Fair Graph Learning with Limited Sensitive Attribute Information

Zichong Wang[1], Jie Yang[2], Jun Zhuang[3], Puqing Jiang[4], Mingzhe Chen[5], Ye Hu[5], Wenbin Zhang[1]*

[1] Florida International University, Miami, United States
[2] University of Wollongong, Wollongong, Australia
[3] Boise State University, Boise, United States
[4] University of Pittsburgh, Pittsburgh, United States
[5] University of Miami, Miami, United States

Abstract

Graph neural networks (GNNs) excel at modeling graph-structured data but often inherit and amplify biases, leading to substantial efforts in developing fair GNNs. However, most existing approaches assume full access to sensitive attribute information, which is often impractical in real-world scenarios due to privacy concerns or risks of discrimination. To address this limitation, this paper focuses on graph fairness with limited sensitive attribute information, ensuring applicability to real-world contexts where current methods fall short. Specifically, we introduce an innovative fairness optimization strategy, propose a novel framework named FGLISA, and provide a theoretical perspective linking limited sensitive attribute information access to fairness objectives, thus enabling fair graph learning in real-world applications with limited sensitive attribute information. Experiments on diverse real-world datasets and tasks validate the effectiveness of our approach in achieving both fairness and predictive performance.

Introduction

Graph-structured data are ubiquitous in real-world applications, such as social networks (Kumar et al. 2022), knowledge graphs (Li et al. 2022), and recommendation (Wang et al. 2025a). The success of deep learning has led to significant advances in graph neural networks (GNNs) for processing graph-structured data (Zhou et al. 2020). However, GNNs often inherit societal biases, particularly those related to *sensitive attributes* (*e.g.*, race or gender), and may amplify these biases (Ma et al. 2022). This limitation hinders their adoption in high-stakes applications such as job applicant ranking (Mehrabi et al. 2021) and crime rate prediction (Jin et al. 2020). To address this concern, numerous approaches for fair GNNs have been developed (Zhang et al. 2025), with most relying on complete sensitive information to guide fair model training (Wang et al. 2025f).

This requirement for complete sensitive information, however, presents a significant challenge in practice, as sensitive information is often incomplete in real-world scenarios due to privacy concerns or fear of discrimination (Chai, Jang, and Wang 2022; Wang et al. 2025b,d). For example, a tech company might utilize GNNs in its hiring process for software engineering roles by analyzing connections within professional networks to assess collaboration potential with internal teams (Liu et al. 2024). Yet, applicants from underrepresented gender groups (*e.g.*, female) may opt not to disclose their gender, especially in male-dominated fields like software engineering (Friedmann and Efrat-Treister 2023). This scenario, where only a limited subset of nodes has known sensitive information, creates a critical gap between theoretical fairness approaches and real-world applications by significantly restricting the applicability of conventional fairness methods.

To this end, several methods have begun exploring fair models with limited sensitive information, either by inferring sensitive information proxies or using strategies like Max-Min fairness (Grari, Lamprier, and Detyniecki 2021; Hashimoto et al. 2018). However, most of them focus on non-graph data and cannot be directly applied to graph-structured scenarios. In addition, they often assume an extreme case where sensitive information is completely unavailable (Ashurst and Weller 2023; Wang et al. 2025c,e). This assumption can diminish the accuracy of identifying sensitive attributes by overlooking the limited yet valuable sensitive information that is available. For example, in the widely used Facebook dataset, a fairness benchmark for bias mitigation in friend recommendation models, 14% of teen users have made their full profiles public (Madden et al. 2013), revealing sensitive information that warrants careful consideration.

To fill the gap between theoretical fairness approaches and real-world applications, it is of practical importance to enable fair graph learning in real-world scenarios with limited sensitive attribute information, a highly underexplored area with unique challenges: **i) Efficient Limited Sensitive Information Utilization:** When sensitive information is only partially available, effectively utilizing this limited information becomes challenging. Directly using this limited information without careful consideration may fail to account for irrelevant information that exhibits similar patterns to sensitive attributes, leading to unreliable identification of subgroups and ineffective fairness enhancement. Therefore, the key challenge lies in how to effectively leverage limited sensitive information to establish reliable patterns that can distinguish genuine sensitive attribute information from irrelevant ones. **ii) Mitigate Discrimination with Incomplete Sensitive Information:** When sensitive information is partially missing, measuring and mitigating model discrimi-

*Corresponding author.

nation becomes challenging. Fairness optimization requires complete sensitive information from all instances to calculate group-level statistics, making it impossible to directly evaluate disparities in model predictions across different demographic groups. Therefore, it becomes unclear how to define and measure model bias when group membership information is incomplete. **iii) Prevention of Manipulation in Joint Optimization to Enhance Fairness:** When simultaneously optimizing for sensitive information inference and unbiased prediction, the model might manipulate the inferred sensitive attribute to artificially reduce fairness loss. For instance, it could misclassify samples from one demographic group to another to minimize subgroup disparities.

To address these challenges, this paper introduces a novel problem of fair graph learning with limited sensitive attribute information and proposes a framework, **F**air **G**raph **L**earning with l**I**mited **S**ensitive **A**ttribute (**FGLISA**), which is designed to achieve fair graph learning with limited sensitive information. *To the best of our knowledge, this is the first work that efficiently utilizes limited sensitive information while making the process independent of it, thus enabling fair graph learning with limited sensitive information.* Specifically, we consider a practical scenario where sensitive information is partially missing and propose a disentanglement-based approach that serves two purposes. First, it enables the effective utilization of limited sensitive information by disentangling sensitive attribute-related information from irrelevant ones into two latent variables through variational inference, thereby establishing reliable guidance for identifying missing sensitive information. Second, armed with the identified sensitive information, it ensures model fairness by estimating and minimizing the correlation between sensitive information and predictions, thus achieving both accurate sensitive attribute identification and fair prediction simultaneously. Our contributions are outlined below:

- **Problem Formulation.** We investigate the critical yet underexplored problem of measuring and mitigating bias in graph learning with limited sensitive information. Through theoretical analysis, we demonstrate how the causal effect of sensitive attributes on other variables influences model bias, and identify the corresponding challenges in addressing this problem when sensitive information is partially available.
- **Framework Design and Generalization.** We propose FGLISA, a fair graph learning framework for improving fairness under limited sensitive attribute. To our best knowledge, this is the first work to improve fairness in graph-structured data with limited sensitive attribute via disentanglement.
- **Experimental Evaluations.** We conduct extensive experiments to evaluate FGLISA by comparing it with six state-of-the-art methods across three real-world datasets, achieving a superior performance in both utility and fairness metrics.

