Towards Fair Graph Learning without Demographic Supervision

Zichong Wang[1] , **Zhipeng Yin**[1] , **Mo Sha**[1] ,
Xiaofeng Gao[1] , **Xiaoli Li**[2] and **Wenbin Zhang**[1*]
[1]Florida International University, FL, United States
[2]Singapore University of Technology and Design, Singapore

Abstract

Graph Neural Networks (GNNs) have demonstrated strong predictive performance across a wide range of applications. However, their increasing deployment has raised critical fairness concerns, as these models can inherit and amplify existing biases. Most existing fairness approaches rely on explicit demographic information, either directly available or inferred, to measure and mitigate bias. In real-world settings, however, such information is often unavailable or legally prohibited to infer due to privacy concerns, legal restrictions, or regulatory constraints, which substantially limits the applicability of these methods. To address this challenge, we propose Demographic-Independent Fair Graph Learning (DIFGL), a novel framework for fair graph learning without demographic supervision. DIFGL mitigates group unfairness by minimizing disparities in individual treatment across implicitly identified subgroups, thereby enforcing fairness without requiring explicit demographic information. Extensive experiments on benchmark datasets demonstrate that DIFGL achieves significant improvements in fairness while maintaining competitive predictive performance.

1 Introduction

With the increasing popularity of graph learning models for high-stakes decision-making, fairness has become an important topic in graph learning in recent years [Wang *et al.*, 2026a; Chen *et al.*, 2024; Zhang *et al.*, 2025]. Many efforts have been made to achieve fair graph learning, highlighting algorithmic decisions that neither discriminate against nor unfairly favor specific demographic groups [Wang *et al.*, 2025h]. However, most of these methods assume the complete availability of demographic information (*e.g.*, gender or race), which often does not hold in practice. In many real-world applications, explicit demographic information is unavailable due to privacy concerns, legal limitations, or ethical constraints that restrict the collection or use of demographic information [Ashurst and Weller, 2023]. For example, in employment decision-making scenarios, job applicants may hesitate or refuse to disclose demographic information because of fears of potential discrimination [Arvey, 1979]. Hence, existing fairness methods relying on explicit demographic information become infeasible.

To this end, several methods have started exploring fairness-aware graph models without explicit demographic information. Among these, a popular and well-studied approach involves inferring proxies for the missing demographic information, then treating these inferred proxies as if they were known demographic information for enforcing fairness constraints [Wang *et al.*, 2025i]. For example, fairGNN-WOD [Wang *et al.*, 2025e] addresses this issue through a two-stage framework, explicitly designed to first infer demographic proxies and then use these inferred attributes in downstream fairness-aware graph learning. However, such models overlook critical limitations of proxy inference. For example, inference of sensitive attributes might itself be restricted or prohibited under regulatory frameworks such as the EU General Data Protection Regulation (GDPR) [Voigt and Von dem Bussche, 2017], which explicitly limits or forbids processing of sensitive personal data like race, political opinions, religion, health status, biometrics, and sexual orientation. Moreover, inaccuracies or errors in the proxy estimation process may significantly compromise the reliability and effectiveness of fairness measurements, ultimately leading to unintended biases in downstream decisions.

To this end, it is essential to develop fairness-aware models that neither rely on explicit nor infer demographic information, thus avoiding regulatory conflicts and inaccuracies inherent to proxy-based approaches. Despite the importance of achieve fairness without direct demographic supervision, this remains a highly open research area with several unaddress and unique challenges: **i) Quantifying Bias without Explicit Demographic Information.** The critical challenge in scenarios where demographic inference is explicitly prohibited is to develop novel fairness measures and methodologies capable of comprehensively detecting and quantifying implicit biases across the entire spectrum of nodes with varying prediction costs, without relying on explicit demographic information or subgroup identification. **ii) Mitigating Bias without Demographic Supervision.** Existing fairness constraints depend on explicit demographic information to enforce fair-

