PRELIMINARY PREPRINT VERSION – DO NOT CITE
This is not the version of record. The final version will appear in the Proceedings of IJCAI-ECAI 2026.

Learning Counterfactual Fairness from Authentic Generation

Zichong Wang[1] , **Zhipeng Yin**[1] , **Zhong Chen**[2] ,
Jack Yang[3] , **Jun Liu**[4] and **Wenbin Zhang**[1*]
[1]Florida International University, FL, United States
[2]Southern Illinois University Carbondale, IL, United States
[3]University of Wollongong, NSW, Australia
[4]Carnegie Mellon University, PA, United States

Abstract

Fairness-aware graph learning has become increasingly important amid growing concerns about algorithmic bias in networked data. Among existing approaches, counterfactual fairness is particularly appealing as it seeks to eliminate unfairness at its causal origin by ensuring that predictions remain invariant in counterfactual worlds where sensitive attributes are altered. However, most existing methods assume that all observed variables are directly influenced by sensitive attributes, an overly strong and often unrealistic assumption in real-world graphs. To address this limitation, we propose Graph Counterfactual Fairness (GCFair), a novel framework that achieves counterfactual fairness by explicitly identifying and disentangling the subsets of node features and graph structures genuinely affected by sensitive attributes. This principled joint disentanglement enables the generation of authentic counterfactual instances that selectively modify only sensitive-related information while preserving all sensitive-irrelevant factors. Extensive experiments show that GCFair effectively mitigates bias and outperforms state-of-the-art fairness methods in both counterfactual fairness and predictive accuracy.

1 Introduction

Graph Neural Networks (GNNs) have demonstrated effectiveness across various graph tasks [Kipf and Welling, 2016; Zhao *et al.*, 2022; Wu *et al.*, 2020], leading to their successful application in high-risk decision-making systems such as social media analysis [Leonelli *et al.*, 2021], and criminal justice [Berk *et al.*, 2021]. For example, GNNs can assist financial institutions in critical functions like evaluating credit card applications or making loan approval decisions [Wang *et al.*, 2021]. Despite their success, there is a growing concern that GNNs may inherit or even amplify discrimination and social bias present in the training data, leading to unfair treatment of subgroups defined by sensitive attributes such as gender, age, region, and race [Liang *et al.*, 2022]. For instance, a hiring system powered by GNNs unfairly disadvantaged applicants from certain ethnic groups, raising serious ethical risks about the deployment of these models in real-world decision-making scenarios [Chen *et al.*, 2024a; Wang *et al.*, 2025g].

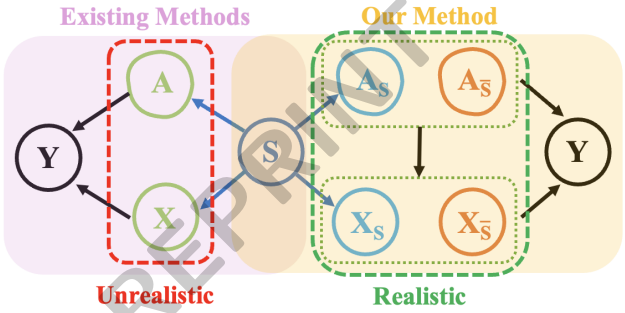

Figure 1: Comparison of existing works for modeling the causal relations among sensitive attribute and observed variables.

To this end, many efforts have recently been devoted to ensuring fairness in graph-based prediction models [Wang *et al.*, 2026b; Wang *et al.*, 2026a; Wang *et al.*, 2026d; Wang *et al.*, 2025b]. Among these efforts, counterfactual fairness stands out by explicitly leveraging causal mechanisms, enabling the modeling and elimination of discrimination at the individual level using Pearl's structural causal models [Pearl, 2009]. The core idea behind the counterfactual fairness is that predictions for an individual should remain consistent across hypothetical scenarios where their sensitive attributes are altered [Kusner *et al.*, 2017]. For example, consider a scenario in which a loan application from an applicant belonging to a minority group is rejected. To evaluate counterfactual fairness, one would need to generate a realistic "what-if" scenario (a counterfactual instance) by altering only the sensitive attribute (such as ethnicity) while keeping all other relevant factors fixed, and verify whether the decision remains unchanged [Zhang and Weiss, 2022]. To operationalize this concept in graph data, most existing methods generate these counterfactual instances either by directly flipping sensitive attributes or by employing graph-generation models [Wang *et al.*, 2023a]. For instance, NIFTY [Agarwal *et al.*, 2021] creates counterfactual samples by perturbing sensitive attributes and maximizing the similarity between

*Corresponding author.

IJCAI-ECAI 2026 PREPRINT

PRELIMINARY PREPRINT VERSION – DO NOT CITE
This is not the version of record. The final version will appear in the Proceedings of IJCAI-ECAI 2026.

original and perturbed representations to achieve invariance concerning sensitive attributes.

Despite their valuable contributions, these existing approaches often rely on an unrealistic assumption, that all observed factors are directly influenced by sensitive attributes, potentially leading to unrealistic counterfactual instances in practical scenarios. As illustrated in Figure 1, existing methods typically assume a simplistic scenario where the sensitive attribute (S) directly affects all node features (X) and graph structures (A). For example, consider gender as a sensitive attribute in a hiring scenario: it can directly influence sensitive-related node features (X_S) such as physical characteristics) and sensitive-related graph structure (A_S) such as gender-based social connections, but it does not causally affect sensitive-irrelevant node features ($X_{\overline{S}}$) and sensitive-irrelevant graph structure ($A_{\overline{S}}$), like professional skills and classmate relationship. Consequently, altering gender without accurately identifying and isolating sensitive-related information could unintentionally modify irrelevant features, leading to unrealistic counterfactual instances.

Despite the importance of learning counterfactual fairness through authentic generation, this area remains significantly underexplored and presents substantial challenges: **i) Difficulty in accurately identifying which information should be modified when generating counterfactuals.** In real-world graph data, sensitive attributes typically influence only a subset of node features and structural patterns, while many other factors are unrelated or only spuriously correlated. Without explicitly distinguishing sensitive-related information from sensitive-irrelevant information, counterfactual generation may unintentionally alter factors that should remain unchanged, resulting in implausible or invalid counterfactual instances. Accurately determining what to change and what to preserve is, therefore, a core challenge for authentic counterfactual generation. **ii) Difficulty in handling complex interactions between node features and graph structures during counterfactual generation.** Even with accurate identification of sensitive-related and sensitive-irrelevant information, generating realistic counterfactual instances remains challenging due to the complex interdependencies between node attributes and graph structure. Node features and graph connectivity mutually influence each other, making it difficult to alter one aspect without inadvertently disrupting coherence in the other. This interconnected nature requires an integrated approach to maintain consistency and realism in counterfactual instances. **iii) Ensuring coherent and realistic counterfactual generation through integrated graph structural and feature-level disentanglement.** Effective counterfactual fairness requires simultaneous interventions on sensitive-related information of both node features and graph connectivity, while preserving sensitive-irrelevant aspects. Achieving this balance is inherently challenging, as it demands coordinated inference and precise alignment between structural and feature-level disentanglement mechanisms to ensure causal consistency and realism in generated counterfactual instances.

