# Towards Fair Graph Learning without Demographic Information

**Zichong Wang**
Florida International University

**Nhat Hoang**
Florida International University

**Xingyu Zhang**
University of Pittsburgh

**Kevin Bello**
Soroco

**Xiangliang Zhang**
University of Notre Dame

**Sundararaja Sitharama Iyengar**
Florida International University

**Wenbin Zhang***
Florida International University

## Abstract

Fair Graph Neural Networks (GNNs) have been extensively studied in graph-based applications. However, most approaches to fair GNNs assume the full availability of demographic information by default, which is often unrealistic due to legal restrictions or privacy concerns, leaving a noticeable gap in methods for addressing bias under such constraints. To this end, we propose a novel method for fair graph learning without demographic information. Our approach leverages a Bayesian variational autoencoder to infer missing demographic information and uses disentangled latent variables to separately capture demographics-related and label-related information, reducing interference when inferring demographic proxies. Additionally, we incorporate a fairness regularizer that enables measuring model fairness without demographics while optimizing the fairness objective. Extensive experiments on three real-world graph datasets demonstrate the proposed method's effectiveness in improving both fairness and utility.

## 1 Introduction

Graph-structured data is prevalent in many real-world scenarios, including financial markets (Zhang et al., 2017), item recommendations (Wu et al., 2021), and social networks (Wan et al., 2019), which has spurred the development of various graph neural networks (GNNs) in recent years. However, despite their strong representation learning capacity, GNNs can unintentionally incorporate and amplify societal biases related to *demographic information*, such as age, gender, and race (Wang et al., 2023a). Consequently, numerous efforts have been directed towards developing fair GNNs (Dai and Wang, 2021; Wang et al., 2023a, 2024b). Most of these approaches focus on ensuring that the outcome statistics of classifiers are equitable across different demographic subgroups (Hardt et al., 2016), assuming that demographic information is available to identify both the deprived group (*e.g.*, female) and the favored group (*e.g.*, male) for bias quantification and mitigation (Wang et al., 2025).

However, this assumption is unrealistic in many real-world scenarios where collecting or using demographic information (*i.e.*, sensitive attribute) is infeasible due to privacy concerns, legal and regulatory restrictions, or fears of discrimination and social desirability (Krumpal, 2013; Lahoti et al., 2020). For instance, a tech company may employ GNNs to enhance its hiring process for software engineering positions by analyzing connections between applicants' professional networks and internal teams (*e.g.*, coding collaboration potential) (Liu et al., 2024). In some cases, applicants from certain gender groups may choose not to disclose their gender, particularly when applying for roles in male-dominated fields like software engineering (Friedmann and Efrat-Treister, 2023). This scenario highlights a significant gap between current fair graph learning models and their application in real-world settings. Indeed, regulations such as the European Union's General Data Protection Regulation (GDPR), introduced in 2018,

Proceedings of the 28$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s). *Corresponding author.

strictly control the collection and use of sensitive personal data (Hoofnagle et al., 2019). Article 9(1) of the GDPR prohibits processing data related to race, ethnicity, political opinions, religion, health, and similar categories to ensure non-discriminatory algorithms.

Despite the practical importance of achieving fairness without demographic information, this area remains largely unexplored due to several critical challenges: **i) Difficulties of excluding presumed interfering information:** Inferring demographics from observed graph data often relies on unsupervised learning, which requires careful exclusion of irrelevant information. For example, significant issues arise when models infer demographic information from labels (*e.g.*, assuming that loan applicants are male if their applications are approved). A robust design is thus required to prevent interference from noisy or misleading information during the inference process. **ii) Inability to define fairness loss without demographics:** In scenarios where demographic information is missing, the fairness loss across different subgroups cannot be directly calculated. Consequently, a differentiable fairness loss needs to be integrated into the learning framework to simultaneously optimize performance and fairness without relying on explicit demographic information. **iii) Disentangling demographic inference and label prediction:** Unlike existing fair graph models that can focus solely on bias mitigation for fairness, the absence of demographic information necessitates simultaneously inferring missing demographic information and predicting node labels. In addition, it is crucial to avoid model manipulation that skews demographic inference to artificially enhance the fairness of the classifier during this dual process.

In response to these challenges, this paper introduces a novel framework, ***T**owards fair grap**H** l**E**arning without de**M**ograph**I**c**S*** (**Themis**), designed to *tackle the open research question of mitigating bias in graph learning algorithms without demographic information—marking the first work, to the best of our knowledge, to enable fair graph learning under these practical conditions.* Specifically, Themis first conducts a comprehensive causal analysis to construct a latent causal model, which forms the basis for variable separation. Guided by the causal insights, Themis decomposes the different causal effects into multiple latent variables, allowing the separation of *demographics-related information* (such as height when the demographics is gender) from *label-related information* (such as job performance when the label is income). With this disentangled structure, demographics-related information is effectively prevented from leaking into label prediction, ensuring that the inference of demographic information remains unaffected by noisy inputs. Additionally, a differentiable fairness loss is introduced, which enables measuring model fairness without demographics while optimizing the fairness objective. This fairness loss can be directly integrated into other methods that tackle fairness without demographic information, enabling regularization of the learned posterior distribution to minimize risks in both utility and fairness.

