fairGNN-WOD: Fair Graph Learning Without Demographics

Zichong Wang[1], **Fang Liu**[2], **Shimei Pan**[3], **Jun Liu**[4], **Fahad Saeed**[1], **Meikang Qiu**[5] and **Wenbin Zhang**[1*]
[1]Florida International University, FL, USA
[2]University of Notre Dame, IN, USA
[3]University of Maryland Baltimore County, MD, USA
[4]Northeastern University, MA, USA
[5]Augusta University, GA, USA

Abstract

Graph Neural Networks (GNNs) have excelled in diverse applications due to their outstanding predictive performance, yet they often overlook fairness considerations, prompting numerous recent efforts to address this societal concern. However, most fair GNNs assume complete demographics by design, which is impractical in most real-world socially sensitive applications due to privacy, legal, or regulatory restrictions. For example, the Consumer Financial Protection Bureau (CFPB) mandates that creditors ensure fairness without requesting or collecting information about an applicant's race, religion, nationality, sex, or other demographics. To this end, this paper proposes fairGNN-WOD, a first-of-its-kind framework that considers mitigating unfairness in graph learning without using demographic information. In addition, this paper provides a theoretical perspective on analyzing bias in node representations and establishes the relationship between utility and fairness objectives. Experiments on three real-world graph datasets illustrate that fairGNN-WOD outperforms state-of-the-art baselines in achieving fairness but also maintains comparable prediction performance.

1 Introduction

Graph Neural Networks (GNNs) have emerged as a powerful approach for learning from graph-structured data, finding applications in areas such as social network analysis [Peng *et al.*, 2016], financial markets [Wang *et al.*, 2023b], and item recommendations [Wu *et al.*, 2021]. Despite their significant success, GNNs, like many ML algorithms, have been observed to potentially discriminate against certain populations as identified by the *demographics* (*e.g.*, gender or race) [Zhang *et al.*, 2025]. To this end, many efforts have been taken towards fair graph learning [Dai and Wang, 2021; Wang *et al.*, 2024c; Wang *et al.*, 2024a; Wang *et al.*, 2025a; Ling *et al.*, 2023] that aim to ensure similar outcome statistics for the algorithmic decisions (*e.g.*, prediction accuracy and true positive rate) across certain demographic subgroups, assuming complete availability of demographics.

However, this assumption is unrealistic in many real-world scenarios where collecting or using demographics (*i.e.*, protected features) is infeasible due to privacy, legal, regulatory restrictions [Lahoti *et al.*, 2020] or out of fear of discrimination and social desirability [Wang *et al.*, 2025b]. This discrepancy highlights the gap between the design of a "fair" model in research environments and their real-world scenarios. In this paper, we explore such a fair graph learning problem, where the assumption of guaranteed demographics does not hold, but models are still required to treat different demographic groups fairly and equally. Below is an illustration of such a real-life fair graph problem.

Example 1: A tech company employs a fair GNN to enhance its hiring processes by integrating applicants' social network data (*e.g.*, potential team fit) [Liu *et al.*, 2024] to identify top candidates efficiently while ensuring equitable treatment across all demographic groups and preventing biases in distributing key social benefits such as employment opportunities. However, the availability of demographics is not guaranteed due to: i) Individuals may choose to withhold their demographic information when applying for jobs if they feel underrepresented or could be potentially discriminated against. For instance, women applying for software engineering roles, which have historically been dominated by men, may choose not to disclose their gender on job boards (*e.g.*, Linkedin) [Friedmann and Efrat-Treister, 2023]. ii) Legal constraints can lead to the complete absence of certain demographics. For example, financial institutions, including those collaborating with tech firms for automated hiring solutions, are bound by regulations like those from the Consumer Financial Protection Bureau (CFPB). These regulations mandate that fairness be achieved without collecting or using specific demographic details such as an applicant's race, color, religion, nationality, or gender [Chai *et al.*, 2022].

In the aforementioned scenarios where demographic information is missing, existing fair graph methods that rely on complete demographics become inapplicable. This limitation highlights the urgent need for developing new fair graph approaches that can function effectively even when the assumption of complete demographics does not hold. Indeed,

*Corresponding author.

studies [Yan *et al.*, 2020; Grari *et al.*, ; Wang *et al.*, 2025c; Lahoti *et al.*, 2020] have begun to explore achieving fairness without demographics. However, these methods are primarily designed for non-graph data and face significant challenges when adapted to graph data. Furthermore, the core idea behind most mitigation bias approaches is removing the *demographic-relevant information* (*i.e.*, features from which demographics can be inferred, such as the mustache for gender), thereby enforcing GNNs to make decisions independent of the demographic information. While this strategy can enhance fairness, it may degrade performance by eliminating information related to tasks and demographics.

Despite the importance of achieving graph fairness without demographics, this remains a highly open research area with several complex and unique challenges: **i) Difficulties of Preventing Label Information Leakage During Demographics Inference:** Inferring missing demographics from observed data should avoid interference from label information to prevent ethical issues, such as inferring an applicant's race based on their bank's loan decision outcomes. However, label information cannot be easily excluded from the inference process, as it is embedded within label-related features and transformed hidden representations within the model. **ii) Avoiding optimization exploitation in fairness-aware learning:** When reducing model bias without demographics, the model might exploit the subgroup identification step to artificially improve fairness metrics. This could happen by deliberately misclassifying demographic information or altering the assignment of samples from the worst-treated groups. For instance, the model could minimize group disparities by incorrectly assigning samples to different demographic groups instead of achieving true fairness. **iii) Precise imposition of fairness constraints:** Most existing fair-GNN designs enforce fairness constraints across the entire node representation, which can inadvertently remove task-related demographic information, leading to performance losses. An optimal balance between GNN performance and fairness requires targeted constraints that mitigate demographic bias while preserving essential task-related information.