To address the above challenges, we introduce a novel causal framework for graph counterfactual fairness, named GCFair, which explicitly addresses the critical problem of accurately generating authentic counterfactual instances by jointly disentangling sensitive-related from sensitive-irrelevant information in node features and graph structures. *To the best of our knowledge, this is the first work that simultaneously achieves structural and feature-level disentanglement within graph neural networks for authentic counterfactual fairness.* Specifically, GCFair utilizes a causal-effect variational autoencoder framework to explicitly encode and separate sensitive-related from sensitive-irrelevant latent factors in node features, guided by dedicated inference networks and an adversarial correlation regularizer. In parallel, we introduce an information bottleneck objective designed to isolate sensitive-related graph structural patterns within the observed graph, effectively disentangling them from sensitive-irrelevant structural information. Through this coordinated approach, GCFair ensures that counterfactual interventions selectively modify only the latent factors and structural components directly influenced by sensitive attributes, preserving causal consistency and realism in generated counterfactual instances. The major contributions of this paper are summarized as follows: **i)** We introduce a principled disentanglement strategy that accurately identifies and isolates sensitive-related information within both node features and graph structures. **ii)** We propose GCFair, a novel framework for graph counterfactual fairness, which builds on our proposed disentanglement strategy to generate authentic counterfactual graphs by selectively modifying only sensitive-related information. **iii)** We validate the effectiveness of our model on three real-world datasets, and results indicate our model outperforms a variety of state-of-the-art baselines.

2 Related Work

Counterfactual Fairness in Graph Learning. Fairness in graph learning has attracted increasing attention, with numerous studies addressing this challenge in recent years [Wang *et al.*, 2025f; Wang *et al.*, 2022; Wang *et al.*, 2025c]. Among them, counterfactual fairness has attracted a surge of attentions because of its strong capability of modeling how discrimination arises and is exhibited [Zhang *et al.*, 2025]. Most existing counterfactual fairness studies [Ma *et al.*, 2022; Chen *et al.*, 2024b] focus on generating counterfactual instances either by directly modifying sensitive attributes or by employing graph-generation methods. However, these methods commonly assume that all observed factors are directly influenced by sensitive attributes, an unrealistic assumption that can lead to the generation of implausible counterfactual instances. In addition, a few studies [Wang *et al.*, 2023a; Wang *et al.*, 2025a] aim to identify potential counterfactual instances within the input dataset itself, rather than generating new ones. However, unlike theoretically-defined counterfactual instances, identifying appropriate real-world counterfactual examples from observational graph data lacks a clear consensus. Hence, these methods cannot guarantee that suitable counterfactuals will be found or that identified instances truly satisfy the intended fairness properties.

Graph Learning with Disentanglement. Learning disentangled representations, where latent factors underlying observed data are explicitly separated, has gained superior per-

IJCAI-ECAI 2026 PREPRINT

formance in recent graph learning research [Yang *et al.*, 2024]. Existing works have demonstrated the benefits of disentanglement by enhancing interpretability and performance across diverse graph-based tasks. For example, DisenGCN [Ma *et al.*, 2019] pioneered graph disentanglement through a neighborhood routing mechanism, explicitly capturing distinct latent factors influencing node interactions. Further advancing this line of inquiry, IPGDN [Liu *et al.*, 2020] incorporated independence regularization between latent factors, enhancing the clarity and independence of the learned representations. Recently, some studies [Wang and Zhang, 2025] have explored leveraging disentangled learning specifically to enforce fairness constraints more accurately, aiming primarily to minimize performance degradation commonly associated with fairness optimization. However, these methods focus predominantly on sustaining predictive performance rather than directly improving fairness outcomes. To explicitly enhance model fairness in graph learning contexts, there is a need to design fundamentally different disentanglement approaches that account for the inherent relational structures and neighborhood dependencies unique to graph data.

Different from the above works, our work explicitly designs a disentanglement-based framework to achieve graph counterfactual fairness. By jointly disentangling sensitive-related information from both node features and graph structures, our method precisely intervenes on sensitive attributes, selectively modifying sensitive-related components while preserving sensitive-irrelevant information. By doing so, our approach generates authentic counterfactual graphs that directly enhance model fairness, rather than merely preserving predictive performance.

3 Notations

We consider a node classification task on a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$, where $\mathcal{V}$ denotes the node set with $|\mathcal{V}| = n$ nodes and $\mathcal{E}$ denotes the edge set with $|\mathcal{E}| = m$ edges. The node feature matrix is $\mathbf{X} \in \mathbb{R}^{n \times d}$, where the i^{th} row $\mathbf{x}_i$ represents the d-dimensional feature vector of node $v_i \in \mathcal{V}$. The graph structure is represented by an adjacency matrix $\mathbf{A} \in \{0,1\}^{n \times n}$, where $\mathbf{A}_{i,j} = 1$ indicates the existence of an edge $e_{i,j} \in \mathcal{E}$ between nodes v_i and v_j, and $\mathbf{A}_{i,j} = 0$ otherwise. For simplicity, we assume that both node labels and the sensitive attribute are binary. Let $\mathbf{S} \in \{0,1\}^{n \times 1}$ denote the sensitive attribute vector, where s_i is the sensitive attribute value associated with node v_i; we define $S_d = \{v_i \in \mathcal{V} \mid s_i = 0\}$ as the deprived group and $S_f = \{v_i \in \mathcal{V} \mid s_i = 1\}$ as the favored group. Each node v_i is associated with a ground-truth label $y_i \in \{0,1\}$, where $y_i = 1$ denotes a granted label and $y_i = 0$ denotes a rejected label, and $\hat{y}_i$ denotes the predicted label of node v_i.

4 Methodology

4.1 Causal Model

To explicitly identify sensitive-related information and distinguish it from unrelated components, the proposed causal model treats sensitive attributes as direct causes that influence both node features and graph structure. This integrated design enables targeted interventions on sensitive-related features and structural information while preserving components that are independent of sensitive attributes. By isolating the causal effects of sensitive attributes, the framework facilitates the generation of realistic counterfactual samples, thereby supporting fair decision-making in graph-based tasks. In the causal graph, each directed edge represents a direct causal influence, explicitly capturing how sensitive attributes affect features and connectivity. Below, we detail the rationale behind each causal relationship.