Related Work

Fairness without Sensitive Attribute. Due to privacy concerns or risks of discrimination, methods for achieving fairness without sensitive information have emerged, primarily focusing on simplified non-graph domains, which can be broadly categorized into two categories (Ashurst and Weller 2023): i) Max-Min fairness: Methods in this category (Lahoti et al. 2020; Hashimoto et al. 2018; Yan, Kao, and Ferrara 2020) follow the principle of Rawlsian Max-Min fairness (Rawls 1971), aiming to minimize the risk for the most disadvantaged subgroup. For example, Chai *et al.* (Chai, Jang, and Wang 2022) proposed a knowledge distillation framework that improves fairness by optimizing for the worst-case group utility without accessing sensitive information. However, these approaches cannot guarantee that the algorithmically identified subgroups correspond to the actual sensitive groups of interest, potentially undermining the fairness objectives (Zhang et al. 2023). ii) Proxy sensitive attribute: This line of work leverages available features to predict missing sensitive information (Grari, Lamprier, and Detyniecki 2022). However, when inferring sensitive attribute proxies, these methods often fail to filter out irrelevant information adequately. For example, when inferring gender information, other attributes like race may introduce noise if not properly accounted for, leading to unreliable proxies that undermine fairness mitigation (Chai, Jang, and Wang 2022). Furthermore, both categories of approaches face two inherent limitations: i) They rely on overly stringent fairness scenarios, assuming the complete absence of sensitive information while failing to leverage limited but crucial sensitive information, and ii) They overlook the risk of manipulation during the optimization process, where models may artificially enhance fairness metrics by modifying group assignments or sensitive attribute predictions rather than genuinely mitigating bias.

Fairness in Graph Learning. Fairness on graphs has gained significant attention due to the ubiquity of graph data and the critical need to uphold fairness (Li et al. 2024; Wang et al. 2024a; Wang and Zhang 2025). The core idea behind most of the existing fair GNNs is to ensure that algorithmic decisions do not discriminate against or favor specific subgroups explicitly defined by sensitive attributes (Wang et al. 2023b; Wang, Yin, and Zhang 2025; Wang et al. 2023a; Zhang 2024b). For instance, Graphhair (Ling et al. 2023) generates fair graph data through adversarial learning to fool discriminators. While these methods have shown efficacy in enhancing fairness, they all rely on completely sensitive information. To address this limitation, FairGNN (Dai and Wang 2021), the only existing work known to us that considers graph fairness without complete sensitive information, employs the sensitive attribute estimator to predict the missing sensitive attribute while improving fairness via adversarial learning. However, this approach relies on a limited subset of known sensitive attribute data to train the model and predict missing sensitive attribute labels. This dependence assumes that the limited data is unbiased, which is often not the case. For instance, in male-dominated fields like software engineering, men may be more likely to disclose their sensitive information, leading to disparities in representation across demographic groups and, consequently, unreliable sensitive attribute label predictions.

Distinct from prior research, this work tackles these drawbacks comprehensively through three key advancements.

First, by estimating the effect of sensitive information on other variables, FGLISA enables reliable proxy generation for missing sensitive information by extracting sensitive attribute patterns in underrepresented demographic groups (*e.g.*, female), leveraging the consistency of causal effects (*i.e.*, features influenced by sensitive attributes remain similar across groups, such as how height is influenced by gender). Second, through a proposed disentangled learning structure, FGLISA separates latent variables into sensitive attribute-related information and other features, thus excluding irrelevant information during missing sensitive attribute inference. Furthermore, FGLISA takes one more step further to consider potential manipulation during co-optimization (*i.e.*, concurrent inference of missing sensitive attributes and node label prediction) to ensure genuine bias mitigation.

Notation

Given a graph by $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$, where $\mathcal{V}$ is the set of n nodes and $\mathcal{E}$ is the set of m edges. The adjacency matrix is denoted by $\mathbf{A} \in \{0,1\}^{n\times n}$, with $\mathbf{A}_{i,j} = 1$ indicating the presence of an edge between nodes v_i and v_j, and $\mathbf{A}_{i,j} = 0$ otherwise. Each node $v_i \in \mathcal{V}$ is associated with a d-dimensional feature vector $\mathbf{x}_i$, and the collection of all node features forms the feature matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$. Without loss of generality, we consider an undirected and unweighted graph where both node labels and sensitive attributes are binary variables. Let $y_i \in \{0,1\}$ denote the true label of node v_i, where $y_i = 1$ represents the granted class and $y_i = 0$ represents the rejected class; the predicted label is denoted by $\hat{y}_i$. The sensitive attribute of node v_i is represented by $s_i \in \{0,1\}$, and only a small set of nodes $\mathcal{V}_S \in \mathcal{V}$ has are provided with the sensitive attribute, collected into the vector $\mathbf{S} \in \{0,1\}^{|\mathcal{V}_S|}$. In addition, we define the favored group as $S_f = \{v_i \in \mathcal{V} \mid s_i = 1\}$ (*e.g.*, male) and the deprived group as $S_d = \{v_i \in \mathcal{V} \mid s_i = 0\}$ (*e.g.*, female).

FGLISA: Framework and Theories

In this section, the underlying causal structure is first introduced. Building on this foundation, a disentangled variational autoencoder is developed to separate sensitive attribute-related information. We then outline the method for identifying missing sensitive information using the learned representations. Finally, the fairness optimization strategy for handling partially observed sensitive information is presented.

Causal Structure of the Sensitive Disentangled Causal Variational Autoencoder

The proposed causal graph, shown in Figure 1, defines the direct causal relationships originating from the sensitive variable S to various observable variables, where each directed edge represents a direct causal influence between the connected variables. In addition, the graph illustrates how S specifically influences certain node features in X, enabling the partition of X into X_S, which are directly influenced by S, and $X_{\overline{S}}$, which are not. For example, if we treat "Male" as the sensitive attribute (*i.e.*, representing gender), then X_S may include attributes like height and mustache, while $X_{\overline{S}}$ could include attributes such as race or weather-related features that are not directly caused by gender. Furthermore, S also affects the graph structure A. For instance, in social networks, individuals may tend to connect with others of the same gender. It is important to note that while S influences X_S and A, it should not exert a direct causal effect on the ground-truth label Y. For example, a hiring decision should not be directly influenced by the applicant's gender. Moreover, both the node features X and the graph structure A contain critical information that affects Y. For example, in a hiring scenario, the applicant's qualifications (X) and their professional network (A) could both shape the hiring outcome (Y). Last, A can affect the node features X. For instance, if an applicant's professional network predominantly consists of individuals skilled in a specific technology, this network structure may increase the likelihood that the applicant has acquired expertise in that domain. This reflects how the graph structure can indirectly impact node-level attributes through connections.