To tackle the aforementioned challenges, we introduce a novel two-stage conditioning framework, *Fair Graph Neural Network Without Demographics* (fairGNN-WOD), which leverages Bayesian variational autoencoders (VAEs) coupled with causal modeling to infer missing demographic information from observed graph data. The inferred demographic is then used as additional information to help downstream GNNs learn fair representations while retaining task-related demographic information. *To our knowledge, this is the first work to enable fair graph learning without complete demographics, while also achieving demographic-dependent fairness-aware learning.* Specifically, our approach begins with a comprehensive bias analysis that examines how demographic information can propagate through node representations and potentially lead to disparate outcomes across different subgroups. Building on these insights, fairGNN-WOD generates accurate demographic proxies by filtering out non-causal and superfluous relations from the observed graph. These proxies serve as a basis for identifying demographic-relevant information. Subsequently, fairGNN-WOD implements fairness constraints specifically designed to achieve de-identification of demographics within these representations. This method not only enhances fairness but also ensures that critical task-related demographic information is retained, thereby maintaining the predictive power of the model. The key contributions of this work can be summarized as follows: **i)** We present a novel perspective on how demographic information propagates through graph structures to cause disparate treatment, offering theoretical insights into how demographic information can disproportionately affect node embeddings and ultimately lead to unfair classification outcomes. **ii)** Building on this bias analysis, we introduce a two-stage framework, fairGNN-WOD, that achieves graph fairness without demographics while preserving task-related information for improved node classification performance. **iii)** Extensive experiments on benchmark datasets to demonstrate the effectiveness of fairGNN-WOD in mitigating unfairness and maintaining comparable performance.

2 Related Work

Graph Neural Networks. GNNs have demonstrated widespread utility across various tasks involving graph-structured data [Kipf and Welling, 2016; Zhao *et al.*, 2022b; Wu *et al.*, 2020]. Their remarkable success has propelled GNNs to the forefront of both research and practical applications, extending their reach into high-risk decision-making systems [Zhang and Weiss, 2022; Zhang *et al.*, 2023; Wang and Zhang, 2024]. For example, GNNs can assist financial institutions in critical functions like evaluating credit card applications or making loan approval decisions [Wang *et al.*, 2024b]. The applications in these areas require GNNs to be not only effective but also fair [Zhang *et al.*, 2025]. Therefore, there is a demand to design fair GNNs to mitigate biases and ensure fair outcomes in graph-based tasks [Zhang, 2024].

Fairness in Graph Learning. In the context of fairness in graph learning, most existing studies aim to ensure similar outcome statistics across demographic groups to prevent disparities in favorable outcomes [Guo *et al.*, 2023; Wang *et al.*, 2023c; Wang *et al.*, 2023a; Wang and Zhang, 2025]. Despite these methods achieving some success, they presume the presence of demographics to quantify and mitigate bias. However, this assumption is unrealistic in many real-world applications due to the practicality and regulatory limitations [Grari *et al.*, 2021]. To this end, FairGNN [Dai and Wang, 2021] is proposed to learn fair GNNs with limited demographics using a demographic estimator to predict the demographic information while improving fairness via adversarial learning. However, FairGNN still assumes the availability of partial demographic information. Group-Free [Liu *et al.*, 2023] aims to use homophily in social networks to reduce inequality in outcome prediction solely based on the similarities of individuals without defining groups. However, there is no guarantee that the uncovered groups are consistent with the real demographics of interest. In addition, several approaches [Chen *et al.*, 2019; Kallus *et al.*, 2022; Lahoti *et al.*, 2020; Wang *et al.*, 2025c] have explored fairness without demographic information in non-graph data, but these methods cannot be easily extended to graph data.

To this end, our work addresses a new fair graph learning problem where the assumption of complete demographic information does not hold. Furthermore, we explore a demographic-dependent fair graph learning paradigm that relaxes the requirement of demographic independence to enhance its predictive power while preserving fairness.

3 Notations

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ denote an undirected attributed graph, comprised of a set of $|\mathcal{V}| = n$ nodes and a set of $|\mathcal{E}| = m$ edges. $X \in \mathrm{R}^{n \times d}$ is node feature matrix whose i-th row represents a d-dimensional feature vector of the i-th node v_i. $\mathbf{A} \in \{0,1\}^{n \times n}$ is the adjacency matrix where $\mathbf{A}_{i,j} = 1$ indicates that there exists edge $e_{i,j} \in \mathcal{E}$ between node v_i and v_j, and $\mathbf{A}_{i,j} = 0$ otherwise. In this paper, we assume that both ground-truth labels and demographics are binary variables for convenience. We let $S \in \{0,1\}^{n \times 1}$ denote the binary demographic, where s_i is the demographic value of v_i. We use $S_d = \{\forall\ v_i \in \mathcal{V} | s_i = 0\}$ denotes the deprived group (*e.g.,* female) and $S_f = \{\forall\ v_i \in \mathcal{V} | s_i = 1\}$ denotes the favored group (*e.g.,* male). For node classification, each node is also associated with a one-hot ground-truth node label y_i where $\hat{y}_i$ is the label of v_i. We also assume $y_i = 1$ denotes the granted label and $y_i = 0$ denotes the rejected label.

4 Methodology

4.1 Root Bias in Graph Learning

This section examines two sources of bias inherent in fair graph learning without demographics, establishing the groundwork for corresponding bias mitigation strategies.

Bias in demographic identification. We begin by examining the factors that cause biased inferences of missing demographic information. Existing methods, primarily in non-graph domains, reconstruct missing demographics using observed data and prior knowledge (*i.e.*, $X \rightarrow S$) [Grari *et al.*, 2021], while in graphs, the reconstruction incorporates both data and structures (*i.e.*, $\{X, A\} \rightarrow S$), but they can unintentionally embed implicit biases. Specifically, if the distribution of nodes receiving favorable outcomes overlaps disproportionately with features indicative of the "male" group, the posterior for S can end up assigning a high "male" likelihood to those nodes, thereby propagating biased associations. To address this issue, we need to minimize the correlation between the label and the latent space, ensuring that the inferred demographics remain independent of outcomes.