*Corresponding author: wenbin.zhang@fiu.edu

ness. When demographic inference is explicitly prohibited, bias mitigation becomes even more challenging, as fairness constraints must be enforced solely based on observed prediction outcomes without referencing demographics at all. The core challenge in this setting is designing novel and efficient optimization strategies capable of reliably reducing implicit biases identified from predictive disparities alone, without incurring substantial losses in predictive performance. **iii) Meeting User-Specified Fairness Thresholds.** Ensuring that user-defined fairness constraints are reliably satisfied becomes particularly challenging without demographic supervision, since the absence of explicit group information makes fairness measurement and constraint-aware optimization nontrivial. Hence, designing mechanisms that can both verify fairness compliance and enforce fairness guarantees remains difficult in practical deployments.

To address these challenges, this paper studies a novel problem of fair graph learning without explicit or inferred demographics. We propose a novel framework, termed **Demographic-Independent Fair Graph Learning (DIFGL)**, specifically designed to enable fairness evaluation and mitigation by addressing treatment disparities without demographic supervision. *To best of our knowledge, this is the first work that achieve fair graph learning without relying on demographic supervision.* Specifically, we propose a novel fairness notion based on principled variance analysis, which identifies implicit biases by quantifying disparities in individual-level treatment. Building upon our proposed fairness notion, we develop a structured optimization framework grounded in the Rawlsian difference principle to systematically mitigate these disparities. Our proposed framework iteratively improves outcomes for implicitly identified least-advantaged nodes while ensuring no adverse impacts on other nodes. Additionally, we introduce a randomized optimization strategy explicitly designed to reduce node-level treatment variance, thus consistently maintaining compliance with predefined fairness threshold. Our contributions are outlined below:

- We propose a novel fairness notion that accurately identifies implicit biases without relying on explicit demographic information or inferred proxies.
- We design a novel optimization framework that builds on the fairness notion we proposed, explicitly targeting and reducing disparities among the least advantaged nodes through iterative variance-aware updates, thereby achieving graph fairness without demographic information or inferred proxies.
- We validate the effectiveness of our proposed methods through extensive experiments on three graph datasets, demonstrating significant improvements in fairness results without compromising model utility.

2 Related Work

2.1 Fairness in Graph Learning

Fairness has been widely studied in graph learning, and numerous fairness notions have been proposed in recent years [Wang *et al.*, 2026d; Wang and Zhang, 2025; Wang *et al.*, 2025c]. Among these, one of the most extensively studied concepts is group fairness, which emphasizes ensuring algorithmic decisions do not discriminate against or favor specific demographic subgroups based on demographic information like race or gender [Jiang *et al.*, 2024]. In other words, group fairness seeks to make the predictions of GNNs independent of demographic information [Wang *et al.*, 2023c; Wang *et al.*, 2025g; Wang and Zhang, 2024]. For example, FAHT [Zhang and Ntoutsi, 2019], an extension of the Hoeffding Tree algorithm, introduces fairness into online stream-based decision-making by accounting for the evolution of the underlying data population, mitigating discrimination as new instances arrive. Similarly, RFCGNN [Wang *et al.*, 2023a] addresses label bias in GNNs through counterfactual fairness, identifying authentic counterfactual samples within graph structures and applying causal analysis to mitigate bias in downstream predictions. However, these methods rely on the assumption that accurate demographic information is readily available to use, a condition that often does not hold in practical scenarios due to data collection constraints, privacy considerations, and regulatory requirements [Andrus *et al.*, 2021]. Given these requirements, there is an emerging trend in the research community to explore fairness methods without demographic information.