- $A_S \leftarrow S \rightarrow X_S$: The sensitive attribute S directly influences both a subset of node features (X_S) and the graph structure (A_S). For instance, if gender is the sensitive attribute, sensitive-related features could include height, while A_S capture network connections preferentially formed among individuals of the same gender.
- $X_{\overline{S}} \perp S \perp A_{\overline{S}}$: Node features unrelated to the sensitive attribute ($X_{\overline{S}}$) and the corresponding sensitive-irrelevant graph structure ($A_{\overline{S}}$) are explicitly independent of the sensitive attribute S.
- $X_{\overline{S}} \leftarrow A_{\overline{S}}$; $A_S \rightarrow X_S$: The graph structure directly influences both sensitive-related and sensitive-irrelevant node features, capturing how individuals' connections can shape attributes both independently and dependently of direct sensitive-related graph structure causation.
- $X \rightarrow Y \leftarrow A$: Both node feature and graph structure, contain essential information influencing the outcome label, but the sensitive attribute S should not exert a direct causal effect on Y.

4.2 Latent Disentanglement of Sensitive Attribute Dependencies

Building on our proposed causal model, we introduce a latent disentanglement framework designed to explicitly separate the causal dependencies arising from sensitive attributes in both node features and graph structure. Specifically, our approach involves two complementary parts: i) disentangling node features into sensitive-related and sensitive-irrelevant features; ii) disentangling the graph structural dependencies within the graph influenced by sensitive attributes. For the first part, we leverage a causal-effect variational autoencoder to encode observed features into latent factors that explicitly separate variations causally driven by sensitive attributes from those unrelated to them, enabling targeted feature-level interventions. On the other hand, the second part focuses on graph structure, where sensitive-related relational mechanisms are inferred from the observed graph and isolated from task-relevant structural patterns using an information bottleneck objective, allowing counterfactual graphs to be generated by intervening only on sensitive-related structural components while preserving the remaining structure.

Node Features Disentanglement. We model node features through a causal-effect variational autoencoder that explicitly separates sensitive-related and sensitive-irrelevant factors in the latent space. Specifically, observed node features are encoded into two distinct sets of latent variables: sensitive-related latent variables (Z_S), which capture feature

IJCAI-ECAI 2026 PREPRINT

variations directly influenced by the sensitive attribute (S), and sensitive-irrelevant latent variables ($Z_{\overline{S}}$), which represent feature information unaffected by S. Inference over these latent factors is performed using variational Bayesian inference with two dedicated inference networks, $q_\phi(Z_S \mid X_S, A_S, S, Y)$ and $q_\theta(Z_{\overline{S}} \mid X_{\overline{S}}, A_{\overline{S}}, Y)$, each aligned with the causal separation between sensitive-related and sensitive-irrelevant features. This ensures that the learned representations are consistent with the assumed causal structure, enabling reliable counterfactual interventions on node features.

However, relying on architectural design is insufficient for true independence and disentanglement between latent representations, due to the inherent complexity of node features, which exhibit indirect interactions and intricate relational patterns [Xiao *et al.*, 2022]. In addition, conventional metrics relying solely on linear correlations or observable attributes often fail to capture the subtle, non-linear dependencies within latent spaces, making traditional independence measures insufficient for reliably assessing disentanglement quality [Eastwood and Williams, 2018]. To this end, we propose the Hirschfeld-Gebelein-Rényi (HGR) based correlation measure, defined formally in Definition 4.2.1, as a principled way to precisely quantify the independence between sensitive-related and sensitive-irrelevant latent variables. With the Z_S-$Z_{\overline{S}}$ correlation measure established as the foundational metric for evaluating independence, we then integrate it into our optimization framework by introducing a structured min-max adversarial training strategy to rigorously ensure disentanglement between sensitive-related and sensitive-irrelevant latent representations. Starting with the correlation measure, the formal definition is given below,

Definition 4.2.1 Z_S**-**$Z_{\overline{S}}$ **correlation.** Given any two jointly distributed random latent variables Z_S and $Z_{\overline{S}}$, the maximal HGR correlation between them is defined as:

$$\mathrm{HGR}(Z_S, Z_{\overline{S}}) = \sup_{\psi_1, \psi_2} \rho\big(f_1(Z_S), f_2(Z_{\overline{S}})\big) \tag{1}$$

where ρ denotes the Pearson correlation coefficient, while ψ_1 and ψ_2 are measurable functions with finite positive variances. To eliminate arbitrary shifts and scaling differences between ψ_1 and ψ_2, we normalize these functions such that $\mathbb{E}[f_1(Z_S)] = \mathbb{E}[f_2(Z_{\overline{S}})] = 0$ and $\mathbb{E}[f_1^2(Z_S)] = \mathbb{E}[f_2^2(Z_{\overline{S}})] = 1$. This normalization is essential, as it ensures comparability and stability across correlation measurements by standardizing the scale, thereby enabling consistent and meaningful quantification of the dependence between latent variables. Under this normalization, a Z_S-$Z_{\overline{S}}$ correlation of zero indicates perfect independence, whereas a correlation value of one indicates a deterministic relationship.

With the established Z_S-$Z_{\overline{S}}$ correlation, we structure a min-max adversarial optimization framework to ensure effective and precise disentanglement. Specifically, this framework consists of two adversarial subnets parameterized by η_{ψ_1} and η_{ψ_2}. These subnetworks actively estimate and attempt to strengthen the dependency between the latent variables Z_S and $Z_{\overline{S}}$. During training, the subnetworks aim to maximize the Z_S-$Z_{\overline{S}}$ correlation through gradient ascent, thereby identifying potential residual dependencies between sensitive-related and sensitive-irrelevant representations. Simultaneously, the primary inference networks, responsible for encoding the latent representations Z_S and $Z_{\overline{S}}$, are trained to minimize this estimated dependency via gradient descent. This min-max procedure ensures maximal independence between the latent factors, resulting in rigorously disentangled latent representations. By training the model adversarially, we enforce the maximum independence between Z_S and $Z_{\overline{S}}$, achieving precise disentanglement. Overall, this approach establishes a foundation for authentic and precise counterfactual instance generation at the node-feature level by ensuring strong independence between sensitive-related (Z_S) and sensitive-irrelevant ($Z_{\overline{S}}$) latent representations. Such explicit disentanglement allows subsequent counterfactual interventions to selectively modify only causally relevant latent factors, while reliably preserving unrelated feature information.