Bias in node classification. We further analyze how bias emerges in GNN predictions through node representations. When GNNs learn node representations, they inevitably capture and potentially amplify the effect of demographic information from both node features and graph structure. As shown in Figure 1, the node representation can be divided into demographic-relevant representation h_S and demographic-irrelevant representation $h_{\overline{S}}$. Although both of them contain task-related information for predictions (*e.g.*, loan decision), h_S introduces bias into the decision-making process because it is influenced by demographic-relevant features X (*e.g.*, height) and graph structure information A (*e.g.*, neighbor demographic information (gray dashed line)). In addition, while

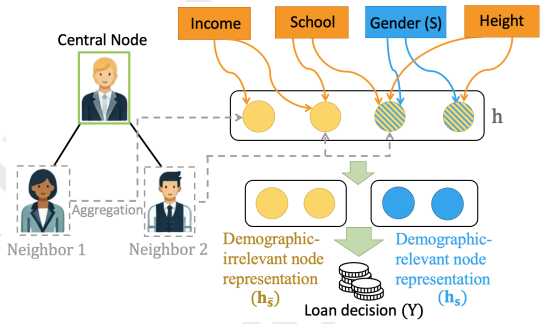

Figure 1: A causal relationship between node demographics and their predictions arises in node representations.

removing h_S and only using $h_{\overline{S}}$ for prediction would prevent demographic bias, this approach leads to suboptimal performance as it also eliminates valuable task-related information contained in h_S. Therefore, achieving fair node classification requires disentangling the demographic information embedded in h_S while preserving its task-relevant components to maintain both fairness and model utility.

4.2 The Proposed Framework - fairGNN-WOD

Based on the above bias analysis, we propose a novel framework called fairGNN-WOD that addresses bias both in demographic identification and node classification. Our framework combines latent representation learning and graph structural modeling to leverage their complementary strengths: the former excels at inferring missing information through learning underlying data distributions but cannot effectively process graph-structured data, while the latter is powerful at capturing graph structural patterns and ensure fairness simultaneously but lacks the capability to infer the necessary missing information. Specifically, as illustrated in Figure 2, fairGNN-WOD employs a two-stage framework. In the first stage, it leverages observed prior knowledge to infer missing demographic information while ensuring fairness by excluding node label information. In the second stage, it uses the inferred sensitive attributes as proxies to mitigate bias in graph learning. With some simplifying design choices, our framework uses a VAE to infer missing demographic information in the first stage, followed by disentangling node representations into demographic-relevant and demographic-independent components in the second stage. This disentanglement allows us to maintain task-relevant information that is relevant to the demographic information while mitigating bias in predictions. The following subsections detail each stage of our framework.

4.3 Missing Demographic Identification

We begin by introducing how fairGNN-WOD infers missing demographic information. The core idea is that the estimates of the latent variable representing the true demographic information are consistent with the underlying data generation process. By learning this process, we can derive a latent representation Z that encapsulates as much information about the true demographic information S as possible. We achieve this through Bayesian inference approximation,

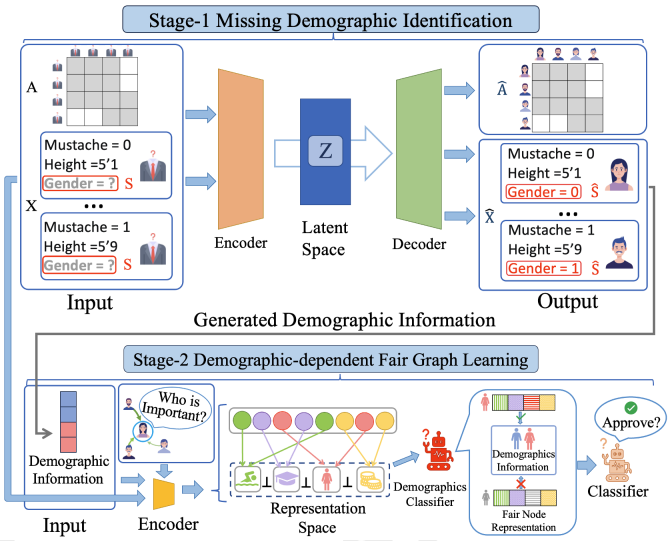

Figure 2: The overview of proposed fairGNN-WOD.

using observed graph data (*i.e.*, graph structure information A and node features X) and prior structural assumptions. Specifically, if we can accurately recover the joint distribution $P(Z, S, A, X)$, we can effectively recover the missing demographic information [Louizos *et al.*, 2017]. Hence, we factorize $P(Z, S, X, A)$ as follows:

$$P(Z, S, X, A) = P(Z)\, P(S \mid Z)\, P(A \mid S)\, P(X \mid S, A) \quad (1)$$

where $P(Z)$ denotes the prior over Z, typically modeled as a standard Gaussian distribution $\mathcal{N}(0, I)$, where I denotes the identity matrix. Moreover, $P(A|S)$ and $P(X|A, S)$ represent the decoders of the structural information and node features.

To approximate the intractable joint distribution $P(Z, S, X, A)$, we employ variational inference with neural parameterizations. Specifically, we introduce a variational distribution $Q(Z \mid A, X)$ to approximate $P(Z \mid A, X)$, and maximize the Evidence Lower Bound (ELBO) [Kingma and Welling, 2013] of the marginal data likelihood:

$$\begin{aligned}
&\log P(X, A) \geq \\
&\mathbb{E}_{q_{\phi,\psi}(Z,S|X,A)}[\log P(X|Z, S, A) + \log P(A|Z, S) \\
&= \mathbb{E}_{q_\phi(Z|X,A)}[\mathbb{E}_{q_\psi(S|Z)}[\log P(X|S, A) + \log P(A|S) \quad (2)\\
&+\log P(S|Z)]+\log P(Z)-\log q_\psi(S|Z)]-\log q_\phi(Z|X, A)]
\end{aligned}$$

where q_ϕ and q_ψ denote the encoders parameterized by ϕ and q_ψ that formulate the variational distributions of Z given (X, A) and S given Z, respectively. In addition, the maximization can be performed using stochastic gradient ascent and the reparameterization trick [Kingma and Welling, 2013].

