classification task pipeline, the node representations are obtained through a GNN g, and then use a classifier f to make the final predictions. In the context of fair graph representation learning, we aim to disentangle the causal factors C from the sensitive factors S that influence the distribution $P(\mathbf{X}, \mathbf{A}, Y)$.

Theorem 1 (Reichenbach's Common Cause Principle [Reichenbach, 1991])**.** *If two random variables X and Y are statistically dependent ($X \not\perp Y$), then there exists a third variable Z that causally influences both. (As a special case, Z may coincide with either X or Y.) Furthermore, this variable Z screens X and Y from each other in the sense that given Z, they become independent, $X \perp Y \mid Z$.*

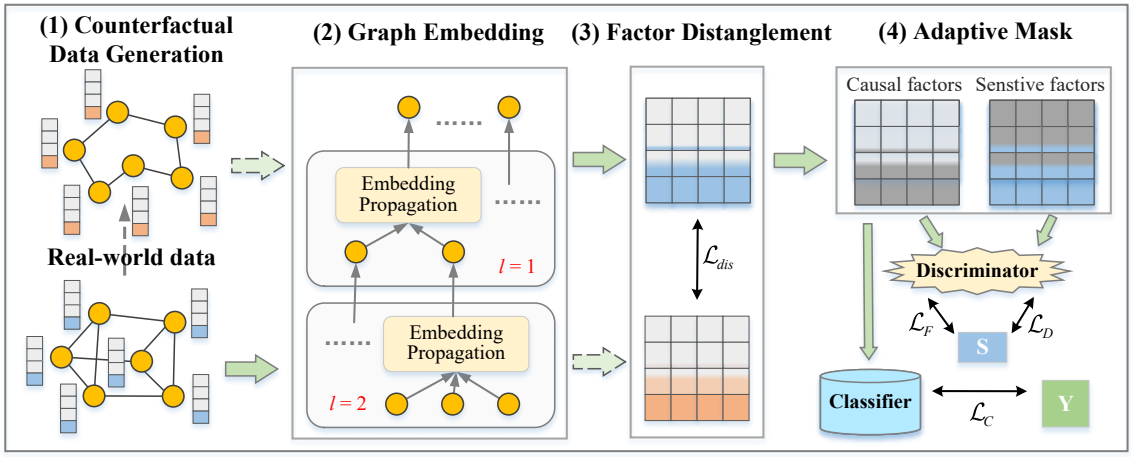


Figure 2: The overview of CDFG. After inputting the data, the whole framework is divided into four modules: (1) counterfactual data generation, (2) graph embedding, (3) factor disentanglement, and (4) adaptive mask. The final prediction and constraints are performed through the classifier and discriminator. The path of using counterfactual data is represented by dashed lines in the figure, and that of real-world data is represented by solid lines.

Based on Fig. 1, we can deduce that if we consider $\mathbf{X}$ and Y, as well as $\mathbf{A}$ and Y, are dependent. According to Theorem 1, there must be a variable C that causally influences both of them and makes $(\mathbf{X}, \mathbf{A})$ and Y conditionally independent given C, i.e., $(\mathbf{X}, \mathbf{A}) \perp Y \mid C$. If we can ensure that C can effectively predict Y and is not affected by sensitive factors, then we can achieve fair node classification through $P(Y \mid C)$. Fair GNNs need to eliminate the influence brought by sensitive information. Therefore, we refer to the variable related to sensitive information as sensitive factors S and hope that S and C are independent and non-redundant.

Theorem 2 (Independent Causal Mechanisms Principle [Parascandolo *et al.*, 2018])**.** *In a directed acyclic graph G, p(x) can be written as:*

$$p(x) = p(x_1, \ldots, x_d) = \prod_{j=1}^{d} p(x_j | PA_j^G) \tag{1}$$

where PA_j^G denotes the parents of variable x_j in the graph G.

From the Theorem 2, we can learn that each conditional $p(x_j|\mathrm{PA}_j^G)$ is considered a physical mechanism generating x_j from its parents and is referred to as a causal conditional [Lv *et al.*, 2022]. The ICM principle states that the causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other. Thus, causal factors C and sensitive factors S are independent of each other, which provides sufficient conditions for us to identify causal factors. To identify the causal factors C, we propose CDFG. This allows us to achieve fair graph representation learning, where the classifier f makes predictions based solely on the causal factors C.

4 The Proposed CDFG Method

In this section, we propose a new framework for fair GNNs.

4.1 Counterfactual Data Generation

To effectively disentangle the causal factor C from the mixture with the sensitive factor S, it is crucial to identify which components of the representation are associated with the causal factors. Counterfactual data, which intervene on the sensitive factor S of real-world data (i.e. the original data from the dataset), can aid in separating the causal factors C from S. The detailed method for achieving this disentanglement will be discussed in Section 3.2 and 3.3.