The main contributions of this paper are summarized as follows:

- We address new challenges in fair graph learning when demographic information is unavailable. We then propose Themis, a Bayesian variational autoencoder that infers demographic information and integrates it with a differentiable fairness loss to achieve fairness without relying on demographic information.
- A new differentiable fairness loss is proposed to estimate fairness loss in the absence of demographic information, enabling the assessment of model bias.
- A novel strategy is proposed to improve demographic recovery by mitigating noisy information, thereby accurately inferring missing demographic information.
- Extensive experiments on real-world datasets demonstrate Themis outperforms existing baselines in multiple fairness metrics while achieving comparable utility.

## 2 Related Work

### 2.1 Fairness-Aware Graph Learning

Fairness-aware graph learning has gained significant attention as the use of GNNs has expanded into high-risk decision-making scenarios (Zhang and Ntoutsi, 2019; Zhang and Weiss, 2022; Zhang et al., 2023; Wang et al., 2023b,c, 2024b). Numerous approaches have been proposed to achieve fairness in graph learning, with most focusing on ensuring that algorithmic decisions neither favor nor disadvantage specific demographic groups. For instance, FairOT (Laclau et al., 2021) employs preprocessing techniques to eliminate bias in the graph or node features before training the GNN. GEIF (Wang et al., 2024c) and FDGNN (Wang et al., 2024a) modify the objective function by incorporating fairness constraints during training to ensure the learning of fair representations. However, the widespread assumption in these methods that demographic information is readily available is often unrealistic in many real-world applications due to practical challenges and regulatory restrictions (Grari et al., 2021). FairGNN (Dai and

Wang, 2021) is the initial approach that addresses the challenge of learning fair GNNs with limited demographic information by employing a demographic estimator to predict missing data while enhancing fairness through adversarial learning. However, it still relies on the availability of partial demographic information.

### 2.2 Fairness without Demographic Information

There is a growing trend in the research community for achieving fairness without demographic information, primarily focusing on simplified non-graph domains, which can be broadly divided into two categories: using proxy features (Zhao et al., 2022) or adhering to Rawlsian Max-Min fairness Ashurst and Weller (2023). Specifically, proxy features (Gupta et al., 2018; Chen et al., 2019; Kallus et al., 2022) are designed to approximate missing demographics and correlate these with predictions to enhance fairness. However, these methods rely on the notion of binary group membership fairness but are not modeled by the method, leading to potential inferred biases. Another line of research, Max-Min fairness Hashimoto et al. (2018); Lahoti et al. (2020), focuses on minimizing harm to the most disadvantaged subgroup but can significantly degrade overall model utility. In addition, this approach enforces uniform utility across all subgroups and may not necessarily enhance group fairness, as improving the worst-group utility does not always reduce inter-group disparities. Note that these methods are not specifically designed for graph data and cannot be easily applied to such contexts.

Different from cited works, our work takes a fresh look at achieving fairness in graph learning without demographic information, rather than focusing on fairness in tabular data. Additionally, our approach explicitly considers the impact of noisy information on the inference of missing demographic data, minimizing potential inference errors. Furthermore, we take one more step to address the risk of model manipulation during the inference process, which could artificially enhance perceived fairness, a concern often overlooked by existing methods.

## 3 Notations

We denote a graph by $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ with the node set $|\mathcal{V}| = n$ 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 $S_d = \{\forall\ v_i \in \mathcal{V} \mid s_i = 0\}$ denotes the deprived group (*e.g.*, female) and $S_f = \{\forall\ v_i \in \mathcal{V} \mid s_i = 1\}$ 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 predicted label of $v_i$. We also assume $y_i = 1$ denotes the granted label and $y_i = 0$ denotes the rejected label.

## 4 Methodology

This section introduces the novel framework, Themis, designed to enable fair graph learning without demographic information. It begins with the causal model presented in Section 4.1, which lays the foundation for the framework. Following that, Section 4.2, we explain how to accurately impute missing demographic information from observational data. We focus on two key aspects: i) identifying demographics-related information while excluding noisy data during the inference process; ii) measuring model fairness without demographic information, and ensuring that the model does not artificially improve fairness by manipulating the inferred demographic information.

### 4.1 Causal Method

This section presents the Structural Causal Model (SCM), essential for addressing bias in graph learning without demographic information, focusing on four key variables: demographic information ($S$), node features ($X$), graph structure ($A$), ground-truth label ($Y$), and two exogenous variables ($U_S$ and $U_Y$), representing the stochastic elements of a variable, as shown in Figure 1. Each connection in SCM reflects a causal relationship, with reasoning and explanations detailed below:

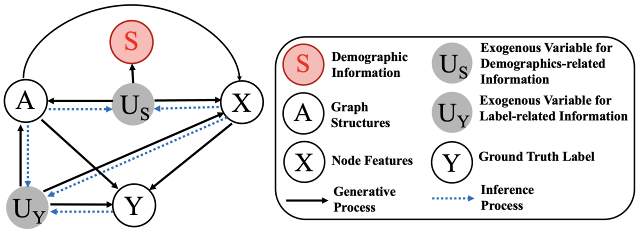
S
A
U_S
X
U_Y
Y
S Demographic Information
U_S Exogenous Variable for Demographics-related Information
A Graph Structures
U_Y Exogenous Variable for Label-related Information
X Node Features
Y Ground Truth Label
Generative Process
Inference Process

Figure 1: The causal model of Themis includes unobserved demographic information $S$, observed attributes $(Y, A, X)$, and two exogenous variables $U_S$ and $U_Y$, where solid-line arrows denote the generative process and dashed-line arrows represent the inference process.