On the other side, a fair inference of S would be based on independence between S (or Z) and Y. To impose this constraint, we add a penalty term onto the ELBO in Equation (2) that penalizes high dependency between Z and Y. Specifically, we extend the Hirschfeld-Gebelein-Rényi (HGR) maximal correlation [Gebelein, 1941] to quantify the dependency, which can be linear or non-linear, as defined in Definition 4.1.

Definition 4.1 (*ZY-correlation*). Given latent space Z and outcome Y, HGR maximal correlation is defined as:

$$\mathrm{HGR}(Z, Y) = \sup_{p_Z, p_Y} \frac{\mathbb{E}(p_Z(Z) p_Y(Y))}{\sqrt{\mathbb{E}({p_Z}^2(Z))\mathbb{E}({p_Y}^2(Y))}} \quad (3)$$

where ρ denotes the Pearson correlation coefficient, p_Z and p_Y are measurable probability density functions with positive and finite variance. We apply normalization $\mathbb{E}(p_Z(Z)) = \mathbb{E}(p_Y(Y)) = 0$ and $\mathbb{E}({p_Z}^2(Z)) = \mathbb{E}({p_Y}^2(Y)) = 1$ before the calculation of the HGR maximal corelation before Y and Z, which equals 0 if Z and Y are independent, and 1 otherwise.

With the HGR correlation, the objective function in Equation (2) can be reformulated as:

$$\begin{aligned}
&\mathbb{E}_{q_\phi(Z|X,A)}[\mathbb{E}_{q_\psi(S|Z)}[\log P(X|S, A)+\log P(A|S)+\log P(S|Z)] \\
&+ \log P(Z)-\log q_\psi(S|Z)]-\log q_\phi(Z|X, A)] - \lambda \cdot \mathrm{HGR}(Z, Y)
\end{aligned} \quad (4)$$

where λ is the hyperparameter that balances the maximization of the ELBO and the minimization of the penalty term.

To maximize the updated objective function, a dual-phase maximum optimization strategy is employed. Specifically, in the max phase, we use gradient ascent to estimate HGR(Z, Y), where p_Z and p_Y in Equation (3) are approximated via two interconnected neural networks that are optimized via, say stochastic gradient ascent. In the max phase, we maximize the updated objective function, where the penalty term would promote independence between Y and Z. This alternate optimization scheme allows us to capture and refine the estimated HGR between Z and Y with each iteration, enhancing the stability and accuracy of the learning process.

4.4 Demographic Disentangled Fair Graph Learning

The first stage yields the complete demographic information, which is then utilized in the second stage for fair prediction. Specifically, the second stage aims to remove the demographic-relevant information h_S, enabling GNNs to produce predictions that do not depend on demographic information. Following existing fairness approaches [Zhang *et al.*, 2025], a straightforward solution is to remove h_S completely and use only $h_{\overline{S}}$ for prediction, as this ensures decisions are made independently of demographic information. However, while this approach used by prior work can enhance fairness, it often leads to unnecessary performance degradation by removing task-relevant information that correlates with demographics, such as the school shown in Figure 1. To maintain the model's effectiveness, we thus keep task-related information in h_S while removing its connection to demographic information. This method allows us to use more predictive information while keeping the model fair. To implement this idea, we need to: i) find demographic-relevant information in node representations, and ii) remove demographic information while keeping task-related information for demographic-relevant node representations.

In the first task, we aim to find demographic-relevant node representations from the whole node representations. Inspired by disentangled representation learning [Ma *et al.*,

2019], we decompose the node representation into multiple ($N_c > 1$) independent components (represented by green, purple, and red circles in Figure 2) and identify the demographic-relevant components among them. Each component represents a subspace and focuses on a specific latent factor (*e.g.*, shared interests) in the node representation. When the separation is complete, only one component contains the demographic information.

To this end, we introduce an adaptive assigner, implemented via a multilayer perceptron that assigns different weights to the neighbors $\mathcal{N}(v_i)$ of a node v_i in a graph with respect to each latent factor of node v_i in the disentangled node aggregation. This mechanism allows learning separate representations for distinct latent factors underlying the graph structure. Specifically, suppose $v_j \in \mathcal{N}(v_i)$, their node features x_i and x_j are inputs to the adaptive assigner F_ψ: $\psi_{v_i,v_j} = F_\psi([x_i; x_j])$, where $[x_i; x_j]$ denotes the concatenation of features x_i and x_j, and $\psi_{v_i,v_j} \in \mathbb{R}^{N_c}$ is a vector representing the importance scores of node v_j in the latent factors $c = 1, \ldots, N_c$ associated with node v_i. To ensure disentanglement among the entries in ψ_{v_i,v_j}, we add a regularization term to encourage independence among the learned components (*e.g.*, minimizing the mutual information between different subspaces). We then apply a softmax function over the latent factors ψ_{v_i,v_j} to obtain the normalized weights: $\omega_{v_i,v_j} = \text{softmax}\left(\psi_{v_i,v_j}\right)$. The elements $\omega^c_{v_i,v_j} \in \omega_{v_i,v_j}$ for $c = 1, \ldots, N_c$ represents the probability that the connection between i and j is influenced by latent factor c.