Graph Structure Disentanglement. Beyond node-feature disentanglement, we explicitly address sensitive attribute dependencies embedded within the graph structure information. In other words, even if sensitive-related and sensitive-irrelevant node features are successfully separated, sensitive attributes may still influence the graph structure through connectivity biases such as homophily. These biases manifest as multi-hop relational patterns rather than direct node-level features, making node feature-based disentanglement alone insufficient for fully capturing relational influences. To this end, we propose a graph structural disentanglement strategy leveraging an information bottleneck objective to identify and disentanglement the structural component driven by the S from sensitive attribute-related and sensitive-irrelevant structural patterns. By explicitly modeling and extracting these sensitive-related relational dependencies, our framework ensures that counterfactual graph generation can selectively intervene only on sensitive-related structural patterns, while keeping all remaining structural information unchanged. Specifically, we introduce a latent decomposition of the graph G into two complementary structural views: the sensitive-related subgraph $G_S = (\mathcal{V}_S, \mathbf{A}_S, \mathbf{X}_S)$ and the sensitive-irrelevant structure $G_{\overline{S}} = (\mathcal{V}_{\overline{S}}, \mathbf{A}_{\overline{S}}, \mathbf{X}_{\overline{S}})$, where edges satisfy $\mathbf{A}_S + \mathbf{A}_{\overline{S}} = \mathbf{A}$ and $\mathbf{A}_S \odot \mathbf{A}_{\overline{S}} = \emptyset$. In other words, each edge is assigned to exactly one structural component. Our goal is to learn G_S such that it captures the relational patterns induced by S (the part we should modify under counterfactual intervention), while $G_{\overline{S}}$ contains the remaining structure and should be as insensitive to S as possible.

To learn this disentanglement, we parameterize the extraction of G_S by a stochastic structural extractor $P_\psi(G_S \mid G)$, where ψ are trainable parameters. Hence, P_ψ outputs candidate sensitive-related structural views from the input graph, using a stochastic extractor avoids brittle hard decisions early in training and allows end-to-end optimization. To achieve this, the structural information bottleneck is used to guide what kind of structure should be captured by G_S: the extracted structure is encouraged to be concise rather than a direct copy of the original graph, while still retaining the structural patterns that are relevant for predicting the sensitive attribute S. This trade-off is enforced by an information bottleneck objective that penalizes redundant structural information from G while encourage sensitivity-related structural in G_S,

IJCAI-ECAI 2026 PREPRINT

as formalized in Definition 4.2.2.

Definition 4.2.2 Sensitive-related structural information bottleneck. Given a jointly distributed graph random variable G and sensitive attribute S, and a stochastic structural view $G_S \sim P_\psi(G_S \mid G)$, we define the sensitive-related graph structural bottleneck objective as:

$$\mathcal{L}_{\mathrm{IB}}(\psi) \;=\; I(G, G_S) - \alpha\, I(S, G_S) \tag{2}$$

where $I(\cdot,\cdot)$ denotes mutual information and $\alpha > 0$ controls the trade-off. In addition, the compression term $I(G, G_S)$ penalizes copying the entire structure into G_S, while the sensitive-information term $I(S, G_S)$ encourages G_S to retain the structural patterns that are predictive of S.

Building on this, the optimal sensitive-related graph structure is obtained via:

$$G_S(\psi^*) = \arg\min_{\psi}\; I(G, G_S) - \alpha I(S, G_S) \tag{3}$$

where $G_S(\psi^*)$ is the sensitive-related graph structural component produced by the optimal extractor. As a result, this optimization leads to a structural disentanglement in which counterfactual graph generation interventions modify only sensitive-related relational structures, producing realistic and causally consistent structural adjustments.

4.3 Counterfactual Generation via Disentangled Learning

Building upon the disentangled latent representations that clearly separate sensitive-related from sensitive-irrelevant information in both node features and graph structure, GCFair generates counterfactual graph instances via a structured three-stage causal inference procedure. The core idea is that counterfactual interventions on S should exclusively modify components directly influenced by the sensitive attribute, specifically sensitive-related node features and graph structure, while preserving all sensitive-irrelevant information to maintain coherence and causal consistency. To realize this principle, GCFair adopts the canonical three-stage counterfactual inference pipeline, explicitly augmented with disentangled representations for node features and structural information. Specifically, these stages include: i) **Abduction**, where the model infers latent variables from the observed graph data. Different from existing work, GCFair infers disentangled latent variables $(Z_S, Z_{\overline{S}})$ for node features, and additionally extracts a sensitive-related structural component from the observed adjacency via the learned structural disentangler. Concretely, given an observed graph $G = (\mathcal{V}, \mathbf{A}, \mathbf{X})$, we obtain a sensitive-related structural view $G_S = (\mathcal{V}, \mathbf{A}_S, \mathbf{X})$ through the trained extractor $P_\psi(G_S \mid G)$, and define the residual structure as $\mathbf{A}_{\overline{S}} = \mathbf{A} - \mathbf{A}_S$; ii) **Action**, where the sensitive attribute is intervened upon by assigning it a counterfactual value $(s \to s')$. In this stage, the inferred latent factors are held fixed, and only the sensitive-related components, both structural and feature-based, are re-sampled under the new sensitive attribute value; and iii) **Generation**, where the decoder reconstructs a realistic counterfactual instance $(\mathbf{A}^{cf}, \mathbf{X}^{cf}, Y^{cf})$ by integrating the re-sampled sensitive-related components with the preserved sensitive-irrelevant residual information. This structured inference procedure is underpinned by a foundational causal assumption, explicitly defining the permissible causal pathways within the proposed framework.

Assumption 4.3 The sensitive attribute S does not directly affect the ground-truth label Y; its influence on Y is mediated only through sensitive-related node features X_S and the sensitive-related structural component of the graph. In addition, S has no causal influence on sensitive-irrelevant node features $X_{\overline{S}}$ and no causal influence on the sensitive-irrelevant residual structure $A_{\overline{S}}$.

Formally, under this assumption, our model characterizes the full counterfactual generation process through the following joint distribution factorization:

$$\begin{aligned} P(S, \mathbf{A}_S, \mathbf{A}_{\overline{S}}, \mathbf{X}_S, \mathbf{X}_{\overline{S}}, Y) = \;& P(S)P(Z_{\overline{S}})\, P(\mathbf{A}_S \mid S, Z_S) \\ & P(Z_S)P(\mathbf{A}_{\overline{S}} \mid Z_{\overline{S}})P(\mathbf{X}_S \mid \mathbf{A}_S, S, Z_S) \\ & P(\mathbf{X}_{\overline{S}} \mid \mathbf{A}_{\overline{S}}, Z_{\overline{S}})\, P\big(Y \mid Z_S, Z_{\overline{S}}\big) \end{aligned} \tag{4}$$

where $P(Z_S)$ and $P(Z_{\overline{S}})$ represent prior distributions for sensitive-related and sensitive-irrelevant latent variables, respectively; $P(S \mid Z_S)$ ensures that latent variables sufficiently encode sensitive attribute information. Structural mechanisms $P(\mathbf{A}_S \mid S, Z_S)$ and $P(\mathbf{A}_{\overline{S}} \mid Z_{\overline{S}})$ explicitly model graph structure information, ensuring that structural changes under intervention (*i.e.*, $do(S)$) affect only the sensitive-related graph structural information, while preserving the residual structure. Similarly, the feature mechanisms $P(\mathbf{X}_S \mid \mathbf{A}_S, S, Z_S)$ and $P(\mathbf{X}_{\overline{S}} \mid \mathbf{A}_{\overline{S}}, Z_{\overline{S}})$ control node feature modifications, guaranteeing that counterfactual interventions selectively alter only sensitive-related features and leave sensitive-irrelevant features intact.