- $A \leftarrow S \rightarrow X$: Demographic information $S$, typically determined at birth, influences both the node features $X$ and the graph structure $A$. For example, a demographic attribute like "gender" can affect features like "height", as well as the graph structure in social networks, where individuals may form connections with others of the same gender. However, these connections do not influence the demographic attribute itself.

- $A \rightarrow Y \leftarrow X$: Both the node features $X$ and the graph structure $A$ provide critical information for determining the ground-truth label $Y$. For instance, in a job application scenario, an applicant's qualifications $(X)$ and their professional network $(A)$ can both significantly influence the hiring decision $(Y)$.
- $S \perp Y$: Demographic information $S$ should be independent of the ground-truth label $Y$ to prevent biased predictions. For example, individuals of different genders (represented by $S$) should have equal chances of acceptance in job applications.
- $A \rightarrow X$: The graph structure $A$ affects the node features $X$. For example, if an applicant's professional network is primarily composed of individuals with experience in a specific technology or field, it increases the likelihood that the applicant has developed expertise in that area as well, reflecting the influence of their connections on their skills and qualifications.

Building on these deterministic causal relationships, we introduce two exogenous variables, $U_S$ and $U_Y$. Specifically, $U_S$ captures latent information related to demographic information (*i.e.*, demographics-related information) and serves as a proxy for missing demographic information $S$, while $U_Y$ captures latent information relevant to label prediction (*i.e.*, label-related information). This design allows us to effectively model positional and structural representations, improving the inference of missing demographic information.

## 4.2 Fairness-Aware Missing Demographics Imputation

Building on the proposed causal model, we then take the problem of identifying missing demographic information by using variational autoencoders to infer $U_S$ to be used as the proxy of the demographic information and approximately recover joint distribution $P(X, A, S, Y)$ among node features $X$, graph structures $A$, ground-truth label $Y$, and the unobservable demographics $S$. Specifically, we leverage Bayesian inference approximation based on the proposed causal model in Section 4.1, where the two exogenous variables, $U_S$ and $U_Y$, in the causal graph are represented as two latent variables in the Bayesian network. Mathematically, this is represented as follows:

$$\begin{aligned} P(X, A, Y, U_S, U_Y) = & P(U_S)P(U_Y)P(Y|U_Y, X, A) \\ & P(X|U_S, A, U_Y)P(A|U_S, U_Y) \end{aligned} \tag{1}$$

where $P(U_S)$ represents the marginal distribution of $S$, typically modeled as a standard Gaussian distribution $\mathcal{N}(0, I)$, where $I$ denotes the identity matrix, and the prior for $P(U_Y)$ is defined similarly.

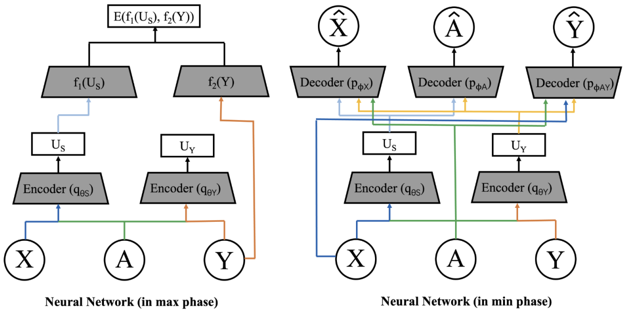

Figure 2: Neural network architecture of Themis.

Building on the Equation 1, the structure of Themis is illustrated in Figure 2, which provides an overview of the proposed neural network architecture. Specifically, the neural network consists of an inference structure on two latent random variables, *i.e.*, $q_{\theta_S}$ and $q_{\theta_Y}$, and the decoder distribution $p_\phi(X, A, Y|U_S, U_Y))$ can be factorized as below:

$$\begin{aligned} p_\phi(X, A, Y|U_S, U_Y) = & p_{\phi_X}(X|U_S, A, U_Y) \\ & p_{\phi_A}(A|U_S, U_Y)p_{\phi_Y}(Y|X, A, U_Y) \end{aligned} \tag{2}$$