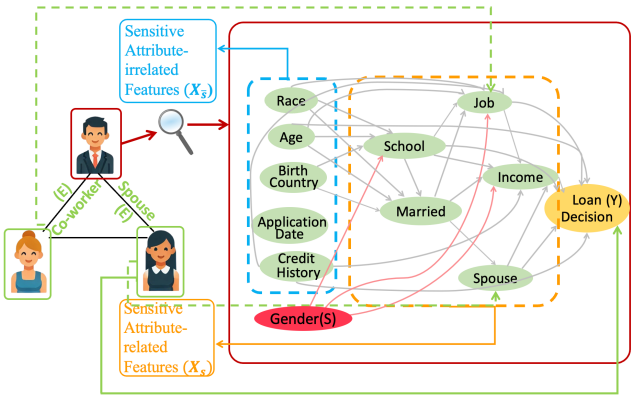

Figure 1: The causal graph of FGLISA.

Disentangled Causal Effect Variational Autoencoder

Building upon the proposed causal model, we present the first component of the FGLISA framework, which aims to learn two distinct latent representations, Z_S and $Z_{\overline{S}}$. Here, Z_S serves as a proxy for the missing sensitive attribute S, capturing information that is directly related to S, while $Z_{\overline{S}}$ encodes aspects unrelated to S. This disentanglement helps ensure that the inference of S remains unaffected by extraneous factors. To implement this approach, we leverage variational Bayesian inference guided by the predefined causal graph in Figure 1. Specifically, Z_S and $Z_{\overline{S}}$ are modeled as two latent variables within a corresponding Bayesian network, defining the structure of our Disentangled Causal Effect Variational Autoencoder, as depicted in Figure 2.

Our neural framework encompasses two inference networks, $q_\phi(Z_S \mid X)$ and $q_\theta(Z_{\overline{S}} \mid X)$, which approximate the posterior distributions over Z_S and $Z_{\overline{S}}$, respectively. In addition, we incorporate a mechanism to encourage disentanglement by measuring and minimizing the correlation between Z_S and $Z_{\overline{S}}$. Specifically, we adopt an adversarial training strategy that builds upon Definition 1, which we extend from the Hirschfeld-Gebelein-Rényi (HGR) correla-

tion (Kamath and Anantharam 2012), denoted as the Z_S-$Z_{\overline{S}}$ correlation. As formalized in Definition 1, this measure captures the maximal degree to which two random variables can be related via measurable functions.

Definition 1 (Z_S-$Z_{\overline{S}}$ correlation). Given any two jointly distributed random latent variables Z_S and $Z_{\overline{S}}$. The HGR maximal correlation between them is defined as:

$$\mathrm{HGR}(Z_S, Z_{\overline{S}}) = \sup_{f_1, f_2} \rho\big(f_1(Z_S), f_2(Z_{\overline{S}})\big) = \sup_{f_1, f_2} \frac{\mathbb{E}\big(f_1(Z_S)\, f_2(Z_{\overline{S}})\big)}{\sqrt{\mathbb{E}\big({f_1}^2(Z_S)\big)\, \mathbb{E}\big({f_2}^2(Z_{\overline{S}})\big)}} \tag{1}$$

where ρ denotes the Pearson correlation coefficient, and f_1, f_2 are measurable functions with finite positive variances. To ensure consistency, we normalize them such that $\mathbb{E}[f_1(Z_S)] = \mathbb{E}[f_2(Z_{\overline{S}})] = 0$ and $\mathbb{E}[f_1^2(Z_S)] = \mathbb{E}[f_2^2(Z_{\overline{S}})] = 1$. If Z_S and $Z_{\overline{S}}$ are independent, the HGR maximal correlation is zero; if there is a deterministic relationship, it is one.

We use this correlation measure in a min-max optimization setting to disentangle Z_S and $Z_{\overline{S}}$ and to infer the missing sensitive attribute. Concretely, we train adversarial subnetworks parameterized by ω_{f_1} and ω_{f_2} to increase the estimated dependence between Z_S and $Z_{\overline{S}}$ via gradient ascent, while the primary model seeks to minimize this dependence. As a result, Z_S and $Z_{\overline{S}}$ become maximally independent, ensuring that Z_S focuses on sensitive information and $Z_{\overline{S}}$ remains sensitive-irrelevant information.

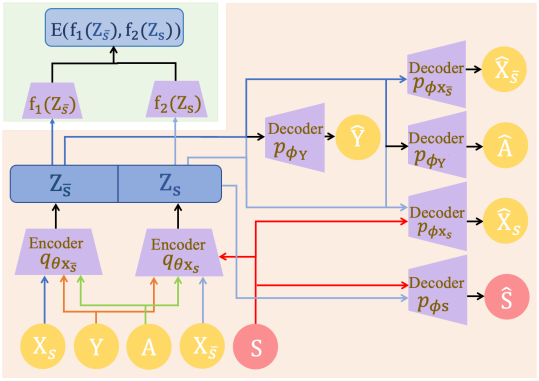

Figure 2: The overview of FGLISA.

Missing Sensitive Attribute Information Identification

In this section, we describe how to identify the missing sensitive attribute information from observed graph data (*i.e.*, X_S, $X_{\overline{S}}$, A, and Y) and partially available sensitive attribute S. Our goal is to develop a systematic approach that leverages prior knowledge from a causal model to infer missing sensitive information. We begin by stating a foundational assumption of the proposed framework:

Assumption 1 Sensitive attributes are not caused by any other variables (*e.g.*, school or height); rather, they only act as parent variables. This is a common perspective in fairness research (Mehrabi et al. 2021; Ma et al. 2022; Zhu et al. 2024), as sensitive attributes (*e.g.*, race or gender) are usually considered inherent to individuals.

A straightforward way to perform inference is to posit a nondeterministic structural model with specific distributions for all causal relationships and then estimate the latent space. Specific to this work, leveraging Assumption 1 and following Louizos *et al.* (Louizos et al. 2017), we can recover missing sensitive information if we can recover the joint distribution of $P(S, A, X_S, X_{\overline{S}}, Y)$. Note that we model exogenous variables in the causal graph differently from the existing counterfactual fairness works. Counterfactual fairness ensures individual-level fairness through the concept of the "do" operator and by intervening on sensitive attributes in the causal relationship (*i.e.*, flipping sensitive attributes $S \to S'$). In contrast, we aim to capture stochastic exogenous variables by generating a sensitive attribute proxy distribution for each individual. As mentioned earlier, using our Bayesian network, we can model the joint probability $P(S, A, X_S, X_{\overline{S}}, Y)$ using the following factorization:

$$\begin{aligned} P(S, A, X_S, X_{\overline{S}}, Y) = {} & P(Z_S)\, P(Z_{\overline{S}})\, P(S \mid Z_S) \\ & P(X_S \mid A, S, Z_S)\, P(A \mid S, Z_{\overline{S}}, Z_S) \\ & P(X_{\overline{S}} \mid A, Z_{\overline{S}})\, P(Y \mid A, Z_{\overline{S}}, Z_S) \end{aligned} \tag{2}$$