However, maximizing the joint likelihood of the observed graph variables is computationally difficult, mainly because it requires integrating over high-dimensional latent factors and also over the latent structural decomposition that separates sensitive-related from residual connectivity. To this end, we adopt variational Bayesian inference and approximate the true posterior $P(Z_S, Z_{\overline{S}} \mid S, G, Y)$ by a factorized variational family $q_\phi(Z_S \mid \mathbf{A}_S, \mathbf{X}_S, S, Y)$, $q_\theta(Z_{\overline{S}} \mid \mathbf{A}_{\overline{S}}, \mathbf{X}_{\overline{S}}, Y)$, and a stochastic structural extractor $q_\psi(G_S \mid G)$, as detailed:

$$\begin{aligned} q_{\phi,\theta,\psi}(Z_S, Z_{\overline{S}} \mid G, S, Y) = \;& q_\phi(Z_S \mid \mathbf{A}_S, \mathbf{X}_S, S, Y) \\ & q_\theta(Z_{\overline{S}} \mid \mathbf{A}_{\overline{S}}, \mathbf{X}_{\overline{S}}, Y)\; q_\psi(G_S \mid G) \end{aligned} \tag{5}$$

Under this approximation, we maximize an evidence lower bound (ELBO) associated with our feature-structure generative factorization:

$$\begin{aligned} \mathcal{L}_{\mathrm{ELBO}} = \mathbb{E}_q \Big[& \log P(S \mid Z_S) + \log P(\mathbf{A}_S \mid S, Z_S) \\ & + \log P(\mathbf{A}_{\overline{S}} \mid Z_{\overline{S}}) + \log P(\mathbf{X}_S \mid \mathbf{A}_S, S, Z_S) \\ & + \log P(\mathbf{X}_{\overline{S}} \mid \mathbf{A}_{\overline{S}}, Z_{\overline{S}}) + \log P\big(Y \mid Z_S, Z_{\overline{S}}\big) \Big] \\ & - \mathrm{KL}\big[q_\phi(Z_S \mid \cdot) \,\|\, P(Z_S)\big] - \mathrm{KL}\big[q_\theta(Z_{\overline{S}} \mid \cdot) \,\|\, P(Z_{\overline{S}})\big] \end{aligned} \tag{6}$$

IJCAI-ECAI 2026 PREPRINT

Method	Credit				Bail				Pokec-z			
	AUC	F1-Score	SPD	EOD	AUC	F1-Score	SPD	EOD	AUC	F1-Score	SPD	EOD
GCN	68.57 ± 3.51	80.46 ± 1.16	12.51 ± 3.08	11.83 ± 3.23	88.12 ± 1.21	76.71 ± 1.71	7.61 ± 1.30	3.74 ± 1.17	**75.21 ± 1.74**	68.80 ± 1.47	4.83 ± 1.61	4.97 ± 0.73
NIFTY	72.57 ± 1.61	80.35 ± 3.03	9.83 ± 1.01	8.21±1.23	76.79 ± 1.70	72.18 ± 2.21	5.85 ± 1.23	4.14±1.38	73.41 ± 1.01	67.51 ± 3.61	3.13 ± 1.08	3.91±1.78
GEAR	**74.80** ± 0.96	82.28 ± 2.77	9.37 ± 1.41	7.82±0.98	88.11 ± 3.62	80.68 ± 3.32	5.04 ± 1.38	3.24±1.03	70.87 ± 2.76	65.01 ± 2.11	3.48 ± 0.43	4.13±0.91
CAF	71.83 ± 3.59	82.59 ± 2.28	7.62 ± 2.46	6.64±1.83	**90.64 ± 1.92**	**81.01 ± 1.86**	2.80 ± 1.44	1.61±0.63	OOM	OOM	OOM	OOM
AGCG	73.02 ± 1.33	**85.25 ± 6.15**	6.35 ± 1.09	**4.31±1.68**	86.63 ± 2.99	76.95 ± 3.26	2.35 ± 0.76	1.85±1.27	73.31 ± 2.33	**71.53 ± 2.47**	4.75 ± 2.41	4.48±1.91
Fairwos	73.47 ± 1.33	75.16 ± 3.14	9.28 ± 2.75	7.43 ± 3.21	78.09 ± 2.19	75.16 ± 2.48	3.87 ± 1.14	2.12 ± 1.01	70.73 ± 3.31	64.53 ± 1.47	5.17 ± 1.57	4.92±1.98
GCFair	73.69 ± 3.40	83.22 ± 3.14	**6.09 ± 1.75**	5.12 ± 0.61	89.87 ± 1.61	79.71 ± 3.56	**2.31 ± 0.54**	**1.54±0.87**	73.53 ± 2.91	70.74 ± 1.49	**3.01 ± 1.64**	**3.73±2.63**

Table 1: The comparison of GCFair with the baselines. OOM indicates Out of memory.

where $\mathrm{KL}(\cdot)$ denotes the Kullback-Leibler divergence of the posterior distributions $q_\phi(Z_S)$ and $q_\theta(Z_{\overline{S}})$ from their respective priors $P(Z_S)$ and $P(Z_{\overline{S}})$, which regularizes the latent space for reliable counterfactual generation.