Building on this, we aim to approximate the joint distribution of the observed graph data (*i.e.*, $X, A, Y$) by maximizing a variational lower bound on the log-probability. Specifically, we use two latent variables: $U_S$, which acts as a proxy for the unobserved demographic information $S$, and $U_Y$, which captures latent factors related to predicting $Y$. Since $S$ is unobserved and needs to be inferred, we cannot directly compute the posterior distribution $P(U_S, U_Y|X, A, Y)$ due to its intractability. To approximate the intractable posterior distributions of these latent variables, we introduce variational distributions $Q(U_S|X, A)$ and $Q(U_Y|X, A, Y)$, which use a parametric family of distributions to approximate the true posteriors $P(U_S|X, A)$ and $P(U_Y|X, A, Y)$. The parameters of this inference model are learned by maximizing the evidence lower bound (ELBO) on the marginal likelihood of the data, as described in Section 4.1. To optimize the parameters of the inference model, we maximize the ELBO as follows:

$$
\begin{aligned}
\log P(X, A, Y|U_S, U_Y) \geq & \mathbb{E}_{Q(U_S|X,A),Q(U_Y|X,A,Y)}[\\
& \log P(Y|U_Y, X, A)\\
& + \log P(X|U_S, A, U_Y)\\
& + \log P(A|U_S, U_Y)]\\
& - \mathrm{KL}(Q(U_Y|X, A, Y)\|P(U_Y))\\
& - \mathrm{KL}(Q(U_S|X, A)\|P(U_S))
\end{aligned} \tag{3}
$$

where KL denotes the Kullback-Leibler divergence of the posterior $q_{\theta_S}$ and $q_{\theta_Y}$ from a prior $P(U_S)$ and $P(U_Y)$, showing as follow:

$$
\begin{cases}
\mathrm{KL}(Q(U_Y|X, A, Y)\|P(U_Y)) = \mathbb{E}_{Q(U_Y|X,A,Y)}[\log Q(U_Y|X, A, Y) \\
\qquad\qquad\qquad\qquad\qquad\qquad - \log P(U_Y)] \\
\mathrm{KL}(Q(U_S|X, A)\|P(U_S)) = \mathbb{E}_{Q(U_S|X,A)}[\log Q(U_S|X, A) \\
\qquad\qquad\qquad\qquad\qquad\qquad - \log P(U_S)]
\end{cases} \tag{4}
$$

The ELBO can now be maximized through stochastic gradient ascent, facilitated by the reparameterization trick (Kingma and Welling, 2013). Additionally, the likelihood term $p_\phi(X, A, Y|U_S, U_Y)$ is parameterized by neural networks, where the parameters $\phi$ represent the model's architecture. Given that $U_S$ is attracted by a standard prior, we must remove task-related information from its posterior distribution. In other words, accurate inference requires decorrelation between $S$ and $Y$ to prevent bias in the inference of $S$. However, optimizing the ELBO alone does not guarantee independence. To address this, we introduce a penalization term in the loss function. Specifically, we extend the Hirschfeld-Gebelein-Rényi (HGR) (Gebelein, 1941) measure to quantify the correlation between $S$ and $Y$, as detailed in Definition 4.1.

**Definition 4.1** (*$U_S$, Y-correlation*). Given a latent space $U_S$ and an outcome variable $Y$, the HGR maximal correlation between $U_S$ and $Y$ is defined as:

$$
\begin{aligned}
\mathrm{HGR}(U_S, Y) &= \sup_{f_1, f_2} \rho(f_1(U_S), f_2(Y))\\
&= \sup_{f_1, f_2} \frac{\mathbb{E}(f_1(U_S) f_2(Y))}{\sqrt{\mathbb{E}({f_1}^2(U_S))\mathbb{E}({f_2}^2(Y))}}
\end{aligned} \tag{5}
$$

where $\rho$ signifies the Pearson correlation coefficient, $f_1$ and $f_2$ are measurable functions with finite, positive variances. To ensure that comparisons are made on a consistent scale, we normalize these functions such that $\mathbb{E}[f_1(U_S)] = \mathbb{E}[f_2(Y)] = 0$ and $\mathbb{E}[f_1^2(U_S)] = \mathbb{E}[f_2^2(Y)] = 1$. It is important to note that the HGR maximal correlation coefficient equals zero when $U_S$ and $Y$ are independent, and equals one when there is a deterministic relationship between them.

Building on this, the ELBO can be reformulated as follows:

$$
\begin{aligned}
\log P(X, A, Y|U_S, U_Y) \geq & \mathbb{E}_{Q(U_S|X,A),Q(U_Y|X,A,Y)}[\\
& \log P(Y|U_Y, X, A)\\
& + \log P(X|U_S, A, U_Y)\\
& + \log P(A|U_S, U_Y)]\\
& - \mathrm{KL}(Q(U_Y|X, A, Y)\|P(U_Y))\\
& - \mathrm{KL}(Q(U_S|X, A)\|P(U_S))\\
& + \lambda \cdot \mathrm{HGR}(U_S, Y)
\end{aligned} \tag{6}
$$

where $\lambda$ denotes the hyperparameter that balances the contribution of the penalization term. To optimize the updated ELBO, our objective becomes the min-max structure correspond to accurately infer missing demographic information and exclude label information. Specifically, as shown in Figure 2, in the max phase, we use gradient ascent to estimate the HGR maximal correlation between $U_S$ and $Y$ by optimizing the functions $f_1$ and $f_2$ in Equation 5. These functions are parameterized by neural networks and are trained to maximize the Pearson correlation $\rho(f_1(U_S), f_2(Y))$. In the min phase, we minimize the overall objective, which includes the ELBO and the penalization term $\lambda \cdot \mathrm{HGR}(U_S, Y)$, with respect to the model parameters. This iterative min-max optimization allows us to effectively reduce the dependency between $U_S$ and $Y$, promoting the independence necessary for accurate inference of the demographic information $S$.