2.2 Fairness without Demographics in Graph Learning

There has been growing interest in achieving fair graph learning in settings where explicit demographic information is unavailable [Ashurst and Weller, 2023; Wang *et al.*, 2025h; Wang *et al.*, 2026b]. A common line of work addresses this problem by inferring proxies for missing demographic information and then treating these inferred proxies as substitutes to satisfy fairness constraints [Kenfack *et al.*, 2024]. For example, the approach proposed in [Wang *et al.*, 2025f] leverages partially available demographic information to infer the missing demographics, while simultaneously evaluating and optimizing model fairness during training. However, such methods still rely on the availability of partial demographic labels to guide the inference process, which limits their applicability. More recent studies have taken initial steps toward fair graph learning without assuming fully observed demographics. For instance, fairGNN-WOD [Wang *et al.*, 2025e] adopts a two-stage framework that first infers demographic information from node features and graph structure, and then applies these inferred proxies in downstream fairness-aware GNN training. Similarly, FairWOS [Wang *et al.*, 2025a] introduces pseudo-demographic information combined with graph counterfactual reasoning to improve fairness without demographics. Despite removing the need for ground-truth demographic information, these approaches still assume that inferring demographic information is permissible and reliable. In practice, however, such assumptions may not hold due to privacy concerns, legal restrictions, or regulatory prohibitions on demographics inference [Wang *et al.*, 2025i]. These limitations highlight a fundamental gap in existing work and underscore the need for fairness-aware graph learning methods that neither require explicit demographic information nor attempt to infer the proxy of missing demo-

graphics indirectly.

Unlike existing demographic-dependent and proxy-inference-based approaches, the proposed method adopts an individual-level treatment perspective to identify and mitigate bias without relying on explicit or inferred demographic information. In addition, a structured optimization framework grounded in the Rawlsian difference principle is developed to systematically improve outcomes for implicitly identified disadvantaged individuals while explicitly enforcing a predefined fairness threshold, yielding a principled and practically applicable solution for fair learning without demographic supervision.

3 Notation

In this paper, we consider an attributed graph $\mathcal{G}$, represented as ($\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$), consisting of a set of nodes ($\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$) and a set of undirected edges ($\mathcal{E} \subseteq v_i, v_j \mid v_i, v_j \in \mathcal{V}$). Each node v_i is characterized by a (d)-dimensional feature vector ($\mathbf{x}_i \in \mathbb{R}^d$), and collectively, these vectors form the node feature matrix ($\mathbf{X} \in \mathbb{R}^{n \times d}$). The connectivity between nodes is encoded by an adjacency matrix ($\mathbf{A} \in \{0,1\}^{n \times n}$), with ($\mathbf{A}_{i,j} = 1$) if an edge connects nodes v_i and v_j, and ($\mathbf{A}_{i,j} = 0$) otherwise. Additionally, each node (v_i) has an associated binary sensitive attribute ($s_i \in \{0,1\}$), collectively represented by the vector ($\mathbf{S} \in \{0,1\}^{n \times 1}$). This sensitive attribute partitions the node set into two groups: the deprived group ($S_d = \{v_i \mid s_i = 0\}$) (*e.g.*, female) and the favored group ($S_f = \{v_i \mid s_i = 1\}$) (*e.g.*, male). Each node v_i also carries a binary outcome label ($y_i \in \{0,1\}$) representing a granted decision outcome such as approval ($y_i = 1$) or rejected decision such as rejection ($y_i = 0$). The outcomes predicted by a model for each node are denoted as $\hat{y}_i$.

4 Methodology

4.1 Quantifying Group Unfairness without Demographics

Existing fairness approaches typically depend on explicit demographic information, either directly available or inferred, to identify and measure bias, thus limiting their applicability when demographic information is unavailable [Ashurst and Weller, 2023]. To this end, we propose a novel fairness notion that evaluates unfairness through the training perspective by examining disparities in model performance as reflected by training losses, without relying on demographics supervision. Specifically, the training loss reflects how accurately a model learns each individual, and systematic differences in these losses indicate consistent advantages afforded to certain individuals over others during the learning process. These persistent disparities can signal underlying structural or societal biases, even when the affected subsets are not explicitly labeled. By shifting evaluation away from final outcomes (*e.g.*, approval or rejection) toward analyzing consistency in training performance, this method enables fairness assessments independent of demographic groupings, offering detailed insights into model behavior. By shifting evaluation from final outcomes (such as approval or rejection) to consistency in training performance, this approach enables fairness assessment without relying on demographic groupings, while providing more fine-grained insights into model behavior. For instance, in a graph-based classifier predicting financial reliability without access to demographic information such as race, if a subset of nodes $\{v_1, \ldots, v_k\}$, possibly representing Black individuals, routinely experiences higher losses, this persistent disparity signals latent unfairness. To quantify this implicit subgroup unfairness without demographic information, the most direct approach is to measure how significantly the average loss among individuals with the highest losses deviates from the other population average. By capturing these disparities, it becomes possible to approximate implicit maximum subgroup-level unfairness, thereby providing a clear and systematic way to evaluate fairness even in the absence of explicit demographic information.