To further ensure that our learned latent representations accurately reflect the intended causal disentanglement between sensitive-related and sensitive-irrelevant features, we explicitly integrate the previously established Z_S-$Z_{\overline{S}}$ independence constraint into the optimization objective. Specifically, while the ELBO encourages effective latent representation learning, it does not inherently guarantee sufficient independence between Z_S and $Z_{\overline{S}}$. By combining the ELBO with the explicit independence constraint (formalized by the Z_S-$Z_{\overline{S}}$ correlation measure), we rigorously enforce the targeted disentanglement established in the previous subsection. This integrated optimization framework simultaneously ensures high-quality latent representation learning and precise feature disentanglement, significantly enhancing the authenticity and precision of generated counterfactual instances. Formally, this joint optimization is defined as follows:

$$\mathcal{L} = -\mathcal{L}_{\text{ELBO}} + \lambda_{\text{HGR}}\,\mathrm{HGR}(Z_S, Z_{\overline{S}}) \tag{7}$$

In addition, the ELBO alone does not guarantee that the extracted structural component G_S concentrates on sensitive-related relational patterns, nor that the residual structure $G_{\overline{S}}$ is minimally predictive of S. This is because there exist many decompositions of the observed adjacency that can achieve similar reconstruction likelihood, including degenerate cases where G_S collapses to the full graph or becomes empty. Directly optimizing mutual information terms such as $I(G, G_S)$ and $I(S, G_S)$ is intractable for graph-structured variables, so we incorporate a variational information bottleneck regularizer based on standard bounds. Specifically, we impose two complementary constraints: i) we enforce compression of the extracted structural view by penalizing its divergence from a simple reference distribution ($R(G_S)$), initially defined as an edge-independent Bernoulli prior with a low edge-inclusion probability; and ii) we ensure that (G_S) maintains predictive power regarding the sensitive attribute, while simultaneously reducing sensitive attribute leakage from the residual structure ($G_{\overline{S}}$), by introducing variational discriminators:

$$\begin{aligned}\mathcal{L}_{\text{IB}} = &-\beta\,\mathbb{E}_G\Big[\mathrm{KL}\big(q_\psi(G_S \mid G)\,\|\,R(G_S)\big)\Big] \\ &+ \alpha\,\max_{\varphi_S}\,\mathbb{E}_{S,G_S}\big[\log q_{\varphi_S}(S \mid G_S)\big] \\ &- \lambda\,\max_{\varphi_{\overline{S}}}\,\mathbb{E}_{S,G_{\overline{S}}}\big[\log q_{\varphi_{\overline{S}}}(S \mid G_{\overline{S}})\big]\end{aligned} \tag{8}$$

where β, α, λ control the strengths of structural compression, sensitive-related information preservation, and reduction of residual sensitive-related information leakage, respectively.

With the generated counterfactual instances, we introduce a counterfactual fairness constraint explicitly designed to penalize prediction changes resulting from interventions on the sensitive attribute S, while maintaining the inferred latent factors constant. Formally, this constraint is defined as:

$$\begin{aligned}\mathcal{L}_F = &\mathbb{E}_{q_{\phi,\theta,\psi}(Z_S,Z_{\overline{S}}\mid S,X_S,X_{\overline{S}},A_S,A_{\overline{S}})} \\ &\Big[\big\|\hat{p}(Y \mid S, \mathbf{Z}_S, \mathbf{Z}_{\overline{S}}) - \hat{p}^{cf}(Y \mid S', \mathbf{Z}_S, \mathbf{Z}_{\overline{S}})\big\|_2\Big]\end{aligned} \tag{9}$$

Building on this, our final training objective augments the ELBO with i) a feature-level independence constraint, ii) a counterfactual fairness regularizer, and iii) a structural information bottleneck regularizer:

$$\mathcal{L}_{\text{Total}} = -\mathcal{L}_{\text{ELBO}} + \lambda_D\,\mathrm{HGR} + \lambda_F\,\mathcal{L}_F - \lambda_{\text{IB}}\,\mathcal{L}_{\text{IB}} \tag{10}$$

where λ_D, λ_F, and λ_{IB} control the strengths of the feature disentanglement, counterfactual fairness, and structural disentanglement regularizers, respectively.

5 Experiments

5.1 Experiment Settings

Datasets and Baselines. Our experiments are conducted on three datasets: the Credit dataset [Yeh and Lien, 2009], Bail dataset [Agarwal *et al.*, 2021], and Pokec-z dataset [Takac and Zabovsky, 2012]. In addition, we compare GCFair with the following baseline methods: i) Vanilla Graph Model: GCN [Kipf and Welling, 2016], ii) Graph Counterfactual Fairness Methods: NIFTY [Agarwal *et al.*, 2021], GEAR [Ma *et al.*, 2022], CAF [Guo *et al.*, 2023], AGCG [Wang *et al.*, 2025d] and Fairwos [Wang *et al.*, 2025a]. The details are provided in the appendix.

Evaluation Metrics. We use AUC and F1-score to evaluate the model performance. To evaluate fairness, we use two commonly used fairness metrics, *i.e.*, Statistical Parity Differences (SPD) [Dwork *et al.*, 2012] and Equal Opportunity Differences (EOD) [Hardt *et al.*, 2016], with values close to zero indicating better fairness.

5.2 Experiment Results

Performance Comparison. Table 1 presents the results comparing GCFair with six baselines on three datasets. Overall, GCFair achieves the best fairness results on all datasets,

IJCAI-ECAI 2026 PREPRINT

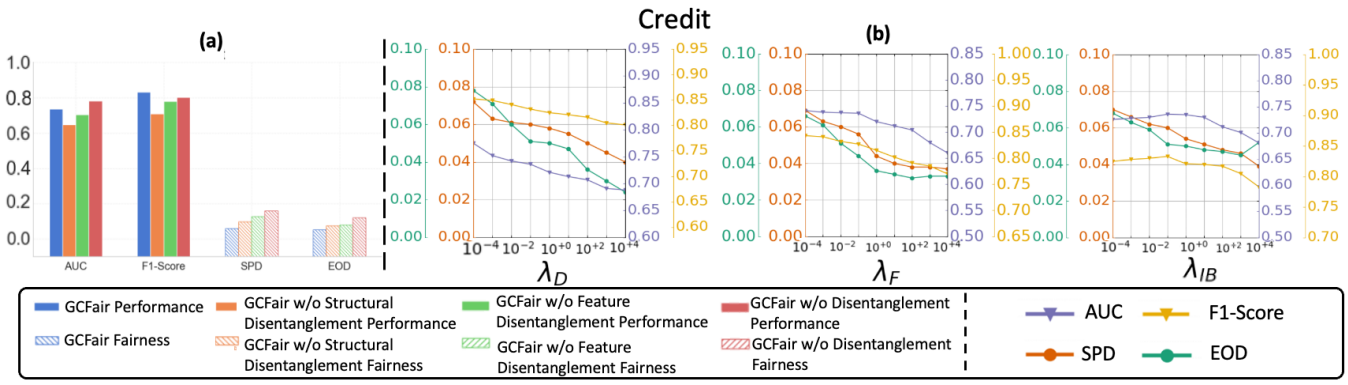

Figure 2: (a) Ablation study results on the Credit dataset. (b) Parameter sensitivity analysis of hyperparameters on the Credit dataset.

while keeping competitive predictive performance. Specifically, i) Fairness improvement: GCFair consistently yields the lowest SPD and EOD across the benchmarks primarily due to its capability to generate authentic counterfactual graphs. By explicitly disentangling sensitive-related information from sensitive-irrelevant components in both node features and graph structures, our method accurately identifies and modifies sensitive-related information, ensuring precise adjustments aligned with counterfactual interventions. Hence, by generating authentic counterfactual instances, we can more effectively enforce fairness constraints, significantly enhancing fairness outcomes. ii) Predictive performance: GCFair maintains competitive AUC and F1-scores because it preserves sensitive-irrelevant information. The selective modification of only causally-linked sensitive components avoids unnecessary removal of task-related information, thus mitigating the accuracy degradation associated with enforcing fairness constraints. Overall, these results demonstrate that GCFair effectively achieve fairness improvements with comparable performance, highlighting its practical utility in graph-based applications.