Next, we employ multiple disentangled layers for graph convolution, where each layer consists of multiple channels (*i.e.*, N_c latent factors) that share the same network architecture. Each channel is tailored to amplify a specific latent factor, enhancing the representation's relevance to that factor. Given node representation h_{v_i} of dimension d_r of node v_i, we apply a linear transformation $F^c_R(\cdot)$ to obtain the initial representations. This step is applied independently to each channel, producing N_c node representations corresponding to the N_c latent factors. This dimensionality reduction is applied independently for each channel, producing a series of reduced node attributes corresponding to the various latent factors. In addition, the node representation at the l^{th} layer for node v_i in channel c is denoted as $h^{(l)}_{c,i}$. We concatenate the outputs from all channels to obtain the full node representation at layer l^{th}: $h^{(l)}_i = [h^{(l)}_{1,i}, h^{(l)}_{2,i}, \ldots, h^{(l)}_{N_c,i}]$. To update each node's representation, we utilize the edge weights predicted by the adaptive assigner. The basic operation between the l^{th} and $(l+1)^{th}$ layers within the c^{th} channel is $h^{(l+1)}_{c,i} = \sigma\left(\sum_{v_j \in \mathcal{N}(i)} \omega^c_{i,j} \cdot h^{(l)}_{c,j} \cdot \mathbf{W}^{c,(l)}\right)$, where $\mathcal{N}(i)$ denotes the neighbors of node v_i, ω^c_{i,v_j} is the weight for channel c from node v_i to node v_j, $\mathbf{W}^{c,(l)}$ is the learnable weight matrix for layer l in channel c, and $\sigma(\cdot)$ is an activation function. Building on this, we can assemble the matrix of disentangled representations for all nodes, *i.e.*, $K = \left[k_1, k_2, \ldots, k_{N_c}\right]$, where k_c corresponds to the channel-c representations gathered across all nodes. For example, $k_1 = \left[h_{1,1}, h_{2,1}, \ldots, h_{i,1}\right]$, indicating the embeddings for channel 1 across all the nodes.

Although the above process can disentangle the node representation in different channels, it may neglect the independence among different latent factors, such as c_1 = "gender" and c_2 = "girls' school". To promote orthogonality and disentanglement among the channels in the node representations in the last layer, we formulate an Independence Constraint using the Maximum Mean Discrepancy (MMD) [Gretton *et al.*, 2006]. Specifically, let k_{c_1} and k_{c_2} denote the embeddings across all n nodes in channels c_1 and c_2, respectively. We define the MMD between these two embedding sets as:

$$\text{MMD}^2\left(k_{c_1}, k_{c_2}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} f\left(k_{c_1,i}, k_{c_1,j}\right) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} f\left(k_{c_2,i}, k_{c_2,j}\right) - \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} f\left(k_{c_1,i}, k_{c_2,j}\right) \tag{5}$$

where $f(\cdot,\cdot)$ is a kernel function (*e.g.*, a Gaussian RBF kernel). To ensure *all* channel pairs are mutually independent, we sum over every distinct pair (c_1, c_2) with $c_1 < c_2$ and formulate an Independence Constraint as:

$$\mathcal{L}_I = -\sum_{c_1=1}^{N_c}\sum_{c_2=c_1+1}^{N_c} \text{MMD}^2\left(k_{c_1}, k_{c_2}\right) \tag{6}$$

Minimizing $\mathcal{L}_I$ drives the empirical distributions of any two channels, c_1 and c_2, to be dissimilar, thus promoting independence in their learned representations. With fully disentangled channels, we need to identify which channels capture demographic information. To achieve this, we utilize a discriminator that predicts the demographic label from the channel-specific embeddings. For each channel c, we feed the corresponding embeddings k_c into the classifier to produce a predicted probability $\hat{y}_{s_i,c}$. We then define the classification loss as follows:

$$\mathcal{L}_D = -\frac{1}{|\mathcal{V}_L|}\sum_{v_i \in \mathcal{V}_L}\sum_{c=1}^{N_c}\left[y_{s_i}\log\left(\hat{y}_{s_i,c}\right) + \left(1 - y_{s_i}\right)\log\left(1 - \hat{y}_{s_i,c}\right)\right] \tag{7}$$

where y_{s_i} is the obtained demographic information from stage 1 for node v_i, and $\hat{y}_{s_i,c}$ is the predicted demographics.

In the second task, we focus on processing the identified demographic-relevant representations to further enhance predictive fairness. Specifically, we employ a learnable masking mechanism on the identified demographic-relevant representation h_S. Through this procedure, we obtain a de-identified version, denoted by $\not{h}_S = h_S \odot \mathbf{m}$, where $\mathbf{m}$ is a masking vector designed to obscure explicit demographic information. The masked representation $\not{h}_S$ (represented by a gray circle in Figure 2) is then used for downstream prediction tasks. To ensure that the mask effectively eliminates demographic cues from h_S, we penalize the covariance between the obfuscated demographic attribute and the label prediction. Formally, we minimize the absolute covariance:

$$\mathcal{L}_F = |\text{Cov}(S, \hat{y})| = |\mathbb{E}[(S - \mathbb{E}(S))(\hat{y} - \mathbb{E}(\hat{y}))]| \tag{8}$$

where $|\cdot|$ indicates the absolute value.

We also include a performance loss $\mathcal{L}_P$ for utility maximization, as in Equation 9:

Dataset	Metrics \ Methods	GCN	GIN	FairKD	KSMOTE	FairRF	Reckoner	fairGNN-WOD
Credit	**Accuracy** (↑)	0.781 ± 0.016	$\mathbf{0.787 \pm 0.018}$	0.711 ± 0.012	0.736 ± 0.009	0.735 ± 0.007	0.736 ± 0.021	0.754 ± 0.052
	F1-Score (↑)	0.868 ± 0.023	$\mathbf{0.877 \pm 0.018}$	0.796 ± 0.023	0.817 ± 0.012	0.809 ± 0.022	0.817 ± 0.015	0.861 ± 0.018
	SPD(↓)	0.117 ± 0.013	0.106 ± 0.011	0.094 ± 0.036	0.071 ± 0.003	0.067 ± 0.017	0.068 ± 0.017	$\mathbf{0.036 \pm 0.015}$
	EOD(↓)	0.096 ± 0.017	0.088 ± 0.013	0.075 ± 0.042	0.055 ± 0.013	0.057 ± 0.018	0.055 ± 0.014	$\mathbf{0.027 \pm 0.013}$
Pokec-z	**AUC** (↑)	0.699 ± 0.024	0.691 ± 0.015	0.673 ± 0.021	0.697 ± 0.024	0.690 ± 0.014	0.692 ± 0.020	$\mathbf{0.703 \pm 0.041}$
	F1-Score (↑)	$\mathbf{0.622 \pm 0.012}$	0.613 ± 0.007	0.592 ± 0.013	0.611 ± 0.018	0.617 ± 0.019	0.603 ± 0.021	0.621 ± 0.032
	SPD(↓)	0.075 ± 0.025	0.061 ± 0.014	0.045 ± 0.014	0.037 ± 0.017	0.032 ± 0.012	0.036 ± 0.018	$\mathbf{0.028 \pm 0.013}$
	EOD(↓)	0.062 ± 0.013	0.057 ± 0.007	0.048 ± 0.009	0.039 ± 0.010	0.034 ± 0.012	0.033 ± 0.010	$\mathbf{0.029 \pm 0.015}$
Pokec-n	**AUC** (↑)	0.689 ± 0.015	0.685 ± 0.018	0.663 ± 0.016	0.669 ± 0.013	0.673 ± 0.013	0.675 ± 0.028	$\mathbf{0.691 \pm 0.024}$
	F1-Score (↑)	$\mathbf{0.631 \pm 0.022}$	0.629 ± 0.008	0.603 ± 0.023	0.611 ± 0.018	0.616 ± 0.032	0.619 ± 0.032	0.626 ± 0.029
	SPD(↓)	0.084 ± 0.013	0.078 ± 0.017	0.067 ± 0.015	0.061 ± 0.005	0.056 ± 0.027	0.042 ± 0.008	$\mathbf{0.028 \pm 0.013}$
	EOD(↓)	0.078 ± 0.019	0.071 ± 0.027	0.064 ± 0.013	0.066 ± 0.013	0.061 ± 0.016	0.052 ± 0.011	$\mathbf{0.038 \pm 0.014}$

Table 1: Comparison results of fairGNN-WOD with baseline methods across real-world datasets. In each row, the best result is indicated in bold, while the runner-up result is marked with an underline.

$$\mathcal{L}_P = \frac{1}{|\mathcal{V}_L|} \sum_{v_i \in \mathcal{V}_L} -\Big[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \Big] \tag{9}$$

The final objective function of demographic-dependent fair graph learning, as presented in Equation 10, brings together the above loss functions.

$$\min \ \mathcal{L}_{GNN} = \mathcal{L}_P + \alpha \mathcal{L}_I + \alpha \mathcal{L}_D + \beta \mathcal{L}_F \tag{10}$$

where α and β are tunable hyperparameters controlling the weights of the various elements: i) $\mathcal{L}_P$ aims to minimize the prediction loss, ii) $\mathcal{L}_I$ encourages the decomposition of learned representations into different independent channels and distinguishes between demographics relevant and irrelevant representations, and iii) $\mathcal{L}_F$ aims to mitigate demographics-related information in node representation thereby improving the fairness of the model. Note that fairGNN-WOD is trained in a sequential two-stage manner, *i.e.*, first optimizing VAE and then optimizing for fair GNN in the second stage while freezing the VAE model.

5 Experiment

5.1 Experimental Setup

Datasets. Our experiments are conducted on three widely used datasets: the Credit dataset [Yeh and Lien, 2009], Pokec-z and Pokec-n datasets [Takac and Zabovsky, 2012]. The **Credit** dataset contains default payment records for credit card holders, where nodes are connected based on similarities in their purchase and payment patterns. The age of the individuals serves as the demographics. The **Pokec-z** and **Pokec-n** datasets are derived from a popular social network in Slovakia, representing user networks from two different provinces. In these social networks, nodes represent users with features including gender, age, and interests, while edges represent friendships between users. The demographics for both datasets are the region. To simulate cases of missing demographics, we mask all demographic information in the training and validation sets.

Baselines. We compare our fairGNN-WOD method with the following baseline methods, categorized into two groups: i) Vanilla graph model: GCN [Kipf and Welling, 2016], GIN [Xu *et al.*, 2018]. ii) Extend Fairness Method: FairKD [Chai *et al.*, 2022]: Enhances fairness without demographics through partial knowledge distillation. KSMOTE [Yan *et al.*, 2020]: Creates pseudo-groups through clustering and enforces fairness through prediction regularization. FairRF [Zhao *et al.*, 2022a]: Minimizes correlation between predictions and demographic-related features. Reckoner [Ni *et al.*, 2024]: Achieves group fairness through learnable noise and knowledge-sharing in a dual-model architecture. For the methods not originally designed for graph data, we adapt them to work with our GNN backbone using the authors' original implementations.

Metrics: We use Accuracy and F1-score to evaluate the utility 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 Result

Performance Comparison. Table 1 presents a comparison between fairGNN-WOD and baseline methods, demonstrating that fairGNN-WOD consistently outperforms all six baselines across various metrics in both predictive performance and fairness. Specifically, i) fairGNN-WOD shows a significant improvement in fairness over the vanilla method (no fairness considerations). This improvement is due to the ability of fairGNN-WOD to effectively infer a demographic information proxy that highly correlates with true demographics, providing a foundation for the subsequent unfairness mitigation. ii) fairGNN-WOD demonstrates better fairness performance than methods adapted from non-graph domains. This superior performance is attributed to fairGNN-WOD's ability to identify and segregate demographic-related information in the node representations so that it can apply fairness constraints in a more precise and targeted manner, thereby reducing the impact of biases in the predictions. iii) fairGNN-WOD surpasses the other methods in utility performance in most cases and fairness performance in all cases. Overall, the experimental results demonstrate the effectiveness of fairGNN-WOD in improving fairness while achieving comparable performance. Unlike the existing fairness methods that might apply fairness constraints indiscrimi-