With the strategy to infer demographic information, we proceed to mitigate bias. However, existing bias mitigation methods cannot be directly applied because they typically assume the prior availability of complete demographic information, which is not the case during this integrated process until it is fully completed. To address this, we introduce a differentiable fairness loss that utilizes the inferred demographic information during the training process. Specifically, we estimate the demographic information $S$ by leveraging $U_S$ through an auxiliary classifier (*i.e.*, $P(\hat{S} = 1 \,|\, U_S) = \sigma(f_{\mathrm{aux}}(U_S)))$, without requiring complete demographic information. Notable, $\sigma(\cdot)$ is the Sigmoid function. The auxiliary classifier predicts the demographic information $\hat{S}$ based on $U_S$, which serves as a proxy for the incomplete demographic information. This allows us to compute the fairness loss using the inferred demographic information in a differentiable manner. Notable, we design the auxiliary classifier using unsupervised methods to improve the estimation of $\hat{S}$. Specifically, we impose an entropy minimization loss on the outputs of the auxiliary classifier to encourage it to produce confident and consistent predictions. The entropy minimization loss is defined as follows:

$$\begin{aligned}\mathcal{L}_a = -\sum_i [&P(\hat{S}_i = 1 \,|\, U_S) \log P(\hat{S}_i = 1 \,|\, U_S) \\ &+ P(\hat{S}_i = 0 \,|\, U_S) \log P(\hat{S}_i = 0 \,|\, U_S)]\end{aligned} \tag{7}$$

Building on this, the proposed fairness loss can be defined as follows:

**Definition 4.2 (Estimated Fairness Loss).** Given an input graph $\mathcal{G}$, the Estimated Fairness Loss (EFL) is formally defined as:

$$\begin{aligned}\mathrm{EFL} = &\left( \frac{\sum_i P(\hat{s}_i = 1 \,|\, U_S) \cdot p_{\phi_Y}(\hat{y}_i = 1 \,|\, v_i)}{\sum_i P(\hat{s}_i = 1 \,|\, U_S)} \right) \\ &- \left( \frac{\sum_i P(\hat{s}_i = 0 \,|\, U_S) \cdot p_{\phi_Y}(\hat{y}_i = 1 \,|\, v_i)}{\sum_i P(\hat{s}_i = 0 \,|\, U_S)} \right)\end{aligned} \tag{8}$$

where $P(\hat{s}_i = s \,|\, U_S)$ denotes the probability that the auxiliary classifier assigns node $v_i$ to demographic group $S_i$, and $p_{\phi_Y}(\hat{y}_i = 1 \,|\, v_i)$ is the probability that node $v_i$ is predicted to have a granted label by the model. In practice, EFL computes the estimated fairness loss by comparing the probability means of the predicted positive outcome between different demographic groups as identified by the inferred $S$. We will begin this approach by primarily using simple averaging while also exploring alternative estimation methods, such as kernel density estimation (Cho et al., 2020).

However, simple incorporation of EFL may lead to manipulated predictions in the model's inference of demographic information. Specifically, when inferring $S$, because the ready-to-use solution simultaneously infers prediction outcome and demographic information, there is a risk that the model might manipulate an individual's demographic information to improve its fairness (*e.g.*, randomly changing a male with positive outcome to female to achieve statistical parity). This issue arises specifically when jointly optimizing $P_S$ and $p_{\phi_Y}$, but it does not occur when demographic information is fully available, either through the complete demographic assumption or demographic information identification. To this end, to prevent the learning of $P_S$ from being influenced by $p_{\phi_Y}$ through shared backbones, the stop-gradient technique will be employed. Specifically, during backpropagation, the gradient of the EFL with respect to $P_S(\hat{S}|U_S)$ will be halted (*i.e.*, set to zero), effectively treating it as a constant. This strategy helps maintain separation in the learning pathways to avoid potential fairness manipulation while simplifying the training process.

Furthermore, during the inference of $Y$, the expectation $\mathbb{E}_{y_i \sim p_{\phi_Y}(y_i|x_i, A_i, \tilde{y}_i)}$ can also benefit from the use of gradient stopping, though this introduces additional complexity and design considerations. Specifically, the prediction of $Y$ should not be guided solely by its predictive performance, as $Y$ also contributes to fairness performance. Essentially, as $p_{\phi_Y}$ is subject to both fairness and performance losses, applying stop-gradient prevents the simultaneous optimization of these objectives. Consequently, we plan to train a separate semi-supervised classifier to impute $y_i$ without incorporating fairness constraints directly. Next, the predictions from $p_{\phi'_Y}$ clamps $\hat{y_i}'$ to $p_{\phi_Y}$ as the observed labels for training. For the remaining samples without observed labels, we infer class probabilities $p_{\phi'_Y}(Y|x_i)$ using semi-supervised learning strategy, leading to the following update in the EFL:

$$\begin{aligned}EFL \approx &\mathbb{E}_{y_i \sim p_{\phi'_Y}(Y|x_i, A_i, \tilde{y}_i = \emptyset)} \left( \frac{\sum_i P(\hat{S}_i = 1 \,|\, U_S) \cdot y_i}{\sum_i P(\hat{S}_i = 1 \,|\, U_S)} \right) \\ &- \left( \frac{\sum_i P(\hat{S}_i = 0 \,|\, U_S) \cdot y_i}{\sum_i P(\hat{S}_i = 0 \,|\, U_S)} \right)\end{aligned} \tag{9}$$