Ablation Study. To evaluate the effectiveness of each constraint in our method, we provide ablation study with the following variants: i) **w/o Feature Disentanglement:** we remove the feature-level disentanglement constraint by dropping the Z_S-$Z_{\overline{S}}$ correlation regularization which allows sensitive and sensitive-irrelevant information to mix in the latent space; ii) **w/o Structural Disentanglement:** we remove the structural information bottleneck regularizer and learn the model without explicitly isolating sensitive-related structural patterns, so sensitive-related information can remain in the graph structure; and iii) **w/o Disentanglement:** we remove both the feature-level disentanglement constraint and the structural bottleneck regularizer, keeping the remaining training objective unchanged. The results on the Credit dataset is illustrated in Figure 2 (a), and additional results are provided in the appendix. Overall, we observe consistent fairness degradation once any disentanglement constraint is removed. Specifically, removing feature disentanglement allows sensitive information to leak into the inferred latent space, making counterfactual interventions on S imprecise. Similarly, removing structural disentanglement causes residual sensitive information to remain within graph structure, limiting the authenticity of generated counterfactual graphs. Therefore, the counterfactual instances become less reliable for fairness training, negatively impacting both fairness metrics and predictive performance. Finally, the removal of both disentanglement components leads to the largest fairness drop, clearly indicating that both feature-level and structural-level disentanglement are complementary and essential for authentic counterfactuals generation.

Parameters Study. To investigate the effect of λ_D, λ_{IB}, and λ_F on the performance of GCFair, we vary the hyperparameters within the set $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4\}$. We compare GCFair's performance in terms of predictive and fairness under the different settings. The results in the Credit dataset are illustrated in Figure 2 (b), and additional results are provided in the appendix. We observe that increasing λ_F and λ_{IB} generally improves fairness but may reduce predictive performance when set too large, while increasing λ_D strengthens feature disentanglement but can also introduce an accuracy drop if overly constrained. Overall, the parameter study confirms that GCFair is stable over a broad range of hyperparameters and achieves consistent fairness gains with moderate regularization strengths.

6 Conclusion

Given the observed gap between assumptions underlying existing counterfactual fairness methods and their inability to generate authentic counterfactuals in graph-structured data, we introduced GCFair, a novel causal and variational framework explicitly designed for genuine counterfactual fairness. Our method leverages a causal model that effectively disentangles sensitive-related information from sensitive-irrelevant components in both node features and graph structures. Through novel disentanglement strategies explicitly designed to enhance model fairness, our approach ensures accurate and meaningful structural decomposition. Extensive evaluations demonstrated that GCFair effectively generates authentic counterfactual graphs, significantly improving fairness without sacrificing model performance. This work provides a practical solution, establishing a strong foundation for future research on counterfactual fairness in graph learning.

IJCAI-ECAI 2026 PREPRINT

Acknowledgements

This work was supported in part by the National Science Foundation (NSF) under Grant No. 2404039.

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] April Chen, Ryan A Rossi, Namyong Park, Puja Trivedi, Yu Wang, Tong Yu, Sungchul Kim, Franck Dernoncourt, and Nesreen K Ahmed. Fairness-aware graph neural networks: A survey. *ACM Transactions on Knowledge Discovery from Data*, 2024.

[Chen *et al.*, 2024b] Wei Chen, Yiqing Wu, Zhao Zhang, Fuzhen Zhuang, Zhongshi He, Ruobing Xie, and Feng Xia. Fairgap: Fairness-aware recommendation via generating counterfactual graph. *ACM Transactions on Information Systems*, 42(4):1–25, 2024.

[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*, 2012.

[Eastwood and Williams, 2018] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *6th International Conference on Learning Representations*, 2018.

[Guo *et al.*, 2023] Zhimeng Guo, Jialiang Li, Teng Xiao, Yao Ma, and Suhang Wang. Towards fair graph neural networks via graph counterfactual. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pages 669–678, 2023.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 2016.

[Kipf and Welling, 2016] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[Kusner *et al.*, 2017] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[Leonelli *et al.*, 2021] Sabina Leonelli, Rebecca Lovell, Benedict W Wheeler, Lora Fleming, and Hywel Williams. From fair data to fair data use: Methodological data fairness in health-related social media research. *Big Data & Society*, 8(1):20539517211010310, 2021.

[Liang *et al.*, 2022] Fan Liang, Cheng Qian, Wei Yu, David Griffith, and Nada Golmie. Survey of graph neural networks and applications. *Wireless Communications and Mobile Computing*, 2022(1):9261537, 2022.

[Liu *et al.*, 2020] Yanbei Liu, Xiao Wang, Shu Wu, and Zhitao Xiao. Independence promoted graph disentangled networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4916–4923, 2020.

[Ma *et al.*, 2019] Jianxin Ma, Peng Cui, Kun Kuang, Xin Wang, and Wenwu Zhu. Disentangled graph convolutional networks. In *International conference on machine learning*, pages 4212–4221. PMLR, 2019.

[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*, 2022.

[Pearl, 2009] Judea Pearl. Causal inference in statistics: An overview. 2009.

[Takac and Zabovsky, 2012] Lubos Takac and Michal Zabovsky. Data analysis in public social networks. In *International scientific conference and international workshop present day trends of innovations*, volume 1, 2012.

[Wang and Zhang, 2025] Zichong Wang and Wenbin Zhang. Fdgen: A fairness-aware graph generation model. In *Forty-second International Conference on Machine Learning*, 2025.

[Wang *et al.*, 2021] Jianian Wang, Sheng Zhang, Yanghua Xiao, and Rui Song. A review on graph neural network methods in financial applications. *arXiv preprint arXiv:2111.15367*, 2021.

[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.

[Wang *et al.*, 2023a] Zichong Wang, Giri Narasimhan, Xin Yao, and Wenbin Zhang. Mitigating multisource biases in graph neural networks via real counterfactual samples. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 638–647. IEEE, 2023.