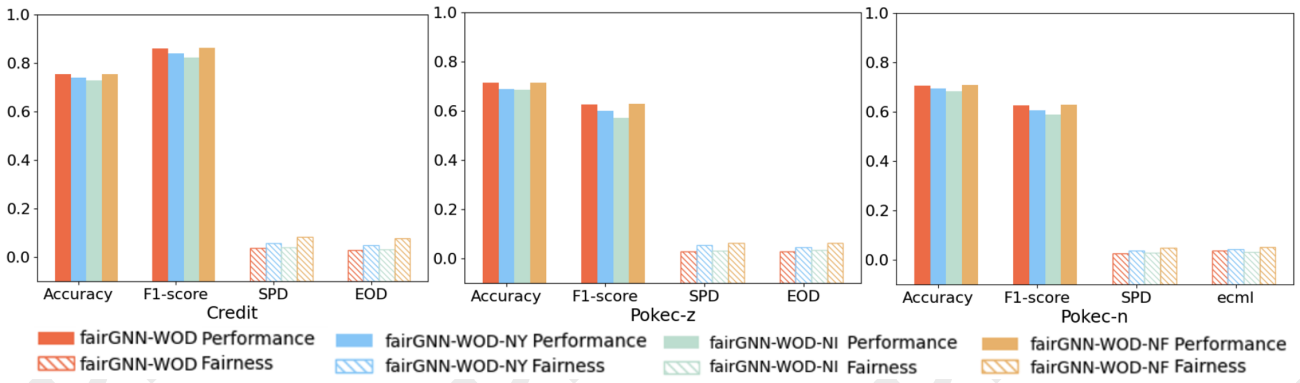

Figure 3: Ablation study results for fairGNN-WOD, fairGNN-WOD-NY, fairGNN-WOD-NI and fairGNN-WOD-NF.

nately, fairGNN-WOD disentangles node representations and applies a more targeted fairness correction. This approach prevents the loss of task-relevant demographic information, ensuring that the model remains both fair and useful.

Ablation Studies. We conducted ablation studies to understand the contributions of individual components in the fairGNN-WOD model. Specifically, we first explored the impact of the downstream GNN by excluding labeled information interference during demographic inference (*i.e.*, fairGNN-WOD-NY variant). The results on the three datasets are illustrated in Figure 3. We observed a decrease in fairness for the fairGNN-WOD-NY variant. This reduction in fairness is attributed to the model's failure to exclude demographic-irrelevant information during demographic inference, which compromised the quality of the demographic proxy and adversely affected downstream bias mitigation. Further, we examined the effects of removing disentanglement by introducing the fairGNN-WOD-NI variant that imposes the fairness constraint on the entire node representation by setting $N_c = 1$ and excluding both $\mathcal{L}_I$ and $\mathcal{L}_D$. The results showed a decline in prediction performance, indicating that the direct application of fairness constraints across the entire node representation space without disentanglement, inevitably removes some task-related information. The impact on fairness is not obvious. Lastly, to evaluate the effectiveness of our fairness constraints, we tested the fairGNN-WOD-NF variant, which drops the $\mathcal{L}_F$ penalty (*i.e.*, setting $\beta = 0$) in the objective function. The comparative analysis with the original fairGNN-WOD setup revealed a significant drop in fairness, underscoring the essential role that the fairness constraint plays. There is minimal impact on the prediction performance. In summary, the ablation studies affirm the importance of each component in the fairGNN-WOD.

Parameters Sensitivity Analysis. We examine the sensitivity of fairGNN-WOD by adjusting the parameters α and β across the values $\{1e^{-3}, 1e^{-2}, 1e^{-1}, 1e^{0}, 1e^{1}, 1e^{2}, 1e^{3}\}$. The results on the Credit dataset are presented in Figure 4. An increase in α and β tends to improve model fairness but may lead to a reduction in predictive performance. This can be attributed to the increase in these parameters, which strengthens the model's ability to disentangle node representation accurately and mitigate their correlation with demographic information. Hence, this diminishes the influence of demographics on node representation, thereby advancing model fairness.

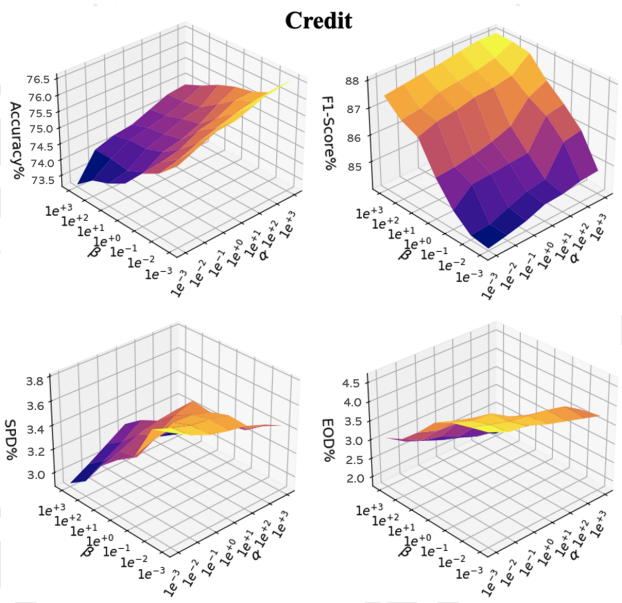

Figure 4: Exploring hyperparameters study results.

6 Conclusion

Given the observed gap between the prevailing real-world applications and the assumption of demographic information availability of existing AI fairness methods, this paper made an initial investigation into achieving graph fairness without complete demographic information. In addition, this work also took a step further to explore how to improve fairness while also preserving task-related information from demographics-related information. The proposed algorithms can achieve graph fairness across scenarios with both complete and incomplete demographic information and are readily extensible to existing fair GNN frameworks. Experiments on three real-world datasets demonstrate that fairGNN-WOD outperforms all baseline methods in terms of both fairness and utility performance.