where $P(S \mid Z_S)$ corresponds to the generation function $\mathcal{F}_S$ that produces Z_S, a latent proxy for the missing sensitive attribute S. The priors $P(Z_S)$ and $P(Z_{\overline{S}})$ follow Gaussian distributions. The terms $P(X_S|A, S, Z_S)$ and $P(X_{\overline{S}}|A, Z_{\overline{S}})$ are decoders for node features related and unrelated to the sensitive attribute, respectively. The term $P(A|S)$ describes the generative process of the graph structure, and $P(Y|A, Z_{\overline{S}}, Z_S)$ predicts the node labels. Since S, A, X, and Y differ in their nature, we assign appropriate distributions. We assume that if X is continuous, it follows a Gaussian distribution. If Y is binary, it follows a Bernoulli distribution. Finally, the adjacency matrix A is follows the Bernoulli distribution. Building on this, the decoder distribution $p_\phi(X_S, X_{\overline{S}}, S, Y \mid U_S, U_{\overline{S}})$ can be factorized accordingly, as shown in detail below:

$$\begin{aligned} p_\phi(S, X_S, X_{\overline{S}}, A, Y) = {} & p(Z_S)\, p(Z_{\overline{S}})\, p_{\phi_S}(S \mid Z_S) \\ & p_{\phi_{X_S}}(X_S \mid Z_S, S, A)\, p_{\phi_A}(A \mid Z_S, Z_{\overline{S}}, S) \\ & p_{\phi_{X_{\overline{S}}}}(X_{\overline{S}} \mid Z_{\overline{S}}, A)\, p_{\phi_Y}(Y \mid Z_S, Z_{\overline{S}}, A) \end{aligned} \tag{3}$$

Following the above factorization, we define the approximate posterior (*i.e.*, $q_\phi(Z_S, Z_{\overline{S}} \mid X_S, X_{\overline{S}}, A, S, Y)$) for the latent variables as:

$$\begin{aligned} q_\phi(Z_S, Z_{\overline{S}} \mid X_S, X_{\overline{S}}, A, S, Y) = {} & q_\phi(Z_S \mid X_S, A, S, Y) \\ & q_\phi(Z_{\overline{S}} \mid X_{\overline{S}}, A, S, Y) \end{aligned} \tag{4}$$

We employ variational inference, parameterized by neural networks, to learn the parameters of this model. The key is to maximize the variational lower bound (ELBO) (Kingma and Welling 2013) on the log probability of the observed data $(S, A, X_S, X_{\overline{S}}, Y)$. Specifically, we introduce a variational distribution $Q(S \mid A, X_S, X_{\overline{S}}, Y)$ to approximate the

intractable posterior $q_\phi(Z_S, Z_{\overline{S}} \mid X_S, X_{\overline{S}}, A, S, Y)$. Then, to learn the parameters of FGLISA, we maximize the variational lower bound of the log probability of the observed data distribution, detailed as follows:

$$\log P(S, A, X_S, X_{\overline{S}}, Y) \geq \mathbb{E}_{q_\phi(Z_S, Z_{\overline{S}} \mid S, A, X_S, X_{\overline{S}}, Y)} \left[\log \frac{P(S, X_S, X_{\overline{S}}, A, Y, Z_S, Z_{\overline{S}})}{q_\phi(Z_S, Z_{\overline{S}} \mid S, A, X_S, X_{\overline{S}}, Y)} \right] \tag{5}$$

The values of $\log P(S, X_S, X_{\overline{S}}, A, Y | U_S, U_{\overline{S}})$ correlates positively with the reality of observed data. The $P(X_S, X_{\overline{S}}, A, Y, S, Z_S, Z_{\overline{S}})$ represents the joint distribution between the observed data, while $q_\phi(Z_S, Z_{\overline{S}} | S, X_S, X_{\overline{S}}, Y)$ denote the posterior distribution of the sensitive attribute. Furthermore, to approximate the intractable posterior distribution of these latent variables, we introduce a variational distribution $Q(Z_S | X_S, A, S, Y)$ and $Q(Z_{\overline{S}} | X_{\overline{S}}, S, Y)$, which uses a parametric family of distributions to approximate the true posterior distribution $P(Z_S | X_S, A, S, Y)$ and $P(Z_{\overline{S}} | X_{\overline{S}}, A, S, Y)$.

However, simply maximizing the ELBO does not ensure the independence between Z_S and $Z_{\overline{S}}$. Such independence is critical for accurately separating sensitive-related information from other factors. To address this, we include an additional independence constraint loss into our model based on Definition 1 (the Z_S-$Z_{\overline{S}}$ correlation), which measures the dependence between the latent variables. Building on this, the ELBO can be reformulated as follows:

$$\begin{aligned} \log P(S, A, X_S, X_{\overline{S}}, Y) \geq\ & \mathbb{E}_{q_\phi} \Big[\log P(S|Z_S) + \log P(A|Z_S, Z_{\overline{S}}, S) \\ & + \log P(X_S|Z_S, A, S) + \log P(X_{\overline{S}}|Z_{\overline{S}}, A) \\ & + \log P(Y|Z_S, Z_{\overline{S}}, A) \\ & - Q(Z_S|S, A, X_S) - Q(Z_{\overline{S}}|S, A, X_{\overline{S}}) \\ & + \log P(Z_S) + \log P(Z_{\overline{S}}) \Big] \\ & + \lambda\, \mathrm{HGR}(Z_S, Z_{\overline{S}}) \end{aligned} \tag{6}$$

where λ denotes the hyperparameter that balances the contribution of the penalization term. To optimize this updated ELBO, we adopt a min-max structure to accurately infer missing sensitive attributes while discouraging any label-related entanglement. Specifically, during the maximization phase, we use gradient ascent to estimate the HGR maximal correlation between Z_S and $Z_{\overline{S}}$ by optimizing the functions f_1 and f_2 in Equation 1. These functions, parameterized by neural networks, are trained to maximize the Pearson correlation $\rho(f_1(Z_S), f_2(Z_{\overline{S}}))$. In the minimization phase, we update the main model parameters, which include the ELBO and the penalty term $\lambda \cdot \mathrm{HGR}(Z_S, Y)$. This iterative optimization reduces the dependency between Z_S and $Z_{\overline{S}}$, promoting the independence required for accurately inferring the S.