To rectify this issue, we have developed a method for generating counterfactual data that modifies both the adjacency matrix and the sensitive attributes simultaneously. We first generate random sensitive attributes for each node to create counterfactuals at the attribute level. Specifically, the modified feature matrix $\mathbf{X}'$ can be represented as:

$$\mathbf{x}'_{i,j} = \begin{cases} r_i, & \text{if } j = \text{the sensitive attribute id} \\ \mathbf{x}_{i,j}, & \text{otherwise} \end{cases} \tag{2}$$

Here, i denotes the sample index, j denotes the attribute index, and r_i is a binary value randomly sampled from the uniform distribution $\mathcal{U}(0, 1)$.

The adjacency matrix for counterfactual data is constructed based on the attributes of nodes. First, we calculate the node feature similarity using cosine similarity in the encoding stage. The feature similarity between two nodes is defined as:

$$M_{i,j} = \frac{\boldsymbol{x}'_i \cdot \boldsymbol{x}'_j}{|\boldsymbol{x}'_i||\boldsymbol{x}'_j|}, \tag{3}$$

where $\boldsymbol{x}'_i$ and $\boldsymbol{x}'_j$ are the feature vectors of nodes i and j, respectively.

Next, we calculate the adjacency matrix that reflects node feature relationships using a K-Nearest neighbours (KNN) algorithm [Zhang *et al.*, 2017]. We sort the nodes according to their similarity and select the K nodes with the highest similarity as the K nearest neighbours of v_i to construct the KNN graph:

$$\mathrm{KNN}(v_i) = \{v_i | v_i \in \mathrm{argmax}(M_{i,j})[0:\alpha]\}, \quad (4)$$

$$\mathbf{A'}_{i,j} = \begin{cases} 1, & \text{if } v_j \in \mathrm{KNN}(v_i) \text{ and } M_{i,j} > \tau, \\ 0, & \text{otherwise}, \end{cases} \quad (5)$$

where α denotes the number of neighbours of each node.

We can obtain representations of real-world data and counterfactual data through GNNs:

$$\mathbf{H} = \mathrm{GNN}(\mathbf{X}, \mathbf{A}), \mathbf{H}' = \mathrm{GNN}(\mathbf{X}', \mathbf{A}'). \quad (6)$$

Due to space constraints, we have included a detailed description of GNN in the Appendix A .

4.2 Factor Disentanglement

Counterfactual data essentially involve a complete modification of the sensitive factor S, resulting in two sets of data that possess similar causal factors but entirely distinct sensitive factors. To achieve the disentanglement of causal and sensitive factors, it is crucial to ensure that a portion of the representation can represent the causal factors. To accomplish this, we should ensure that (1) the dimensions of the node representations are mutually independent, and (2) the causal factors remain unchanged despite interventions on the sensitive factors.

For the representations $\mathbf{H}$ of the real-world data and $\mathbf{H}'$ of the counterfactual data, we first establish the cross-correlation matrix between the two representations:

$$R = \frac{1}{n}\mathbf{H'}^T\mathbf{H}, \quad (7)$$

where n denotes the number of nodes.

To maintain independence among the dimensions of the representations, we expect the off-diagonal elements of C to be close to 0:

$$\mathcal{L}_{off-diag} = \frac{1}{d(d-1)} \sum_{i=1}^{d} \sum_{j \neq i} R_{ij}^2, \quad (8)$$

where d is the dimension of the representation.

To effectively disentangle the causal factors from the mixture of non-causal factors, we want the positions of the dimensions influenced by the causal factors to remain unchanged despite interventions on the sensitive factors. Considering that the causal factors in both representations should be similar, we constrain the similarity between each dimension of the two representations:

$$\mathcal{L}_{diag} = \frac{1}{d} \sum_{i=1}^{d} (R_{ii} - 1)^2. \quad (9)$$

https://github.com/shawn-dm/CDFG/blob/main/Appendix.pdf

By minimizing $\mathcal{L}_{dis}$, we can ensure that the representations are disentangled:

$$\mathcal{L}_{dis} = \gamma * \mathcal{L}_{diag} + \delta * \mathcal{L}_{off-diag}. \quad (10)$$

4.3 Adaptive Mask Module

With the Factor Disentanglement module, we obtain representations that may have some dimensions that carry more causal factors but include fewer sensitive factors. Therefore, we design a learnable mask to identify which channels are mainly associated with causal factors. We use the Gumbel-Softmax trick [Jang *et al.*, 2017] to optimise masks for fair representation learning by finding these dimensions that approximate causal factors:

$$\mathbf{m} = \text{Gumbel-Softmax}(\hat{w}(\mathbf{h}_i), \kappa d), \quad (11)$$

where $\hat{w}$ denotes the contribution to learn each dimension, and the dimensions corresponding to the maximum $\kappa \in (0, 1)$ ratio is considered as the causal dimensions, while the rest of the dimensions are considered as the sensitive dimensions.