Armed with the proposed EFL, we augment the ELBO as follows:

$$\begin{aligned}\log P(X, A, Y | U_S, U_Y) \geq &\mathbb{E}_{Q(U_S|X,A), Q(U_Y|X,A,Y)}[ \\ &\log P(Y|U_Y, X, A) \\ &+ \log P(X|U_S, A, U_Y) \\ &+ \log P(A|U_S, U_Y)] \\ &- \mathrm{KL}(Q(U_Y|X,A,Y) \| P(U_Y)) \\ &- \mathrm{KL}(Q(U_S|X,A) \| P(U_S)) \\ &+ \lambda \cdot \mathrm{HGR}(U_S, Y) - \gamma \cdot \mathrm{EFL}\end{aligned} \tag{10}$$

where $\gamma$ denotes a tradeoff hyperparameter.

Overall, Themis employs a disentangled structure to prevent noisy information from affecting the inference of missing demographic data. It measures fairness loss between the predicted output and the inferred demographic proxy while simultaneously optimizing both prediction loss and fairness loss. By doing so, Themis can improve fairness performance while maintaining comparable accuracy.

# 5 Experiment

**Datasets.** We assess the effectiveness of Themis using three widely used graph datasets in fairness research: Credit, Pokec-z, and Pokec-n (Zhang et al., 2025). In **Credit** dataset (Yeh and Lien, 2009), each node represents a default payment record of a credit card holder, with connections formed based on similarities in purchase and payment patterns. The demographic information here is "age", and the task involves predicting

Table 2: Comparison results of Themis 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.

| **Dataset** | **Methods**<br>**Metrics** | **GCN** | **GIN** | **FairGNN** | **RFCGNN** | **Graphair** | **FairAGG** | **Themis** |
|---|---|---|---|---|---|---|---|---|
| **Credit** | **Accuracy** (↑) | 0.737 ± 0.071 | 0.751 ± 0.013 | 0.687 ± 0.012 | 0.723 ± 0.023 | 0.531 ± 0.054 | 0.653 ± 0.028 | **0.767** ± **0.032** |
| | **F1-Score** (↑) | 0.812 ± 0.028 | 0.831 ± 0.018 | 0.783 ± 0.043 | 0.821 ± 0.062 | 0.728 ± 0.082 | 0.747 ± 0.008 | **0.865** ± **0.018** |
| | **SPD**(↓) | 0.111 ± 0.010 | 0.133 ± 0.037 | 0.123 ± 0.036 | 0.076 ± 0.043 | 0.085 ± 0.077 | 0.074 ± 0.044 | **0.016** ± **0.021** |
| | **EOD**(↓) | 0.097 ± 0.007 | 0.128 ± 0.047 | 0.115 ± 0.042 | 0.078 ± 0.033 | 0.088 ± 0.068 | 0.056 ± 0.045 | **0.026** ± **0.013** |
| **Pokec-z** | **Accuracy** (↑) | 0.813 ± 0.014 | 0.792 ± 0.001 | 0.689 ± 0.091 | 0.813 ± 0.041 | 0.655 ± 0.024 | 0.735 ± 0.032 | **0.846** ± **0.041** |
| | **F1-Score** (↑) | 0.782 ± 0.002 | 0.787 ± 0.007 | 0.687 ± 0.033 | 0.794 ± 0.038 | 0.647 ± 0.019 | 0.716 ± 0.002 | **0.835** ± **0.032** |
| | **SPD**(↓) | 0.041 ± 0.005 | 0.048 ± 0.004 | 0.038 ± 0.022 | 0.023 ± 0.021 | 0.035 ± 0.008 | 0.022 ± 0.014 | **0.015** ± **0.007** |
| | **EOD**(↓) | 0.053 ± 0.003 | 0.057 ± 0.007 | 0.033 ± 0.029 | 0.028 ± 0.018 | 0.032 ± 0.006 | 0.025 ± 0.011 | **0.010** ± **0.013** |
| **Pokec-n** | **Accuracy** (↑) | 0.789 ± 0.007 | 0.809 ± 0.008 | 0.726 ± 0.013 | 0.811 ± 0.022 | 0.641 ± 0.009 | 0.694 ± 0.017 | **0.831** ± **0.024** |
| | **F1-Score** (↑) | 0.661 ± 0.001 | 0.686 ± 0.012 | 0.637 ± 0.019 | 0.683 ± 0.026 | 0.553 ± 0.012 | 0.603 ± 0.032 | **0.756** ± **0.029** |
| | **SPD**(↓) | 0.049 ± 0.004 | 0.054 ± 0.006 | 0.036 ± 0.012 | 0.053 ± 0.013 | 0.055 ± 0.028 | 0.049 ± 0.011 | **0.022** ± **0.016** |
| | **EOD**(↓) | 0.036 ± 0.003 | 0.045 ± 0.007 | 0.044 ± 0.020 | 0.049 ± 0.017 | 0.053 ± 0.016 | 0.046 ± 0.015 | **0.020** ± **0.005** |