Fairness Optimization with Limited Sensitive Attribute Information

Armed with the identity proxy for missing sensitive information, the goal is now to use it to train a fair predictive function. The core idea behind this is to remove the sensitive attribute-related information, thereby enforcing a predictive function to make decisions independent of the sensitive attribute. One straightforward approach is to remove the latent variable Z_S and only use $Z_{\overline{S}}$ to predict the node label. However, this may inevitably remove some label-related information due to its correlation with the sensitive attribute. As a result, this leads to performance degradation of FGLISA in downstream tasks. To this end, we aim to keep Z_S and mitigate the bias via an adversarial penalization during the training phase. Specifically, we propose to find a mapping $p_{\phi_Y}(Y | Z_S, Z_{\overline{S}}, A)$ which both minimizes the deviation from the ground truth label Y and does not imply too much dependency with the sensitive attribute. To achieve this, we further mitigate the dependency between $p_{\phi_Y}(Y | Z_S, Z_{\overline{S}}, A)$ and $\hat{S}$, where $\hat{S}$ is generated as mentioned above from the posterior distribution $p_{\phi_S}(S | Z_S)$. Building on this, the proposed optimization strategy is to minimize the prediction loss while maximizing the independence of $p_{\phi_S}(S | Z_S)$ and $\hat{S}$. To this end, we proposed a broad fairness loss, which extends from statistical parity into missing sensitive attribute situations, as shown in Definition 2.

Definition 2 (Broad Fairness Loss). Given a classifier p_{ϕ_Y} and sensitive attribute predictor p_{ϕ_S}, if the classifier's predictions have consistent relevance to different predicted sensitive attribute subgroups, then the classifier is considered unbiased. Mathematically, this is represented as:

$$\begin{aligned} \text{BFL} := \Bigg| & \sum_{v_i \in \mathcal{V}} f\Big(p_{\phi_Y}(Y = 1 \mid Z_S^i, Z_{\overline{S}}^i, A_i),\, p_{\phi_S}(S = 0 \mid Z_S^i)\Big) \\ & - \sum_{v_i \in \mathcal{V}} f\Big(p_{\phi_Y}(Y = 1 \mid Z_S^i, Z_{\overline{S}}^i, A_i),\, p_{\phi_S}(S = 1 \mid Z_S^i)\Big) \Bigg| \end{aligned} \tag{7}$$

where $f(\cdot)$ denotes the relevance estimate function (the Maximum Mean Discrepancy (MMD) (Gretton et al. 2006) in our experiments). MMD is a popular estimator for measuring the distribution discrepancy of latent variables. Therefore, given a classifier p_{ϕ_Y}, the MMD-based fairness loss measures the discrepancy between predictions for different sensitive attribute groups and is defined as follows:

$$\mathcal{L}_{\text{BFL}} = \frac{|D_{S_0,S} - D_{S_1,S}|}{D_{S_0,S_1}} \tag{8}$$

where $D_{S_0,S}$ and $D_{S_1,S}$ denote the dependency between model predictions p_{ϕ_Y} and sensitive attribute predictions for subgroups S_0 and S_1 respectively, and D_{S_0,S_1} represents the dependency between predictions of these two subgroups. Notably, D_{S_0,S_1} targets enlarging the similarity between every two subgroups defined by sensitive attributes obfuscated by $|D_{S_0,S} - D_{S_1,S}|$. Mathematically, these terms are computed as:

$$\begin{aligned} D_{S_0,S} =& \frac{1}{|\mathcal{V}|^2} \sum_{v_i \in \mathcal{V}} \sum_{v_j \in \mathcal{V}} k(p_{\phi_Y}(Y = 1 | Z_S^i, Z_{\overline{S}}^i, A_i), p_{\phi_S}(S = 0 | Z_S^j)) \\ & + \frac{1}{|S_0|^2} \sum_{v_i \in S_0} \sum_{v_j \in S_0} k(p_{\phi_Y}(Y = 1 | Z_S^i, Z_{\overline{S}}^i, A_i), p_{\phi_S}(S = 0 | Z_S^j)) \\ & - \frac{2}{|\mathcal{V}||S_0|} \sum_{v_i \in \mathcal{V}} \sum_{v_j \in S_0} k(p_{\phi_Y}(Y = 1 | Z_S^i, Z_{\overline{S}}^i, A_i), p_{\phi_S}(S = 0 | Z_S^j)) \end{aligned} \tag{9}$$

$D_{S_1,S}$, and D_{S_0,S_1} is defined similarly. In addition, $k(\cdot, \cdot)$ is a positive definite kernel function (*e.g.*, Gaussian kernel).

Dataset	Metrics \ Methods	GCN	FairGNN	KSMOTE	FairRF	FairGKD	FairAC	FGLISA
Credit	**Accuracy (↑)**	**0.781 ± 0.016**	0.687 ± 0.012	0.736 ± 0.009	0.735 ± 0.007	0.741 ± 0.017	0.697±0.022	0.758 ± 0.021
	F1-Score (↑)	**0.868 ± 0.023**	0.793 ± 0.043	0.812 ± 0.062	0.809± 0.022	0.819 ± 0.018	0.815±0.018	0.825 ± 0.018
	SPD(↓)	0.117 ± 0.013	0.123 ± 0.036	0.071 ± 0.003	0.067 ± 0.017	0.066 ± 0.014	0.064±0.009	**0.058 ± 0.013**
	EOD(↓)	0.096 ± 0.017	0.115 ± 0.042	0.055 ± 0.013	0.057± 0.018	0.050 ± 0.015	0.053±0.011	**0.047 ± 0.021**
Pokec-z	**Accuracy (↑)**	**0.699 ± 0.024**	0.689 ± 0.091	0.687±0.024	0.690 ± 0.014	0.694 ± 0.022	0.684 ± 0.021	0.695 ± 0.018
	F1-Score (↑)	0.622 ± 0.012	0.631 ± 0.033	0.611±0.018	0.617 ± 0.019	**0.637 ± 0.022**	0.633±0.012	0.631 ± 0.016
	SPD(↓)	0.075 ± 0.025	0.058 ± 0.022	0.037±0.017	0.032 ± 0.012	0.043 ± 0.014	0.030±0.011	**0.027 ± 0.010**
	EOD(↓)	0.062 ± 0.013	0.045 ± 0.029	0.039±0.010	0.034 ± 0.012	0.033 ± 0.011	0.031±0.021	**0.028 ± 0.006**
Pokec-n	**Accuracy (↑)**	0.689 ± 0.015	**0.726 ± 0.013**	0.669±0.013	0.673 ± 0.013	0.680 ± 0.017	0.675±0.021	0.673 ± 0.024
	F1-Score (↑)	0.631 ± 0.022	**0.637 ± 0.019**	0.611±0.018	0.616 ± 0.032	0.627 ± 0.032	0.631±0.028	0.622 ± 0.029
	SPD(↓)	0.084 ± 0.013	0.036 ± 0.012	0.061±0.005	0.056 ± 0.027	0.023 ± 0.011	0.020±0.014	**0.016 ± 0.013**
	EOD(↓)	0.078 ± 0.019	0.044 ± 0.020	0.066±0.013	0.061 ± 0.016	0.034 ± 0.006	0.033±0.021	**0.025 ± 0.005**

Table 1: Comparison results of FGLISA 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.