However, while conceptually intuitive, directly using the average loss of individuals with the highest training losses introduces practical limitations. In particular, this measurement uses the experience of these particular individuals as a proxy for group-level disparity, which may overlook the experiences of other individuals across the population. To address this limitation, we propose quantifying fairness using the empirical variance of training losses instead. The core idea is that variance, by accounting for the loss of every individual in the population, reflects how consistently the model performs across implicitly defined regions of the data and offers a more stable measure of latent unfairness. Moreover, variance is smooth and differentiable, enabling gradient-based optimization. Motivated by these properties, we define the *Latent Disparity Index* (LDI), a variance-based fairness measure that quantifies the extent to which the individuals with the highest training losses diverge from the rest of the population mean, providing a principled and practical framework for evaluating fairness without explicit demographic information. Formally, LDI is defined as follows:

Definition 4.3 (Latent Disparity Index). Given a graph learning model parameterized by θ, the latent disparity index quantifies implicit bias by measuring the extent to which the subset of individuals with the highest training losses deviates from the average loss of the remaining population:

$$\mathrm{LDI}(\theta) = \left| \frac{1}{k} \sum_{i \in \mathcal{Q}_k} \left(\ell(i;\theta) - \bar{\ell}(\theta)\right)^2 - \frac{1}{N-k} \sum_{i \in \mathcal{Q}_{\mathrm{rest}}} \left(\ell(i;\theta) - \bar{\ell}(\theta)\right)^2 \right| \quad (1)$$

where $\mathcal{Q}_k$ denotes the subset containing the k nodes with the highest training losses, and $\mathcal{Q}_{\mathrm{rest}}$ denotes the complementary subset with N denotes the total number of the nodes. In addition, $\ell(i;\theta)$ represents the training loss of one individual node i, while $\bar{\ell}(\theta)$ denotes the mean training loss over all nodes, defined as:

$$\bar{\ell}(\theta) = \frac{1}{|\mathcal{Q}_{\mathrm{train}}|} \sum_{i \in \mathcal{Q}_{\mathrm{train}}} \ell(i;\theta) \quad (2)$$

where $\mathcal{Q}_{\text{train}}$ denotes the empirical training set across all N nodes.

In summary, LDI provides a without demographic supervision mechanism to quantify latent disparity by contrasting the loss dispersion of the worst-off (top-k) nodes against the remaining population. Due to the top-k operator defining $\mathcal{Q}_k$ and the absolute-value operation, LDI inherently becomes non-smooth when node membership in $\mathcal{Q}_k$ changes or when the disparity gap reaches zero. To this end, in our framework, LDI primarily serves as a fairness criterion to be monitored or enforced at the objective or constraint level. To achieve stable optimization, we instead utilize node-level residual signals, defined as the standardized deviation of each node's loss from the global mean loss, for reweighting. These node-level residuals continuously vary with individual losses, enabling stable gradient-based updates. This method progressively reduces the underlying loss imbalance and, as a result, robustly decreases the LDI value throughout training iterations. This strategy effectively addresses the non-smoothness inherent in LDI, ensuring a practical and stable approach to fairness evaluation and enforcement.

4.2 Mitigating Unfairness without Explicit Demographics

Building on the proposed LDI, a novel framework is introduced to systematically mitigate bias without requiring explicit demographic information. The core idea is inspired by the Rawlsian difference principle [Rawls, 1971], which aims to improve the outcomes of the least advantaged group without reducing the welfare of others. In this framework, welfare is represented by individual training losses, with disparity measured through variations in these losses. Under the absence of explicit demographic information, the "least advantaged group" is naturally defined as the top-k highest-loss nodes identified by LDI.