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

| **Dataset** | **Credit** | **Pokec-z** | **Pokec-n** |
|---|---|---|---|
| # Vertices | 30,000 | 67,797 | 66,569 |
| # Edges | 200,526 | 882,765 | 729129 |
| Feature Dimension | 13 | 59 | 59 |
| Demographic information | Age | Region | Region |

the likelihood of a user defaulting on their credit card payments. The **Pokec-z** and **Pokec-n** datasets (Takac and Zabovsky, 2012) are derived from a popular social networking platform in Slovakia, capturing social network data from two different provinces. Nodes correspond to users with features such as gender, age, and interests, while edges represent friendships between users. Using "region" as the demographic information, the objective is to predict the working field of each user.

For all datasets, we exclude nodes that are not connected to others. We randomly allocate 20% of the nodes for validation and 30% for testing. To simulate scenarios with demographic information is completely absent, we mask all demographic information in the dataset (*i.e.*, $S = \emptyset$). This masking is applied to both the training and validation sets, and no demographic information is utilized during the testing phase. Details of these datasets are summarized in Table 1.

**Baselines.** We evaluate the performance of our proposed Themis[1] method by comparing it with several baseline methods, categorized into two groups: i) Vanilla Graph Models: **GCN** (Kipf and Welling, 2016) employs spatial graph convolutions to aggregate information from neighboring nodes. **GIN** (Xu et al., 2018) enhances node representation learning using multilayer perceptrons. ii) Fairness-aware Methods: **FairGNN** (Dai and Wang, 2021) utilizes a demographic information estimator to predict demographic information and improve fairness through adversarial training. **RFCGNN** (Wang et al., 2023a) learns the fair node representation by masking demographics-related information. **Graphair** (Ling et al., 2023) aims to automatically generate fair graph data to deceive the discriminator via adversarial learning. **FairAGG** (Zhu et al., 2024) implements a fair aggregation scheme based on the Shapley value to ensure group fairness.

Note that the vanilla graph models do not utilize demographic information. In contrast, the fairness-aware methods we provide demographic information, aligning with their design requirements.

**Evaluation Metrics.** We evaluate Themis with respect to two key aspects: prediction performance and fairness. For prediction performance, we use the widely adopted node classification metrics: Accuracy and F1-score (Sokolova and Lapalme, 2009). To assess the fairness, we employ two commonly used fairness metrics: Statistical Parity Difference (SPD) and Equal Opportunity Difference (EOD) (Zhang et al., 2025). Fairness is considered improved when fairness metrics are closer to zero.

### 5.1 Experimental Results

**Prediction Performance and Fairness.** The prediction performance and fairness results, shown in Table 2, indicate that Themis consistently outperforms the six baseline methods across all evaluation metrics. Specifically: i) Themis demonstrates significant improvement in fairness over the vanilla model. This is largely because Themis effectively infers a proxy for demographic information that closely approximates true demographic data, which provides a more reliable basis for mitigating unfairness. ii) Themis achieves better fairness than the state-of-the-art fairGNN. This improvement is due to its separated structure, which

[1] https://github.com/LavinWong/Themis

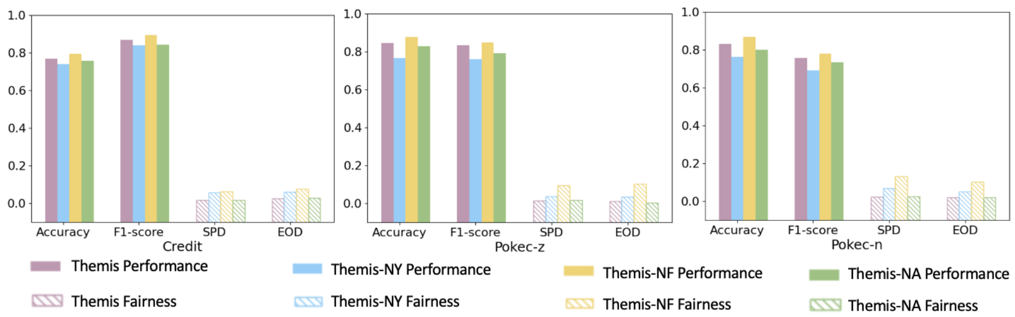

Figure 3: Ablation study results for Themis, Themis-NY, Themis-NF, and Themis-NA.

Table 3: Recovered demographic information results of Themis and Themis-NY.

| Accuracy / Methods / Datasets | Themis | Themis-NY |
|---|---|---|
| **Credit** | 0.745 | 0.723 |
| **Pokec-z** | 0.726 | 0.692 |
| **Pokec-n** | 0.738 | 0.701 |

isolates the demographics-related information in the graph data. By doing so, the model better captures latent factors behind data compared to methods that apply fairness constraints to all information indiscriminately, enabling superior fairness outcomes. iii) Themis also surpasses other methods in utility performance in most cases, underscoring its ability to preserve important task-relevant information while enforcing fairness. Overall, the experimental results demonstrate that Themis effectively improves fairness without compromising performance.