On the other hand, this strategy effectively handles the challenge of evaluating model bias when sensitive information is partially missing during training, which is a limitation of the existing fairness strategy. This requirement makes this strategy inapplicable in scenarios where sensitive attributes are partially missing, as they cannot assess fairness without the complete sensitive information. In contrast, our strategy overcomes this limitation by leveraging inferred sensitive information during the training, enabling continuous fairness evaluation throughout the sensitive information that is only partially available.

In addition, directly incorporating the broad fairness loss into the final loss function may lead to unintended manipulation of the inferred sensitive attribute. Specifically, when simultaneously optimizing for sensitive attribute inference and prediction fairness, the model might manipulate the inferred sensitive attribute to reduce the fairness loss. For example, it could misclassify a male positive instance as female to reduce the disparity between groups. This issue stems from the joint optimization of the inference model q_θ and predictor p_{ϕ_Y}. To address this challenge, we propose a two-stage training strategy with gradient isolation, *i.e.*, stop-gradient technique. First, we train the inference model q_θ to learn accurate sensitive attribute predictions. Then, during fairness optimization, we apply the stop-gradient operator to these predictions, treating them as fixed values. This prevents the fairness loss from influencing the quality of sensitive attribute inference. Furthermore, the prediction of Y not only affects predictive performance but also contributes to fairness metrics. Essentially, since the fairness loss depends on the predictions of Y, applying a stop gradient prevents the simultaneous optimization of both fairness and performance objectives. To address this, we introduce a separate semi-supervised classifier $p_{\phi'_Y}(Y|X, A)$ that estimates Y without considering fairness constraints. This classifier is trained using available labeled data and semi-supervised techniques for unlabeled instances. Building on this, the final loss function of FGLISA is defined as follows:

$$\mathcal{L}_{\text{total}} = -\mathcal{L}_{\text{ELBO}} + \lambda \cdot \text{HGR}(Z_S, Z_{\overline{S}}) + \omega \cdot \mathcal{L}_{\text{BFL}} \quad (10)$$

where ω is a hyperparameter that balances the contribution of fairness, and λ encourages further independence between Z_S and $Z_{\overline{S}}$. Through this design, we preserve necessary label-related information within Z_S while mitigating bias.

Experiment

Datasets and Baselines. We conduct experiments using three real-world datasets: Credit (Yeh and Lien 2009), Pokec-z and Pokec-n (Takac and Zabovsky 2012). FGLISA is compared with several baseline methods, including GCN (Kipf and Welling 2016), FairGNN (Dai and Wang 2021), KSMOTE (Yan, Kao, and Ferrara 2020), FairRF (Zhao et al. 2022), FairGKD (Zhu et al. 2024) and FairAC (Guo, Chu, and Li 2023). Detailed descriptions are provided in the Appendix.

Evaluation Metrics. We use accuracy and F1-score to evaluate the utility, with higher values indicating better performance. To evaluate fairness, we use two widely used fairness metrics, statistical parity (Dwork et al. 2012): $\Delta_{SP} = |P(\hat{y} = 1|s = 0) - P(\hat{y} = 1|s = 1)|$ and equal opportunity (Hardt, Price, and Srebro 2016): $\Delta_{EO} = |P(\hat{y} = 1|y = 1, s = 0) - P(\hat{y} = 1|y = 1, s = 1)|$, with values close to 0 indicate better fairness.

Experiment Result

Fairness and Utility. Table 1 presents a comparison between FGLISA and six baseline methods across utility and fairness metrics, revealing two key findings: i) FGLISA achieves superior fairness with limited sensitive attribute information. Across all three datasets, our method consistently outperforms baseline methods in fairness metrics. This improvement can be attributed to FGLISA's ability to effectively infer missing sensitive attributes proxies that highly correlate with true attributes, providing a solid foundation for bias mitigation. ii) FGLISA maintains a better balance between prediction accuracy and fairness. While traditional node classification methods like GCN achieve higher prediction accuracy at the cost of fairness, and existing fair node classification methods (FairGNN, KSMOTE, FairRF, FairGKD, FairAC) improve fairness at the expense of prediction performance, FGLISA outperforms state-of-the-art methods in both aspects. This superior performance stems from two factors: accurate inference of missing sensitive attributes reduces fairness cost and effective disentangled learning that captures

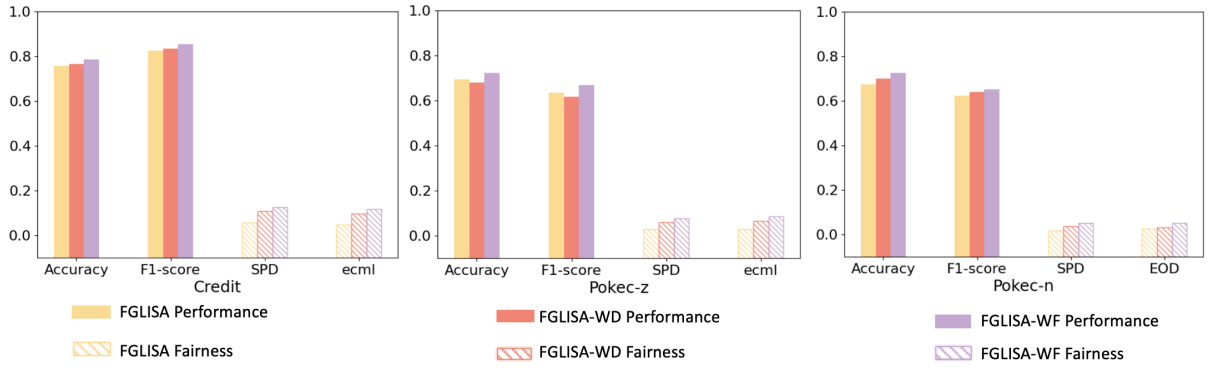

Figure 3: Ablation study results for FGLISA, FGLISA-WD, and FGLISA-WF.