By applying the learned mask $\mathbf{m}$ and its complement $1-\mathbf{m}$ to the representations, we can extract the causal and sensitive representations, respectively. In this section, we utilize two multilayer perceptrons (MLPs) to function as a classifier f and a discriminator r. Specifically, we aim for the causal representation to accurately predict the label Y, and the sensitive representation to accurately identify the sensitive attribute S:

$$\mathcal{L}_C = \mathbb{E}_{v_i \sim \mathcal{V}} \left[\ell(f(\mathbf{h}_i \odot \mathbf{m}), y_i) \right], \quad (12)$$

$$\mathcal{L}_D = \mathbb{E}_{v_i \sim \mathcal{V}} \left[\ell(r(\mathbf{h}_i \odot (1 - \mathbf{m})), s_i) \right], \quad (13)$$

where y_i denotes the label of v_i, s_i denotes the sensitive attribute of v_i and ℓ denotes the cross entropy.

The proposed adaptive mask module can accurately detect causal dimensions, but to ensure that these causal dimensions are not influenced by sensitive factors, we have introduced additional fairness constraints on the causal dimensions:

$$\mathcal{L}_F = -\mathbb{E}_{v_i \sim \mathcal{V}} \left[\ell(r(\mathbf{h}_i \odot \mathbf{m}), s_i) \right]. \quad (14)$$

Finally, we can get the overall training objective:

$$\mathcal{L}_f = \mathcal{L}_{dis} + \mathcal{L}_C + \mathcal{L}_D + \beta * \mathcal{L}_f. \quad (15)$$

5 Experiments

Our experiments are designed to answer the following research questions (RQs): **RQ1**: How effective is our proposed method compared to the state-of-the-art fair graph representation learning method? **RQ2**: How does each module of our proposed method contribute to the final performance? **RQ3**: How do different methods of counterfactual data generation affect final performance? **RQ4**: How do the hyperparameters in the method affect the performance?

5.1 Experimental Settings

Datasets. We conducted experiments on three widely used real-world datasets, namely German [Dua and Graff, 2017], Bail [Jordan and Freiburger, 2015], and Credit [Yeh and Lien, 2009]. The statistical details of the datasets are presented in Table 1 and Table 2.

Dataset	Credit	German	Bail
#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

Table 1: A summary of the datasets.

	Credit		German		Bail	
	s=0	s=1	s=0	s=1	s=0	s=1
y=0	5906	730	191	109	5457	6315
y=1	21409	1955	499	201	3860	3244

Table 2: Different groups in the datasets.

Baselines. In the experiments, we compared the proposed EAGNN method with nine state-of-the-art methods. These methods can be categorized into two groups: **Vanilla GNNs** and **Fair GNNs**. These methods represent a range of approaches to graph representation learning, from traditional GNN architectures to more recent methods that incorporate fairness constraints. The Vanilla GNNs include: GCN [Kipf and Welling, 2017], GIN [Xu *et al.*, 2019] , and SAGE [Hamilton *et al.*, 2017]. The **Fair GNNs** include: FairGNN [Dai and Wang, 2022] uses adversarial training to achieve fairness on graphs; EDITS [Dong *et al.*, 2022] improves fairness through preprocessing; NIFTY [Agarwal *et al.*, 2021] simply flips the sensitive attributes to obtain counterfactual data and trains under fairness constraints; FVGNN [Wang *et al.*, 2022] effectively addresses the changes in feature correlations during propagation through a feature masking strategy, thereby eliminating discriminative bias; FairMILE [He *et al.*, 2023] is a multi-level fair graph representation learning framework. FairSIN [Yang *et al.*, 2024] achieves distributional balance by emphasising neighbours with different sensitive attributes for each node.

Evaluation metrics and implementation details. In this paper, we assess the efficacy of our proposed method, EAGNN, using the F1 score (F1) and accuracy (ACC) as evaluation metrics. The higher the value of these metrics, the more accurate the model's decisions. Accuracy is a measure of the proportion of correctly predicted samples relative to the total number of samples. F1 score offers a balanced measure between precision and recall, which is essential in domains such as medical diagnosis and fraud detection, where both false negatives and false positives are highly undesirable. For fairness evaluation, we focus on group fairness [Berk *et al.*, 2021] and use two quantitative metrics to evaluate it: Δ_{SP} [Dwork *et al.*, 2012] and Δ_{EO} [Hardt *et al.*, 2016]. We place the specific descriptions of the two fairness metrics in Appendix B. Consistent with prior studies [Agarwal *et al.*, 2021; Wang *et al.*, 2022], the datasets are partitioned into three distinct phases: training, validation, and testing. All Fair GNNs utilize SAGE as the encoder, which is described in Appendix A. The Adam optimization algorithm is applied uniformly across all models. We set the hidden layer size uniformly to 16 and κ to 0.5. Hyperparameters were optimized using a grid search approach, with a comprehensive hyperparameter analysis provided in Section 5.3.