Guided by this principle, we develop the *Demographic-Independent Fair Graph Learning (DIFGL)* framework, which implements Rawlsian fairness through two iterative steps. First, at each training iteration, nodes experiencing the greatest disadvantage are identified based on their individual training losses. This criterion directly leverages training loss as a proxy signal for implicit disadvantage in settings where explicit demographic information is unavailable. Second, after this subset of high-loss nodes is identified, the training process explicitly prioritizes these nodes, guiding model updates to reduce their losses toward the overall population average and thereby shrinking the LDI-based disparity. Since the composition of the least advantaged group may change dynamically during training, this identification-and-correction procedure is repeated at each iteration. Specifically, at iteration t, the identification step is implemented by ranking training nodes according to their current loss values and selecting the top-k nodes with the highest losses. Formally, an ordering π_θ over the training nodes is defined as follows:

$$\ell\left(v_i^1;\theta_i\right) \;\geq\; \ell\left(v_i^2;\theta_i\right) \;\geq\; \cdots \;\geq\; \ell\left(v_i^N;\theta_i\right) \tag{3}$$

where $\mathcal{V}_{\text{train}} \subseteq \mathcal{V}$ denotes the set of training nodes with cardinality $N = |\mathcal{V}_{\text{train}}|$, while $\ell(v_i;\theta)$ represents the training loss of node v_i under parameters θ.

Using this sorted order, we dynamically define the least advantaged subset and advantaged subset as:

$$\mathcal{V}_{dis}(\theta) = \left\{v_i^1, \ldots, v_i^k\right\},\ \mathcal{V}_{\text{adv}}(\theta) = \mathcal{V}_{\text{train}} \setminus \mathcal{V}_{dis}(\theta) \tag{4}$$

This subset, $\mathcal{V}_{dis}(\theta)$, is re-identified after every model update, ensuring that the framework consistently targets the nodes currently most disadvantaged under the latest model parameters.

Armed with the identification of the least advantaged group, the next step is to translate this fairness criterion into a stable, node-level update rule. Specifically, while LDI effectively signals the existence of disparities, it remains a global scalar measure, offering no explicit indication of which individual nodes require correction or the extent to which each node should influence subsequent updates. To this end, we propose a two-stage mechanism that explicitly bridges the gap between global fairness signals and actionable node-level guidance. First, a diagnostic component (the separating oracle) translates the overall loss disparity into individual residual scores, clearly indicating how much loss is caused if each node deviates from the population mean. Second, a corrective component (the weighted optimization oracle) transforms these residual scores into training weights, systematically assigning greater emphasis to nodes with above-average losses, thus directly reducing their disadvantage. Guided by this mechanism, DIFGL maintains and iteratively updates a distribution over candidate models, systematically reducing loss variability across all nodes. By repeatedly applying this diagnostic and corrective process, DIFGL ensures that no subgroup consistently experiences disproportionately higher losses and progressively reduces the disparities captured by LDI, resulting in stable and sustained fairness improvements throughout training. Specifically, we first introduce the separate oracle and formally define how it computes the residual scores. Formally, the separating oracle outputs the residual vector, which captures node-level deviations from the global average loss as follows:

$$r_i(\theta) \;=\; \frac{\ell(v_i;\theta) - \bar{\ell}(\theta)}{\sigma_\ell(\theta) + \epsilon} \tag{5}$$

where $\epsilon > 0$ ensures numerical stability. To clarify why the residual is an appropriate node-level signal for reducing dispersion, we view the empirical variance as a function of the loss vector. Let $\ell_i = \ell(v_i;\theta)$ and $\mu = \bar{\ell}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell_i$, so that:

$$\sigma_\ell^2(\theta) = \frac{1}{N}\sum_{i=1}^{N}(\ell_i - \mu)^2 \tag{6}$$

Differentiating $\sigma_\ell^2(\theta)$ with respect to the i^{th} loss component yields:

$$\frac{\partial}{\partial\, \ell(v_i;\theta)}\, \sigma_\ell^2(\theta) \;=\; \frac{2}{N}\big(\ell(v_i;\theta) - \bar{\ell}(\theta)\big) \tag{7}$$

which shows that the centered loss $\ell(v_i;\theta) - \bar{\ell}(\theta)$ is exactly the direction that increases the variance most. The residual