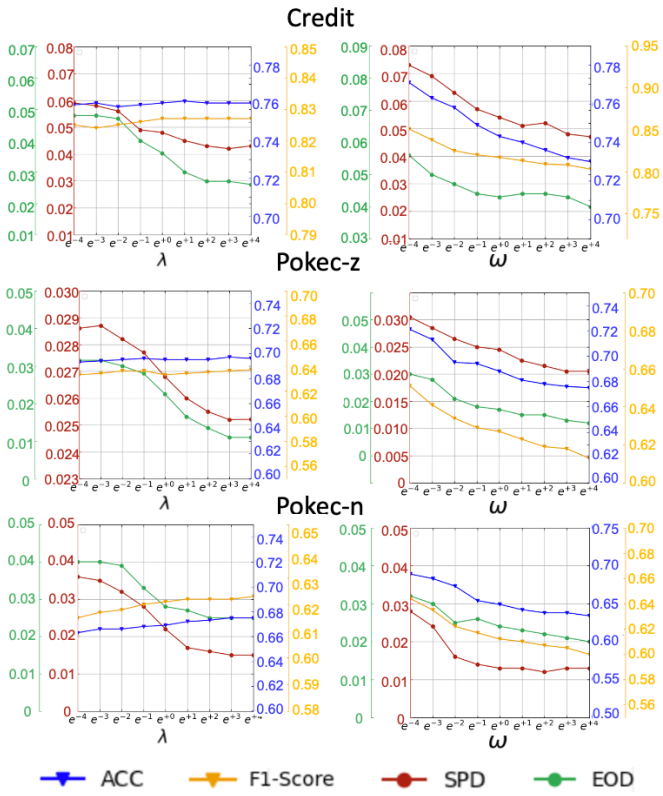

Figure 4: Study on parameter sensitivity analysis.

the underlying data structure. Such observations demonstrate the advance of FGLISA.

Parameters Sensitivity. We investigate the sensitivity of FGLISA to its key parameters λ and ω. First, we analyze the effect of λ by varying it within $\{e^{-4}, e^{-3}, \cdots, e^{4}\}$ while keeping other hyper-parameters fixed. As shown in Figure 4, the impact of λ exhibits three distinct phases: i) When λ is small (less than e^{-2} for Credit, e^{-3} for Pokec-z, and e^{-2} for Pokec-n), disentangled learning has minimal effect on FGLISA's fairness. ii) As λ increases, we observe an improvement in fairness, along with a decrease in the cost of fairness. iii) When λ becomes relatively large (larger than e^{1} across all datasets), further increases show a diminishing impact on FGLISA's fairness. Similarly, we analyze ω by varying it within $\{e^{-4}, e^{-3}, \cdots, e^{4}\}$. As shown in Figure 4, increasing ω improves model fairness but may reduce prediction accuracy, as larger values impose stronger constraints on group differences in predictions.

Ablation Study. We conduct ablation studies to analyze the contribution of each component in FGLISA. We create two variants: FGLISA-WD (without disentangled learning) and FGLISA-WF (without fairness constraints, $\omega = 0$). Figure 3 presents the results on Credit, Pokec-z, and Pokec-n datasets. Our analysis reveals that FGLISA outperforms both ablation variants in fairness and utility metrics, demonstrating the importance of both components. When removing disentangled learning (FGLISA-WD), we observe decreased fairness performance due to less accurate demographic inference, which introduces additional bias and compromises bias mitigation. Without fairness constraints (FGLISA-WF), the model shows reduced fairness due to the absence of fairness constraints, which is particularly evident in the Credit and Pokec-z datasets, where it performs significantly worse than FGLISA in both metrics. In addition, we observe dataset-dependent variations in the contribution of disentangled learning, as shown by the improved appearance in Pokec-n. Overall, these findings underscore the necessity of our design choices, particularly the importance of excluding irrelevant information during demographic inference for effective bias mitigation.

Conclusion

Despite the growing attention to fairness in graph learning, existing fairness studies typically assume that sensitive attributes are either fully available or completely missing, thereby overlooking the real-world scenario of partial sensitive attribute availability in graph-structured data. To address this gap, we present FGLISA, a novel framework for mitigating bias with limited sensitive attributes during training. Through a disentangled variational auto-encoding framework, FGLISA effectively identifies sensitive attribute-related information to infer missing attributes while preventing manipulation during fairness optimization. Extensive experiments demonstrate that our method significantly outperforms both fairness-agnostic and fairness-aware baselines in bias mitigation. This work opens a promising direction for developing fair graph learning algorithms that can operate with partially available sensitive attributes.

Acknowledgements

This work was supported in part by the National Science Foundation (NSF) under Grant No. 2404039, and the Dissertation Fellowship from Florida International University.

References

Ashurst, C.; and Weller, A. 2023. Fairness Without Demographic Data: A Survey of Approaches. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–12.

Chai, J.; Jang, T.; and Wang, X. 2022. Fairness without demographics through knowledge distillation. *Advances in Neural Information Processing Systems*, 35: 19152–19164.

Dai, E.; and Wang, S. 2021. 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*, 680–688.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Friedmann, E.; and Efrat-Treister, D. 2023. Gender bias in STEM hiring: implicit in-group gender favoritism among men managers. *Gender & Society*, 37(1): 32–64.

Grari, V.; Lamprier, S.; and Detyniecki, M. 2021. Fairness without the sensitive attribute via causal variational autoencoder. *arXiv preprint arXiv:2109.04999*.

Grari, V.; Lamprier, S.; and Detyniecki, M. 2022. Fairness without the Sensitive Attribute via Causal Variational Autoencoder. In *Thirty-First International Joint Conference on Artificial Intelligence {IJCAI-22}*, 696–702. International Joint Conferences on Artificial Intelligence Organization.

Gretton, A.; Borgwardt, K.; Rasch, M.; Schölkopf, B.; and Smola, A. 2006. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19.

Guo, D.; Chu, Z.; and Li, S. 2023. Fair attribute completion on graph with missing attributes. *arXiv preprint arXiv:2302.12977*.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Hashimoto, T.; Srivastava, M.; Namkoong, H.; and Liang, P. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, 1929–1938. PMLR.

Jin, G.; Wang, Q.; Zhu, C.; Feng, Y.; Huang, J.; and Zhou, J. 2020. Addressing crime situation forecasting task with temporal graph convolutional neural network approach. In *2020 12th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, 474–478. IEEE.

Kamath, S.; and Anantharam, V. 2012. Non-interactive simulation of joint distributions: The Hirschfeld-Gebelein-Rényi maximal correlation and the hypercontractivity ribbon. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 1057–1064. IEEE.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kumar, S.; Mallik, A.; Khetarpal, A.; and Panda, B. S. 2022. Influence maximization in social networks using graph embedding and graph neural network. *Information Sciences*, 607: 1617–1636.

Lahoti, P.; Beutel, A.; Chen, J.; Lee, K.; Prost, F.; Thain, N.; Wang, X.; and Chi, E. 2020. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33: 728–740.

Li, R.; Zhao, J.; Li, C.; He, D.; Wang, Y.; Liu, Y.; Sun, H.; Wang, S.; Deng, W.; Shen, Y.; et al. 2022. House: Knowledge graph embedding with householder parameterization. In *International conference on machine learning*, 13209–13224. PMLR.