5.2 Performance Comparison

As shown in Table 3, we present the comparative experimental results of CDFG with the current state-of-the-art methods. The results demonstrate that CDFG outperforms several state-of-the-art methods in terms of both utility and fairness, highlighting its potential for practical deployment in sensitive domains. We will analyse the comparison experiment from both the fairness results and the utility results. **Fairness Results.** The causality-inspired disentanglement method achieves the best fairness results on all three datasets. The fairness metrics Δ_{SP} and Δ_{EO} are significantly reduced, indicating that the model's decisions are more equitable. This is crucial in applications where fairness is a critical concern, such as credit assessment and judicial decision-making. This demonstrates that we have effectively disentangled the causal and sensitive factors. CDFG learns causal representations that approximate the causal factors, avoiding the influence of sensitive factors, and thereby ensuring the fairness of the causal representations. **Utility Results.** The causality-inspired disentanglement method achieves satisfactory utility results on all three datasets. Compared to carefully designed fair GNN models, our approach obtains competitive and even superior performance. Particularly in the F1 score, we achieve the best results. Accuracy, while a straightforward measure of the proportion of correctly predicted samples, can be misleading in imbalanced datasets. It may be skewed by the frequent occurrence of certain labels and sensitive attribute subsets, leading to misleading conclusions. Therefore, the F1 score is a more reliable metric in such scenarios. It provides a balanced measure between precision and recall, which is crucial in fields such as medical diagnosis and fraud detection where false negatives and false positives are equally undesirable. Some of the methods have high ACC but low F1 on unbalanced datasets, probably due to overfitting some of the groups and not essentially learning a fair representation .

5.3 Ablation Study

To address **RQ2** and validate the effectiveness of our proposed method, we constructed three variants of CDFG: (1) removing the counterfactual data generation module, only constraining the independence of each dimension and using the adaptive mask to find causal factors (w/o CF); (2) removing the factor disentanglement module, treating counterfactual data as part of the training data (w/o FD); (3) removing the adaptive masking module, directly constraining and predicting the node representations (w/o AM). The experimental results are shown in Table 4. From the experimental results, we can draw the following conclusions: (1) All three variants

Dataset	Metric	GCN	GIN	SAGE	FairGNN	NIFTY	FVGNN	EDITS	FairMILE	FairSIN	CDFG
Credit	ACC (↑)	73.62±0.06	75.30±2.86	74.20±0.60	75.44±3.28	73.80±4.75	76.06±4.37	83.73±0.73	80.18±0.27	78.14±0.39	77.74±0.27
	F1 (↑)	81.88±0.06	84.56±2.17	82.45±0.52	81.35±1.83	81.21±0.59	84.43±4.23	76.93±0.89	87.16±0.17	87.27±0.48	87.32±0.25
	$\Delta_{SP}(\downarrow)$	12.93±0.26	5.14±0.96	16.35±2.36	10.46±5.69	8.09±2.77	6.06±3.63	7.28±0.49	1.21±0.39	1.29±0.54	0.44±0.53
	$\Delta_{EO}(\downarrow)$	10.65±0.18	3.79±0.64	14.12±2.64	9.47±6.10	7.41±1.54	3.90±3.54	5.09±0.78	0.84±0.14	0.61±0.80	0.44±0.53
German	ACC (↑)	72.45±0.75	70.32±1.55	71.63±1.35	70.83±1.66	66.24±4.12	69.60±1.13	65.60±6.81	70.08±1.48	69.52±0.96	69.84±0.20
	F1 (↑)	81.73±2.31	81.58±0.56	81.08±1.04	79.57±2.61	78.27±1.25	81.33±0.55	77.89±6.06	80.87±0.94	81.75±0.92	82.17±0.22
	$\Delta_{SP}(\downarrow)$	20.36±5.27	6.70±4.92	14.33±5.11	6.21±2.34	8.03±7.19	2.50±3.01	4.35±4.29	1.40±0.99	1.55±0.70	0.42±0.75
	$\Delta_{EO}(\downarrow)$	19.71±5.19	5.80±3.32	12.53±7.56	5.36±2.07	4.40±4.18	1.26±1.07	4.41±3.81	0.78±0.61	1.95±1.02	0.57±1.08
Bail	ACC (↑)	82.49±0.82	82.93±0.53	87.44±1.34	83.56±2.70	80.11±5.39	87.61±1.30	83.15±2.96	87.48±0.28	88.35±0.62	89.16±1.60
	F1 (↑)	77.52±1.35	77.28±0.58	81.57±1.19	78.37±1.99	79.85±3.16	82.67±0.87	80.42±2.53	82.52±0.50	83.54±0.68	85.60±1.92
	$\Delta_{SP}(\downarrow)$	9.31±2.12	7.74±1.19	8.14±1.08	6.88±1.41	5.96±2.13	3.49±1.74	6.57±1.35	3.17±0.21	1.27±0.78	0.53±0.43
	$\Delta_{EO}(\downarrow)$	8.59±1.13	6.77±0.81	7.43±1.75	5.77±1.48	5.57±1.69	2.42±1.29	5.61±1.73	1.72±0.56	1.05±0.74	0.83±0.89

Table 3: We conducted comparative experiments on three real-world datasets to evaluate the effectiveness and fairness of the models. The best results for each metric are highlighted in dark brown, and the second-best results are highlighted in light brown.