[Wang *et al.*, 2023b] Zichong Wang, Charles Wallace, Albert Bifet, Xin Yao, and Wenbin Zhang. : Fairness-aware graph generative adversarial networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 259–275. Springer, 2023.

[Wang *et al.*, 2024] Zichong Wang, David Ulloa, Tongjia Yu, Raju Rangaswami, Roland Yap, and Wenbin Zhang. Individual fairness with group constraints in graph neural networks. In *Proceedings of the 27th European Conference on Artificial Intelligence*. IOS Press, 2024.

[Wang *et al.*, 2025a] Xuemin Wang, Tianlong Gu, Xuguang Bao, and Liang Chang. Towards fair graph neural

IJCAI-ECAI 2026 PREPRINT

networks via graph counterfactual without sensitive attributes. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pages 265–277. IEEE, 2025.

[Wang *et al.*, 2025b] Zichong Wang, Zhibo Chu, Thang Viet Doan, Shaowei Wang, Yongkai Wu, Vasile Palade, and Wenbin Zhang. Fair graph u-net: A fair graph learning framework integrating group and individual awareness. In *proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 28485–28493, 2025.

[Wang *et al.*, 2025c] Zichong Wang, Nhat Hoang, Xingyu Zhang, Kevin Bello, Xiangliang Zhang, Sundararaja Sitharama Iyengar, and Wenbin Zhang. Towards fair graph learning without demographic information. In *The 28th International Conference on Artificial Intelligence and Statistics*, pages 2107–2115, 2025.

[Wang *et al.*, 2025d] Zichong Wang, Zhipeng Yin, Fang Liu, Zhen Liu, Christine Lisetti, Rui Yu, Shaowei Wang, Jun Liu, Sukumar Ganapati, Shuigeng Zhou, et al. Graph fairness via authentic counterfactuals: Tackling structural and causal challenges. *ACM SIGKDD Explorations Newsletter*, 26(2):89–98, 2025.

[Wang *et al.*, 2025e] Zichong Wang, Zhipeng Yin, Zhen Liu, Roland HC Yap, Xiaocai Zhang, Shu Hu, and Wenbin Zhang. Redefining fairness: A multi-dimensional perspective and integrated evaluation framework. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 336–353. Springer, 2025.

[Wang *et al.*, 2025f] Zichong Wang, Zhipeng Yin, Liping Yang, Jun Zhuang, Rui Yu, Qingzhao Kong, and Wenbin Zhang. Fairness-aware graph representation learning with limited demographic information. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 354–371. Springer, 2025.

[Wang *et al.*, 2025g] Zichong Wang, Zhipeng Yin, Roland HC Yap, and Wenbin Zhang. Ai fairness beyond complete demographics: Current achievements and future directions. *ECAI*, 2025.

[Wang *et al.*, 2026a] Zichong Wang, Tongliang Liu, and Wenbin Zhang. Guic: Certified graph unlearning with individual fairness guarantees. In *proceedings of the AAAI conference on artificial intelligence*, volume 40, pages 35820–35828, 2026.

[Wang *et al.*, 2026b] Zichong Wang, Jie Yang, Jun Zhuang, Puqing Jiang, Mingzhe Chen, Ye Hu, and Wenbin Zhang. Fair graph learning with limited sensitive attribute information. In *proceedings of the AAAI conference on artificial intelligence*, volume 40, pages 39423–39431, 2026.

[Wang *et al.*, 2026c] Zichong Wang, Zhipeng Yin, Mo Sha, Xiaofeng Gao, Xiaoli Li, and Wenbin Zhang. Towards fair graph learning without demographic supervision. In *The 35th International Joint Conference on Artificial Intelligence*, 2026.

[Wang *et al.*, 2026d] Zichong Wang, Zhipeng Yin, and Wenbin Zhang. A unified framework for fair graph generation: Theoretical guarantees and empirical advances. *Advances in Neural Information Processing Systems*, 2026.

[Wu *et al.*, 2020] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 2020.

[Xiao *et al.*, 2022] Naijia Xiao, Aifen Zhou, Megan L Kempher, Benjamin Y Zhou, Zhou Jason Shi, Mengting Yuan, Xue Guo, Linwei Wu, Daliang Ning, Joy Van Nostrand, et al. Disentangling direct from indirect relationships in association networks. *Proceedings of the National Academy of Sciences*, 119(2):e2109995119, 2022.

[Yang *et al.*, 2024] Shangshang Yang, Mingyang Chen, Ziwen Wang, Xiaoshan Yu, Panpan Zhang, Haiping Ma, and Xingyi Zhang. Disengcd: A meta multigraph-assisted disentangled graph learning framework for cognitive diagnosis. *Advances in Neural Information Processing Systems*, 37:91532–91559, 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.

[Yin *et al.*, 2026] Zhipeng Yin, Zichong Wang, Zhong Chen, Jack Yang, Xin Ning, and Wenbin Zhang. Disentangled graph-enhanced large language models for fair learning. In *Proceedings of the 35th International Joint Conference on Artificial Intelligence*, 2026.

[Zhang and Ntoutsi, 2019] Wenbin Zhang and Eirini Ntoutsi. Faht: An adaptive fairness-aware decision tree classifier. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 1480–1486. International Joint Conferences on Artificial Intelligence Organization, 2019.

[Zhang and Weiss, 2022] Wenbin Zhang and Jeremy C Weiss. Longitudinal fairness with censorship. In *proceedings of the AAAI conference on artificial intelligence*, volume 36, 2022.

[Zhang *et al.*, 2021] Wenbin Zhang, Liming Zhang, Dieter Pfoser, and Liang Zhao. Disentangled dynamic graph deep generation. In *SDM*, 2021.

[Zhang *et al.*, 2023] Wenbin Zhang, Tina Hernandez-Boussard, and Jeremy Weiss. Censored fairness through awareness. In *Proceedings of the AAAI conference on artificial intelligence*, 2023.

[Zhang *et al.*, 2025] Wenbin Zhang, Shuigeng Zhou, Toby Walsh, and Jeremy C Weiss. Fairness amidst non-iid graph data: A literature review. *AI Magazine*, 2025.

[Zhang, 2024a] Wenbin Zhang. Ai fairness in practice: Paradigm, challenges, and prospects. *Ai Magazine*, 2024.

[Zhang, 2024b] Wenbin Zhang. Fairness with censorship: Bridging the gap between fairness research and real-world deployment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2024.

[Zhao *et al.*, 2022] Tong Zhao, Gang Liu, Daheng Wang, Wenhao Yu, and Meng Jiang. Learning from counterfactual links for link prediction. In *International Conference on Machine Learning*, pages 26911–26926. PMLR, 2022.

IJCAI-ECAI 2026 PREPRINT