Li, Y.; Wang, X.; Xing, Y.; Fan, S.; Wang, R.; Liu, Y.; and Shi, C. 2024. Graph Fairness Learning under Distribution Shifts. In *Proceedings of the ACM on Web Conference 2024*, 676–684.

Ling, H.; Jiang, Z.; Luo, Y.; Ji, S.; and Zou, N. 2023. Learning fair graph representations via automated data augmentations. In *International Conference on Learning Representations (ICLR)*.

Liu, P.; Wei, H.; Hou, X.; Shen, J.; He, S.; Shen, K. Q.; Chen, Z.; Borisyuk, F.; Hewlett, D.; Wu, L.; et al. 2024. LinkSAGE: Optimizing Job Matching Using Graph Neural Networks. *arXiv preprint arXiv:2402.13430*.

Louizos, C.; Shalit, U.; Mooij, J. M.; Sontag, D.; Zemel, R.; and Welling, M. 2017. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30.

Ma, J.; Guo, R.; Wan, M.; Yang, L.; Zhang, A.; and Li, J. 2022. Learning fair node representations with graph counterfactual fairness. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 695–703.

Madden, M.; Lenhart, A.; Cortesi, S.; Gasser, U.; Duggan, M.; Smith, A.; and Beaton, M. 2013. Teens, social media, and privacy. *Pew Research Center*, 21(1055): 2–86.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Rawls, A. 1971. Theories of social justice.

Takac, L.; and Zabovsky, M. 2012. Data analysis in public social networks. In *International scientific conference and international workshop present day trends of innovations*, volume 1.

Wang, Z.; Chu, Z.; Blanco, R.; Chen, Z.; Chen, S.-C.; and Zhang, W. 2024a. Advancing Graph Counterfactual Fairness through Fair Disentangled Representation. In *Joint European*

Conference on Machine Learning and Knowledge Discovery in Databases.

Wang, Z.; Chu, Z.; Viet Doan, T.; Wang, S.; Wu, Y.; Palade, V.; and Zhang, W. 2025a. 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, 28485–28493.

Wang, Z.; Hoang, N.; Zhang, X.; Bello, K.; Zhang, X.; Iyengar, S. S.; and Zhang, W. 2025b. Towards Fair Graph Learning without Demographic Information. In *The 28th International Conference on Artificial Intelligence and Statistics*, volume 258, 2107–2115.

Wang, Z.; Liu, F.; Pan, S.; Liu, J.; Saeed, F.; Qiu, M.; and Zhang, W. 2025c. fairGNN-WOD: Fair Graph Learning Without Demographics. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*.

Wang, Z.; Narasimhan, G.; Yao, X.; and Zhang, W. 2023a. Mitigating multisource biases in graph neural networks via real counterfactual samples. In *2023 IEEE International Conference on Data Mining (ICDM)*, 638–647. IEEE.

Wang, Z.; Saxena, N.; Yu, T.; Karki, S.; Zetty, T.; Haque, I.; Zhou, S.; Kc, D.; Stockwell, I.; Bifet, A.; et al. 2023b. Preventing Discriminatory Decision-making in Evolving Data Streams. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Wang, Z.; Ulloa, D.; Yu, T.; Rangaswami, R.; Yap, R.; and Zhang, W. 2024b. Individual Fairness with Group Constraints in Graph Neural Networks. In *27th European Conference on Artificial Intelligence*.

Wang, Z.; Wallace, C.; Bifet, A.; Yao, X.; and Zhang, W. 2023c. FG^{2}AN: Fairness-Aware Graph Generative Adversarial Networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 259–275. Springer Nature Switzerland.

Wang, Z.; Wu, A.; Moniz, N.; Hu, S.; Knijnenburg, B.; Zhu, X.; and Zhang, W. 2025d. Towards fairness with limited demographics via disentangled learning. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*.

Wang, Z.; Yin, Z.; Yang, L.; Zhuang, J.; Yu, R.; Kong, Q.; and Zhang, W. 2025e. Fairness-Aware Graph Representation Learning with Limited Demographic Information. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer Nature Switzerland*.

Wang, Z.; Yin, Z.; Yap, R.; and Zhang, W. 2025f. Ai fairness beyond complete demographics: Current achievements and future directions. In *28th European Conference on Artificial Intelligence*.

Wang, Z.; Yin, Z.; and Zhang, W. 2025. A unified framework for fair graph generation: Theoretical guarantees and empirical advances. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Wang, Z.; Yin, Z.; Zhang, Y.; Yang, L.; Zhang, T.; Pissinou, N.; Cai, Y.; Hu, S.; Li, Y.; Zhao, L.; et al. 2025g. FG-SMOTE: Towards fair node classification with graph neural network. *ACM SIGKDD Explorations Newsletter*, 26(2): 99–108.

Wang, Z.; and Zhang, W. 2025. FDGen: A Fairness-Aware Graph Generation Model. In *Forty-second International Conference on Machine Learning*.

Yan, S.; Kao, H.-t.; and Ferrara, E. 2020. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 1715–1724.

Yeh, I.-C.; and Lien, C.-h. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2): 2473–2480.

Zhang, W. 2024a. AI fairness in practice: Paradigm, challenges, and prospects. *Ai Magazine*, 45(3): 386–395.

Zhang, W. 2024b. Fairness with censorship: Bridging the gap between fairness research and real-world deployment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 22685–22685.

Zhang, W.; Hernandez-Boussard, T.; and Weiss, J. 2023. Censored fairness through awareness. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 14611–14619.

Zhang, W.; and Ntoutsi, E. 2019. FAHT: an adaptive fairness-aware decision tree classifier. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 1480–1486.

Zhang, W.; and Weiss, J. C. 2022. Longitudinal fairness with censorship. In *proceedings of the AAAI conference on artificial intelligence*, volume 36, 12235–12243.

Zhang, W.; Zhou, S.; Walsh, T.; and Weiss, J. C. 2025. Fairness amidst non-IID graph data: A literature review. *AI Magazine*, 46(1): e12212.

Zhang, Z.; Liu, Q.; Jiang, H.; Wang, F.; Zhuang, Y.; Wu, L.; Gao, W.; and Chen, E. 2023. Fairlisa: Fair user modeling with limited sensitive attributes information. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhao, T.; Dai, E.; Shu, K.; and Wang, S. 2022. 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*, 1433–1442.

Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; and Sun, M. 2020. Graph neural networks: A review of methods and applications. *AI open*, 1: 57–81.

Zhu, Y.; Li, J.; Chen, L.; and Zheng, Z. 2024. The Devil is in the Data: Learning Fair Graph Neural Networks via Partial Knowledge Distillation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 1012–1021.