**Quality of Proxy to Recovery Demographics.** To assess the effectiveness of Themis in inferring accurate demographic proxies, we compared the model-generated proxies with the ground-truth demographic information. Specifically, we evaluated the quality of demographic proxies produced by the standard Themis and a variant, Themis-NY, which does not mitigate the influence of label information (*i.e.*, $\lambda = 0$). The results, summarized in Table 3, are based on three real-world datasets. Themis-NY generated significantly lower-quality demographic proxies than the standard Themis. This degradation is attributed to the leakage of label information into the demographic reconstruction process, leading to biases where granted labels are associated with favored groups and refused labels with deprived groups, thus reducing the accuracy of the demographic inference. These findings highlight the critical importance of effectively separating label information from the demographic inference process in Themis to ensure the integrity and accuracy of the generated proxies.

**Ablation Studies.** To evaluate the effectiveness of each part of Themis, an ablation study was conducted. First, we evaluated the impact of excluding label interference during demographic inference using the Themis-NY variant. The results on the Credit, Pokec-z, and Pokec-n datasets, as shown in Figure 3, demonstrate a significant reduction in fairness for Themis-NY. This drop is attributed to the model's inability to filter out noisy label information, compromising the quality of the demographic proxy and affecting bias mitigation. Next, we assessed the effect of removing the fairness constraint (*i.e.*, EFL) by testing the Themis-NF variant, which focuses solely on utility (*i.e.*, $\gamma = 0$). The comparison with the original Themis configuration revealed a marked decline in fairness, emphasizing the essential role of the fairness constraint in mitigating demographic biases in graph datasets. Finally, to evaluate the causal relationship between graph structure and node features established in the causal model, we introduced Themis-NA, which omits graph structure information in the X-reconstruction process (*i.e.*, $\log P(X|U_S, U_Y)$). The results showed declines in both fairness and overall model performance, underscoring the importance of this causal connection. In summary, these ablation studies affirm the critical role of each component within the Themis framework for both fairness and utility.

## 6 Conclusion

Despite the growing attention on fairness in graphs, existing research often assumes the default availability of demographic information, which limits its real-world applicability. Motivated by this, this paper proposes a disentangling structure that identifies demographics-related information from observed graph data while preventing the model from manipulating the inferred demographic information to artificially enhance fairness, thereby achieving graph fairness without demographic information. Extensive experiments demonstrate that the proposed method can achieve state-of-the-art performance on three real-world datasets with respect to the prediction fairness trade-off.

## Acknowledgements

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

### 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*, pages 1–12.

Chen, J., Kallus, N., Mao, X., Svacha, G., and Udell, M. (2019). Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 339–348.

Cho, J., Hwang, G., and Suh, C. (2020). A fair classifier using kernel density estimation. *Advances in neural information processing systems*, 33:15088–15099.

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*, pages 680–688.

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.

Gebelein, H. (1941). 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.

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

Gupta, A., Murali, A., Gandhi, D. P., and Pinto, L. (2018). Robot learning in homes: Improving generalization and reducing dataset bias. *Advances in neural information processing systems*, 31.

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*, pages 1929–1938. PMLR.

Hoofnagle, C. J., Van Der Sloot, B., and Borgesius, F. Z. (2019). The european union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, 28(1):65–98.

Kallus, N., Mao, X., and Zhou, A. (2022). Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 68(3):1959–1981.

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*.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & quantity*, 47(4):2025–2047.

Laclau, C., Redko, I., Choudhary, M., and Largeron, C. (2021). All of the fairness for edge prediction with optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 1774–1782. PMLR.

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.

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*.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Inf. Process. Manage.*, 45(4):427–437.

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.

Wan, H., Zhang, Y., Zhang, J., and Tang, J. (2019). Aminer: Search and mining of academic social networks. *Data Intelligence*, 1(1):58–76.

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. (2025). Fair graph u-net: A fair graph learning framework integrating group and individual awareness. In *proceedings of the AAAI 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)*, pages 638–647. IEEE.

Wang, Z., Qiu, M., Chen, M., Salem, M. B., Yao, X., and Zhang, W. (2024b). Toward fair graph neural networks via real counterfactual samples. *Knowledge and Information Systems*, pages 1–25.

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. (2024c). 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*, pages 259–275. Springer Nature Switzerland.

Wu, J., Wang, X., Feng, F., He, X., Chen, L., Lian, J., and Xie, X. (2021). 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.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2018). How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

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, S., Zhou, D., Yildirim, M. Y., Alcorn, S., He, J., Davulcu, H., and Tong, H. (2017). Hidden: hierarchical dense subgraph detection with application to financial fraud detection. In *Proceedings of the 2017 SIAM international conference on data mining*, pages 570–578. SIAM.

Zhang, W., Hernandez-Boussard, T., and Weiss, J. (2023). Censored fairness through awareness. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 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*, pages 1480–1486.

Zhang, W. and Weiss, J. C. (2022). Longitudinal fairness with censorship. In *proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 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.

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*, pages 1433–1442.

Zhu, Y., Li, J., Chen, L., and Zheng, Z. (2024). Fairagg: Toward fair graph neural networks via fair aggregation. *IEEE Transactions on Computational Social Systems*.