Dataset	Metric	w/o CF	w/o FD	w/o AM	CDFG
Credit	ACC (↑)	74.50±4.02	76.82±2.82	74.26±9.31	77.74±0.27
	F1 (↑)	84.06±4.14	86.41±2.33	86.68±9.63	87.32±0.25
	$\Delta_{SP}(\downarrow)$	2.07±2.79	1.16±0.97	1.67±1.12	0.44±0.53
	$\Delta_{EO}(\downarrow)$	1.50±1.98	1.16±0.99	0.98±0.53	0.44±0.53
German	ACC (↑)	69.92±0.78	69.92±0.78	69.60±0.25	69.84±0.20
	F1 (↑)	81.89±0.37	81.96±0.36	81.94±0.12	82.17±0.22
	$\Delta_{SP}(\downarrow)$	3.53±4.16	1.65±1.25	0.89±0.82	0.42±0.75
	$\Delta_{EO}(\downarrow)$	2.10±3.18	2.08±1.02	0.97±0.57	0.57±1.08
Bail	ACC (↑)	85.91±0.64	88.48±0.79	87.28±0.58	81.62±0.53
	F1 (↑)	80.75±1.04	84.18±0.79	81.62±0.53	85.60±1.92
	$\Delta_{SP}(\downarrow)$	1.36±0.87	0.64±0.55	0.76±0.78	0.53±0.43
	$\Delta_{EO}(\downarrow)$	1.53±1.06	1.37±1.00	1.16±0.97	0.83±0.39

Table 4: Ablation study results. We removed the CounterFactual data generation module (CF), the Factor Disentanglement module (FD), and the Adaptive Masking module (AM), respectively.

Dataset	Metric	F	R	A&F	A&R
Credit	ACC (↑)	76.65±2.72	76.67±2.80	77.30±1.67	77.74±0.27
	F1 (↑)	86.23±2.32	86.18±2.45	86.61±1.50	87.32±0.25
	$\Delta_{SP}(\downarrow)$	1.23±1.23	3.12±3.13	1.05±0.92	0.44±0.53
	$\Delta_{EO}(\downarrow)$	0.93±0.81	2.17±2.04	0.72±0.68	0.44±0.53
German	ACC (↑)	70.00±0.25	69.60±0.72	70.16±0.70	69.84±0.20
	F1 (↑)	81.74±1.05	81.23±1.84	82.11±0.29	82.17±0.22
	$\Delta_{SP}(\downarrow)$	2.45±4.80	4.36±7.19	1.56±1.54	0.42±0.75
	$\Delta_{EO}(\downarrow)$	1.60±2.44	3.47±4.99	1.08±0.98	0.57±1.08
Bail	ACC (↑)	82.44±3.33	85.17±1.20	87.43±0.74	81.62±0.53
	F1 (↑)	76.39±4.61	79.43±1.32	82.98±1.16	85.60±1.92
	$\Delta_{SP}(\downarrow)$	2.04±1.81	1.31±0.67	0.92±0.70	0.53±0.43
	$\Delta_{EO}(\downarrow)$	1.90±1.05	1.01±0.62	0.88±0.30	0.83±0.39

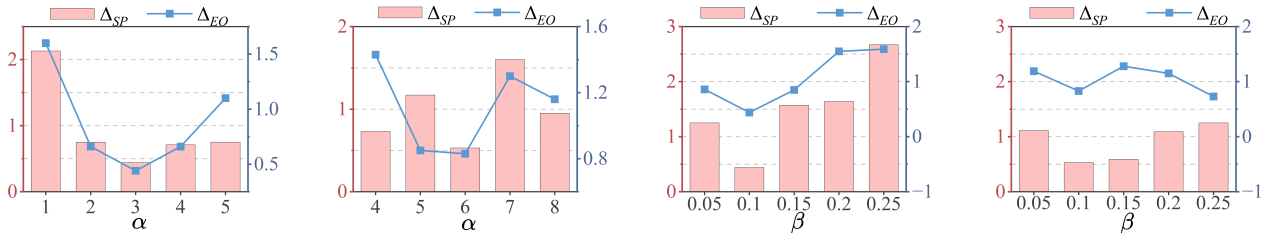
Table 5: (1) **F**lipping sensitive attributes (**F**); (2) **R**andomising sensitive attributes (**R**); (3) reconstructing the **A**djacency matrix with **F**lipped sensitive attributes (**A&F**); (4) reconstructing the **A**djacency matrix with **R**andomised sensitive attributes (**A&R**).