$r_i(\theta)$ is a normalized version of this gradient component (up to a global scaling factor), since:

$$r_i(\theta) = \frac{\ell(v_i;\theta) - \bar{\ell}(\theta)}{\sigma_\ell(\theta) + \epsilon} \;\propto\; \frac{\partial}{\partial\, \ell(v_i;\theta)} \sigma_\ell^2(\theta) \tag{8}$$

Therefore, if we view $\sigma_\ell^2(\theta)$ as a function of the loss vector $\big(\ell(v_1;\theta), \ldots, \ell(v_N;\theta)\big)$, then decreasing the losses of nodes with positive residuals (that is, nodes whose losses lie above the mean) reduces σ_ℓ^2 at the first-order level, making the loss profile more uniform across nodes. Up to this point, we have described the deterministic variant that maintains a single parameter vector θ. By doing so, DIFGL can be implemented in a deterministic form by updating a single parameter vector $\theta \leftarrow \theta^+$. To avoid abruptly replacing the current model after each oracle call, we maintain a mixture state over candidate parameters. Concretely, we represent the current state by a distribution D and evaluate each node by its expected loss:

$$\ell(v_i; D) \;=\; \mathbb{E}_{\theta\sim D}\, \ell(v_i;\theta) \tag{9}$$

Given the residual vector produced by the separating oracle, the weighted optimization oracle performs the second step of the framework by translating these residuals into a concrete model update. Specifically, we update the state by a convex mixture $D_{\text{new}} = (1-\gamma)D_{\text{old}} + \gamma\, \delta_{\theta^+}$, where the mixing rate γ controls how much probability mass is assigned to θ^+ (smaller γ yields a more gradual update). This distributional form includes the single-model case as a special instance: if $D = \delta_\theta$, then $\ell(v_i; D) = \ell(v_i;\theta)$. After receiving the residual vector from the separating oracle, the weighted oracle reduces the loss disparity by solving a weighted training problem that assigns larger influence to nodes whose losses are above the population level. With step size $\eta > 0$ and an optional regularization $\mathcal{R}(\theta)$, we compute:

$$\theta^+ \;=\; \arg\min_\theta \; \frac{1}{N} \sum_{i=1}^{N} \Big(1 + \eta\, r_i\Big)\, \ell(v_i;\theta) + \lambda_{\text{reg}}\, \mathcal{R}(\theta) \tag{10}$$

where r_i can be $r_i(\theta)$ for a single model or $r_i(D)$ for randomized distributions. Notably, we use $w_i = \text{clip}(1 + \eta r_i, w_{\min}, w_{\max})$ to ensure nonnegative, bounded weights. Minimizing $\frac{1}{N}\sum_i w_i\, \ell(v_i;\theta) + \lambda_{\text{reg}}\mathcal{R}(\theta)$ aligns the update direction with $-\nabla\sigma_\ell^2$ at the loss level, systematically reducing node loss variance.

Intuitively, because $r_i > 0$ indicates that node v_i lies above the global mean loss, upweighting such nodes increases the training pressure to reduce their losses, which in turn makes the loss distribution more uniform across nodes. In DIFGL, this idea is applied to the current model state in a gradual manner by maintaining a distribution D_t over candidate parameters, so that node-level quantities are evaluated in expectation rather than relying on a single point estimate. Accordingly, at iteration t the separating oracle first computes each node's expected training loss under D_t, and then derives the population mean and dispersion from these expected losses:

$$\sigma_\ell(D_t) \;=\; \sqrt{\frac{1}{N}\sum_{i=1}^{N} \big(\ell(v_i; D_t) - \bar{\ell}(D_t)\big)^2} \tag{11}$$

where $\bar{\ell}(D_t) \;=\; \frac{1}{N}\sum_{i=1}^{N} \ell(v_i; D_t)$, and $N = |\mathcal{V}_{\text{train}}|$. The separating oracle then recalculates node-level residuals from $\{\ell(v_i; D_t)\}_{i=1}^N$, providing a quantitative assessment of how each node deviates from the global mean. Subsequently, the weighted oracle uses these residual scores to perform a reweighted training step and returns a new candidate parameter vector θ^+.