Acknowledgements

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

References

[Chai *et al.*, 2022] Junyi Chai, Taeuk Jang, and Xiaoqian Wang. Fairness without demographics through knowledge distillation. *Advances in Neural Information Processing Systems*, 35:19152–19164, 2022.

[Chen *et al.*, 2019] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, 2019.

[Dai and Wang, 2021] Enyan Dai and Suhang Wang. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 680–688, 2021.

[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.

[Friedmann and Efrat-Treister, 2023] Enav Friedmann and Dorit Efrat-Treister. Gender bias in stem hiring: implicit in-group gender favoritism among men managers. *Gender & Society*, 37(1):32–64, 2023.

[Gebelein, 1941] Hans Gebelein. Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379, 1941.

[Grari *et al.*,] Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. Fairness without the sensitive attribute via causal variational autoencoder.

[Grari *et al.*, 2021] Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. Fairness without the sensitive attribute via causal variational autoencoder. *arXiv preprint arXiv:2109.04999*, 2021.

[Gretton *et al.*, 2006] A Gretton, K Borgwardt, M Rasch, B Schölkopf, and A Smola. A kernel method for the two-sample-problem advances in neural information processing systems. *MIT Press*, 2006.

[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.

[Kallus *et al.*, 2022] Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 68(3):1959–1981, 2022.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

[Lahoti *et al.*, 2020] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

[Ling *et al.*, 2023] Hongyi Ling, Zhimeng Jiang, Youzhi Luo, Shuiwang Ji, and Na Zou. Learning fair graph representations via automated data augmentations. In *International Conference on Learning Representations*, 2023.

[Liu *et al.*, 2023] David Liu, Virginie Do, Nicolas Usunier, and Maximilian Nickel. Group fairness without demographics using social networks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1432–1449, 2023.

[Liu *et al.*, 2024] Ping Liu, Haichao Wei, Xiaochen Hou, Jianqiang Shen, Shihai He, Kay Qianqi Shen, Zhujun Chen, Fedor Borisyuk, Daniel Hewlett, Liang Wu, et al. Linksage: Optimizing job matching using graph neural networks. *arXiv preprint arXiv:2402.13430*, 2024.

[Louizos *et al.*, 2017] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

[Ma *et al.*, 2019] Jianxin Ma, Peng Cui, Kun Kuang, Xin Wang, and Wenwu Zhu. Disentangled graph convolutional networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4212–4221. PMLR, 09–15 Jun 2019.

[Ni *et al.*, 2024] Hongliang Ni, Lei Han, Tong Chen, Shazia Sadiq, and Gianluca Demartini. Fairness without sensitive attributes via knowledge sharing. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1897–1906, 2024.

[Peng *et al.*, 2016] Sancheng Peng, Guojun Wang, and Dongqing Xie. Social influence analysis in social networking big data: Opportunities and challenges. *IEEE network*, 31(1):11–17, 2016.

[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, 2024] Zichong Wang and Wenbin Zhang. Group fairness with individual and censorship constraints. In *27th European Conference on Artificial Intelligence*, 2024.

[Wang and Zhang, 2025] Zichong Wang and Wenbin Zhang. Fdgen: A fairness-aware graph generation model. In *Proceedings of the 42nd International Conference on Machine Learning*. PMLR, 2025.

[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, Nripsuta Saxena, Tongjia Yu, Sneha Karki, Tyler Zetty, Israat Haque, Shan Zhou, Dukka Kc, Ian Stockwell, Albert Bifet, et al. Preventing discriminatory decision-making in evolving data streams. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023.

[Wang *et al.*, 2023c] Zichong Wang, Charles Wallace, Albert Bifet, Xin Yao, and Wenbin Zhang. Fg^2an: Fairness-aware graph generative adversarial networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 259–275. Springer Nature Switzerland, 2023.

[Wang *et al.*, 2024a] Zichong Wang, Zhibo Chu, Ronald Blanco, Zhong Chen, Shu-Ching Chen, and Wenbin Zhang. Advancing graph counterfactual fairness through fair representation learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 40–58. Springer Nature Switzerland, 2024.

[Wang *et al.*, 2024b] Zichong Wang, Jocelyn Dzuong, Xiaoyong Yuan, Zhong Chen, Yanzhao Wu, Xin Yao, and Wenbin Zhang. Individual fairness with group awareness under uncertainty. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 89–106. Springer Nature Switzerland, 2024.

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

[Wang *et al.*, 2025a] 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.*, 2025b] 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*, 2025.

[Wang *et al.*, 2025c] Zichong Wang, Anqi Wu, Nuno Moniz, Shu Hu, Bart Knijnenburg, Qingquan Zhu, and Wenbin Zhang. Towards fairness with limited demographics via disentangled learning. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*, 2025.

[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*, 32(1):4–24, 2020.

[Wu *et al.*, 2021] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. Self-supervised graph learning for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 726–735, 2021.

[Xu *et al.*, 2018] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[Yan *et al.*, 2020] Shen Yan, Hsien-te Kao, and Emilio Ferrara. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1715–1724, 2020.

[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 and Weiss, 2022] Wenbin Zhang and Jeremy C Weiss. Longitudinal fairness with censorship. In *proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 12235–12243, 2022.

[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*, volume 37, pages 14611–14619, 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*, 46(1):e12212, 2025.

[Zhang, 2024] Wenbin Zhang. Ai fairness in practice: Paradigm, challenges, and prospects. *Ai Magazine*, 45(3):386–395, 2024.

[Zhao *et al.*, 2022a] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. Towards fair classifiers without sensitive attributes: Exploring biases in related features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 2022.

[Zhao *et al.*, 2022b] 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.