perform worse than CDFG in terms of both utility and fairness, demonstrating the effectiveness of each module and the rationality of the combination. This indicates that each component plays a crucial role in achieving the desired outcomes. (2) When the counterfactual data generation module is removed, the model achieves the worst experimental results. This highlights the necessity of counterfactual data for identifying causal factors. It shows that our method effectively intervenes on sensitive factors and successfully distinguishes between sensitive and causal factors through the factor disentanglement module. (3) When either the factor disentanglement module or the adaptive masking module is removed, the model still achieves relatively good results, but not as good as CDFG. This further emphasizes the importance of counterfactual data in the overall framework. The combination of these modules is essential for achieving both high utility and fairness in the learned representations.

To answer **RQ3**, we construct three variants of the counterfactual data generation moudle. The experimental results are shown in Table 5, from which we can get the following conclusions: (1) A&R variant achieves the best fairness results and F1 scores. This indicates that our method effectively intervenes on sensitive factors, providing a favourable condition for subsequent factor disentanglement. By randomizing sensitive attributes and reconstructing the adjacency matrix, we ensure that the model learns representations that are both fair and useful. (2) All variants that involve adjacency matrix reconstruction achieve excellent fairness results and F1 scores. This highlights the importance of the adjacency matrix in graph data, which contains non-negligible sensitive factors. Reconstructing the adjacency matrix helps to mitigate the influence of these sensitive factors, leading to more fair and accurate representations. (3) **F** and **R** results in worse and more fluctuating results compared to **w/o CF**. This suggests that focusing solely on node attributes can mislead the model, causing it to confuse sensitive factors with causal factors. This confusion leads to unstable training and suboptimal performance.

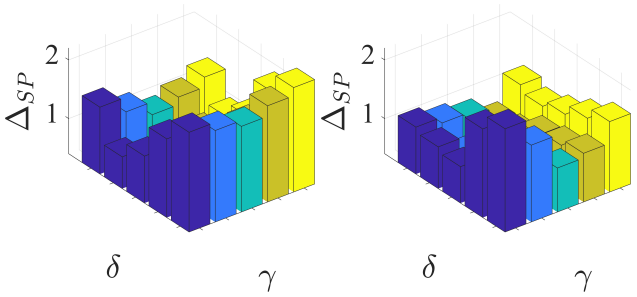
5.4 Hyper-parameter Analysis

To answer **RQ4**, we perform the hyper-parameter sensitivity analyses on the Credit and Bail datasets for α, β, γ and δ. Due to space constraints, we place the experimental results on the German dataset and the analysis of κ in Appendix C.



(a) α on Credit (b) α on Bail (c) β on Credit (d) β on Bail

Figure 3: Fairness performance under different values of α and β.



(a) Credit (b) Bail

Figure 4: Fairness performance under different values of γ and δ.

In the counterfactual data generation module, a critical parameter is the K-value α used for adjacency matrix reconstruction. The experimental results, as shown in Fig. 3 (a) - (b), provide insights into the optimal values of α for different datasets. For datasets with a higher average node degree (Credit and German), a smaller K-value (α = 3) is effective. For datasets with a lower average node degree (Bail), a larger K-value (α = 6) is necessary. This approach ensures that the reconstructed adjacency matrix is sufficiently different from the original, thereby effectively intervening on sensitive factors and improving the fairness and utility of the learned representations. This finding underscores the importance of considering both attribute and structural changes when generating counterfactual data for fair graph representation learning.

The hyperparameter β represents the weight of the fairness constraint for the causal representation. The experimental results for β are shown in Fig. 3 (c) - (d). For all three datasets, the model achieves the best fairness metrics when β = 0.2. When β is set to a larger value, it can lead to suboptimal disentanglement between causal and sensitive factors. This results in a decrease in fairness performance.

In the factor disentanglement module, the hyperparameters γ and δ play crucial roles in achieving effective disentanglement of factors. We set γ to $\{0.001\text{-}0.005\}$ and δ to $\{0.0001\text{-}0.0005\}$. The experimental results for these hyperparameters are shown in Fig 4. Due to space constraints, we present the experimental results for the Δ_{SP} metrics here and place the experimental results for the Δ_{EO} metrics in Appendix C. For the Credit and German datasets, which are relatively dense, lower values of γ (0.002) and δ (0.0003) are sufficient to achieve effective disentanglement. For the Bail dataset, which is relatively sparse, larger values of γ (0.5) and δ (0.001) are required to achieve effective disentanglement. This may signify that causal and sensitive factors in the representation of sparse data are more difficult to disentangle.

6 Conclusion

In this paper, we formalise the causal mechanism of GNN from a fairness perspective and propose a causally inspired CFDG framework. By leveraging counterfactual data with similar causal factors but different sensitive factors, we ensure the independence of causal and sensitive factors in node representations through the factor disentanglement module. Finally, causal representations independent of sensitive factors are extracted using the adaptive mask module. Extensive experiments demonstrate that the CDFG framework has the best results in both fairness and utility. This paper provides a new causal perspective on the area of fair GNNs, and we hope this work inspires further advancements in the research community. In future work, we will continue to explore more complex architectures to achieve superior results.

Acknowledgments

This work was supported in part by the Nature Science Foundation of China under Grant No.62277046, the Key Research and Development Program of Xuzhou under Grant No.KC23296, the Science and Technology Program of Xuzhou under Grant No.KC22047, 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, Research Fund of Guangxi Key Lab of Multi-source Information Mining & Security (MIMS24-13).

Contribution Statement

Guixian Zhang and Debo Cheng contributed equally to this paper, and Guan Yuan is the corresponding author.

References

[Agarwal *et al.*, 2021] Chirag Agarwal, Himabindu Lakkaraju, and Marinka Zitnik. Towards a unified framework for fair and stable graph representation

learning. In *Uncertainty in Artificial Intelligence*, pages 2114–2124. PMLR, 2021.

[Berk *et al.*, 2021] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

[Chen *et al.*, 2024a] Jinsong Chen, Boyu Li, Qiuting He, and Kun He. Pamt: A novel propagation-based approach via adaptive similarity mask for node classification. *IEEE Transactions on Computational Social Systems*, 11(5):5973–5983, 2024.

[Chen *et al.*, 2024b] Jinsong Chen, Hanpeng Liu, John Hopcroft, and Kun He. Leveraging contrastive learning for enhanced node representations in tokenized graph transformers. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems*, pages 85824–85845, 2024.

[Cheng *et al.*, 2024a] Debo Cheng, Ziqi Xu, Jiuyong Li, Lin Liu, Jixue Liu, Wentao Gao, and Thuc Duy Le. Instrumental variable estimation for causal inference in longitudinal data with time-dependent latent confounders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11480–11488, 2024.

[Cheng *et al.*, 2024b] Xu Cheng, Hao Zhang, Yue Xin, Wen Shen, and Quanshi Zhang. Clarifying the behavior and the difficulty of adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11507–11515, 2024.

[Dai and Wang, 2022] Enyan Dai and Suhang Wang. Learning fair graph neural networks with limited and private sensitive attribute information. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7103–7117, 2022.

[Dong *et al.*, 2022] Yushun Dong, Ninghao Liu, Brian Jalaian, and Jundong Li. Edits: Modeling and mitigating data bias for graph neural networks. In *Proceedings of the ACM web conference 2022*, pages 1259–1269, 2022.

[Dua and Graff, 2017] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[Dwork *et al.*, 2012] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[Fu *et al.*, 2023] Xingcheng Fu, Yuecen Wei, Qingyun Sun, Haonan Yuan, Jia Wu, Hao Peng, and Jianxin Li. Hyperbolic geometric graph representation learning for hierarchy-imbalance node classification. In *Proceedings of the ACM Web Conference 2023*, pages 460–468, 2023.

[Fu *et al.*, 2024] Xingcheng Fu, Yisen Gao, Yuecen Wei, Qingyun Sun, Hao Peng, Jianxin Li, and Xianxian Li. Hyperbolic geometric latent diffusion model for graph generation. In *Proceedings of the 41st International Conference on Machine Learning*, pages 14102–14124, 2024.

[Guan *et al.*, 2024] Renxiang Guan, Zihao Li, Wenxuan Tu, Jun Wang, Yue Liu, Xianju Li, Chang Tang, and Ruyi Feng. Contrastive multiview subspace clustering of hyperspectral images based on graph convolutional networks. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

[Hamilton *et al.*, 2017] William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1025–1035, 2017.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3323–3331, 2016.

[He *et al.*, 2023] Yuntian He, Saket Gurukar, and Srinivasan Parthasarathy. 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, 2023.

[He *et al.*, 2025] Ludan He, Debo Cheng, Guixian Zhang, and Shichao Zhang. Leveraging long-range nodes in multi-view graph contrastive learning. *Information Fusion*, page 103186, 2025.

[Jang *et al.*, 2017] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparametrization with gumble-softmax. In *International Conference on Learning Representations (ICLR 2017)*. OpenReview. net, 2017.

[Jordan and Freiburger, 2015] Kareem L Jordan and Tina L Freiburger. 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, 2015.

[Kang *et al.*, 2020] Jian Kang, Jingrui He, Ross Maciejewski, and Hanghang Tong. Inform: Individual fairness on graph mining. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 379–389, 2020.

[Kipf and Welling, 2017] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[Li *et al.*, 2024] Chengyu Li, Debo Cheng, Guixian Zhang, and Shichao Zhang. Contrastive learning for fair graph representations via counterfactual graph augmentation. *Knowledge-Based Systems*, 305:112635, 2024.

[Li *et al.*, 2025] Yi Li, Shichao Zhang, Guixian Zhang, and Debo Cheng. Community-centric graph unlearning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18548–18556, 2025.

[Liu *et al.*, 2022] Shang Liu, Yang Cao, Takao Murakami, and Masatoshi Yoshikawa. A crypto-assisted approach for publishing graph statistics with node local differential privacy. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 5765–5774. IEEE, 2022.

[Liu *et al.*, 2024] Shang Liu, Yang Cao, Takao Murakami, Jinfei Liu, and Masatoshi Yoshikawa. Cargo: Crypto-assisted differentially private triangle counting without trusted servers. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 1671–1684. IEEE, 2024.

[Luo *et al.*, 2024] Renqiang Luo, Huafei Huang, Shuo Yu, Zhuoyang Han, Estrid He, Xiuzhen Zhang, and Feng Xia. Fugnn: Harmonizing fairness and utility in graph neural networks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2072–2081, 2024.

[Luo *et al.*, 2025] Renqiang Luo, Huafei Huang, Ivan Lee, Chengpei Xu, Jianzhong Qi, and Feng Xia. Fairgp: A scalable and fair graph transformer using graph partitioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12319–12327, 2025.

[Lv *et al.*, 2022] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8046–8056, 2022.

[Ma *et al.*, 2022] Jing Ma, Ruocheng Guo, Mengting Wan, Longqi Yang, Aidong Zhang, and Jundong Li. 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, 2022.

[McPherson *et al.*, 2001] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[Meng *et al.*, 2025] Biwen Meng, Xi Long, Wanrong Yang, Ruochen Liu, Yi Tian, Yalin Zheng, and Jingxin Liu. Advancing cross-organ domain generalization with test-time style transfer and diversity enhancement. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2025.

[Parascandolo *et al.*, 2018] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *International Conference on Machine Learning*, pages 4036–4044. PMLR, 2018.

[Pearl, 2009] Judea Pearl. *Causality*. Cambridge university press, 2009.

[Reichenbach, 1991] Hans Reichenbach. *The direction of time*, volume 65. Univ of California Press, 1991.

[Wang *et al.*, 2022] Yu Wang, Yuying Zhao, Yushun Dong, Huiyuan Chen, Jundong Li, and Tyler Derr. 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, 2022.

[Wu *et al.*, 2024] Zongqian Wu, Yujie Mo, Peng Zhou, Shangbo Yuan, and Xiaofeng Zhu. Self-training based few-shot node classification by knowledge distillation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 15988–15995, 2024.

[Xu *et al.*, 2019] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

[Xu *et al.*, 2025] Ziqi Xu, Sevvandi Kandanaarachchi, Cheng Soon Ong, and Eirini Ntoutsi. Fairness evaluation with item response theory. In *Proceedings of the ACM on Web Conference 2025*, pages 2276–2288, 2025.

[Yan *et al.*, 2023] Kailun Yan, Jilian Zhang, and Yongdong Wu. Improving bitcoin transaction propagation efficiency through local clique network. *The Computer Journal*, 66(2):318–332, 2023.

[Yan *et al.*, 2025] Bo Yan, Sihao He, Cheng Yang, Shang Liu, Yang Cao, and Chuan Shi. Federated graph condensation with information bottleneck principles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12990–12998, 2025.

[Yang *et al.*, 2024] Cheng Yang, Jixi Liu, Yunhe Yan, and Chuan Shi. Fairsin: Achieving fairness in graph neural networks through sensitive information neutralization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 9241–9249, 2024.

[Yeh and Lien, 2009] I-Cheng Yeh and Che-hui Lien. 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, 2009.

[Zhang *et al.*,] Haiqi Zhang, Guangquan Lu, Mengmeng Zhan, and Beixian Zhang. Semi-supervised classification of graph convolutional networks with laplacian rank constraints. *Neural Processing Letters*, pages 1–12.

[Zhang *et al.*, 2017] Shichao Zhang, Xuelong Li, Ming Zong, Xiaofeng Zhu, and Debo Cheng. Learning k for knn classification. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(3):1–19, 2017.

[Zhang *et al.*, 2024] Guixian Zhang, Debo Cheng, Guan Yuan, and Shichao Zhang. Learning fair representations via rebalancing graph structure. *Information Processing & Management*, 61(1):103570, 2024.

[Zhang *et al.*, 2025a] Guixian Zhang, Guan Yuan, Debo Cheng, Lin Liu, Jiuyong Li, Ziqi Xu, and Shichao Zhang. Deconfounding representation learning for mitigating latent confounding effects in recommendation. *Knowledge and Information Systems*, pages 1–22, 2025.

[Zhang *et al.*, 2025b] Guixian Zhang, Guan Yuan, Debo Cheng, Lin Liu, Jiuyong Li, and Shichao Zhang. Disentangled contrastive learning for fair graph representations. *Neural Networks*, 181:106781, 2025.