To update the model state in a controlled and gradual manner, DIFGL does not fully replace the current state with the new candidate. Instead, it mixes θ^+ into the current distribution using a tunable mixing rate:

$$D_{t+1} \;\leftarrow\; (1-\gamma_t)\, D_t \;+\; \gamma_t\, \delta_{\theta^+} \tag{12}$$

where $\gamma_t \in (0, 1]$ controls how aggressively the state moves toward θ^+ at iteration t. In practice, we choose γ_t by a short line search along the mixture path $D(\gamma) = (1-\gamma)D_t + \gamma\delta_{\theta^+}$, selecting γ_t so that the variance of the expected losses does not increase:

$$\sigma_\ell^2(D_{t+1}) \;=\; \sigma_\ell^2\big(D(\gamma_t)\big) \;\le\; \sigma_\ell^2\big(D(\gamma=0)\big) \;=\; \sigma_\ell^2(D_t) \tag{13}$$

After updating D_{t+1}, we recompute $\ell(v_i; D_{t+1})$ and the residual vector and repeat the loop.

Finally, we stop once the dispersion is small enough to imply the target LDI threshold. Let $\alpha = k/N$ for the top-k split used in LDI. The mean-loss gap between the top-k set and the remaining set admits the bound:

Therefore, a sufficient stopping condition for ensuring $\text{LDI}(\cdot) \le \varepsilon$ is:

$$\sigma_\ell(\cdot) \;\le\; \varepsilon\, \sqrt{\alpha(1-\alpha)} \tag{14}$$

Under Equation (14), the resulting model state satisfies the LDI-based fairness requirement in Definition 4.3.

4.3 Overall Optimization Objectives

In this subsection, we introduce our overall training objective that jointly considers predictive accuracy and fairness criteria. Specifically, our optimization framework involves two key components: i) predictive performance evaluated through the binary cross-entropy loss on labeled nodes, and ii) fairness constraints measured by disparities between the top-k highest-loss nodes and the rest of the training population. We define the predictive accuracy component using the standard binary cross-entropy loss across labeled training nodes $\mathcal{V}_L$:

$$\mathcal{L}_P = \frac{1}{|\mathcal{V}_L|} \sum_{v_i \in \mathcal{V}_L} -\Big(y_{v_i} \log(\hat{y}_{v_i}) + (1 - y_{v_i}) \log(1 - \hat{y}_{v_i})\Big) \tag{15}$$

In addition, we then formalize a fairness constraint based on the proposed LDI measure. Mathematically, this is represented as:

$$\mathcal{L}_F = \Big[\Big|\frac{1}{k}\sum_{i\in\mathcal{Q}_k} \ell(i;\theta) - \frac{1}{N-k}\sum_{i\in\mathcal{Q}_{\text{rest}}} \ell(i;\theta)\Big| - \varepsilon\Big] \tag{16}$$

Finally, the overall training objective combines predictive loss and fairness loss as:

Dataset	Metrics \ Methods	GCN	GIN	KSMOTE	Reckoner	Themis	fairGNN-WOD	DIFGL
Credit	**AUC** (↑)	0.681 ± 0.056	**0.707 ± 0.048**	0.656 ± 0.059	0.668 ± 0.021	0.675 ± 0.032	0.680 ± 0.052	0.701 ± 0.052
	F1-Score (↑)	0.794 ± 0.023	**0.807 ± 0.028**	0.755 ± 0.052	0.763 ± 0.045	0.760 ± 0.048	0.785 ± 0.027	0.790± 0.038
	$\Delta_{DP}(\downarrow)$	0.118 ± 0.013	0.147 ± 0.030	0.073 ± 0.043	0.068 ± 0.027	0.049 ± 0.021	0.041 ± 0.021	**0.038 ± 0.020**
	$\Delta_{EO}(\downarrow)$	0.098 ± 0.027	0.089 ± 0.023	0.065 ± 0.023	0.055 ± 0.014	0.042 ± 0.023	0.038 ± 0.023	**0.033 ± 0.013**
German	**AUC** (↑)	**0.746 ± 0.045**	0.728 ± 0.030	0.679±0.034	0.662 ± 0.051	0.713 ± 0.041	0.728 ± 0.043	0.733 ± 0.021
	F1-Score (↑)	0.807 ± 0.032	**0.823 ± 0.027**	0.775 ± 0.048	0.738 ± 0.046	0.747 ± 0.032	0.752 ± 0.042	0.783 ± 0.032
	$\Delta_{DP}(\downarrow)$	0.365 ± 0.032	0.159 ± 0.046	0.086± 0.027	0.077 ± 0.018	0.071 ± 0.037	0.063 ± 0.017	**0.057 ± 0.013**
	$\Delta_{EO}(\downarrow)$	0.315 ± 0.041	0.097 ± 0.037	0.062 ± 0.020	0.055 ± 0.018	0.040 ± 0.013	0.044 ± 0.013	**0.038 ± 0.015**
NBA	**AUC** (↑)	**0.691 ± 0.035**	0.683 ± 0.048	0.680 ± 0.053	0.681 ± 0.028	0.679 ± 0.054	0.678 ± 0.024	0.681 ± 0.024
	F1-Score (↑)	**0.631 ± 0.022**	0.629 ± 0.042	0.624 ± 0.019	0.621 ± 0.032	0.616 ± 0.029	0.626 ± 0.029	0.626 ± 0.029
	$\Delta_{DP}(\downarrow)$	0.081 ± 0.033	0.075 ± 0.017	0.032 ± 0.025	0.027 ± 0.018	0.031 ± 0.019	0.023 ± 0.008	**0.021 ± 0.023**
	$\Delta_{EO}(\downarrow)$	0.156 ± 0.049	0.132 ± 0.057	0.039 ± 0.020	0.033 ± 0.011	0.038 ± 0.025	0.034 ± 0.017	**0.031 ± 0.016**

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

$$\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_P + a\,\mathcal{L}_F \tag{17}$$

where a controls the balance between predictive performance and the strength of fairness enforcement.

5 Experiment

5.1 Datasets

Our experiments are conducted on two widely used datasets:, namely Credit, German, and NBA. The **Credit** dataset [Yeh and Lien, 2009] consists of credit card holders represented as nodes, with edges capturing similarities in their purchasing and repayment behaviors. Age is considered the sensitive attribute, and the task aims to determine whether a cardholder is likely to default on future payments. The **German** dataset [Asuncion and Newman, 2007] models clients of a German financial institution as nodes, where edges are constructed between individuals exhibiting similar credit account characteristics. In this dataset, gender is treated as the sensitive attribute, and the learning objective is to predict whether a client presents a high or low credit risk. The **NBA** dataset [Dai and Wang, 2021] is professional basketball players from the 2016–2017 season as nodes, while edges denote social interactions among players on Twitter. Nationality, categorized as American or overseas, serves as the sensitive attribute, and the classification task focuses on predicting whether a player's salary is above the median level.

5.2 Baselines

We evaluate the proposed DIFGL method against a set of representative baseline approaches, which are grouped into two categories. i) Vanilla graph models: **GCN** [Kipf and Welling, 2016] and **GIN** [Xu *et al.*, 2018]. ii) Fairness-enhanced methods: **KSMOTE** [Yan *et al.*, 2020], which constructs pseudo-groups through clustering and enforces fairness using prediction regularization; **Reckoner** [Ni *et al.*, 2024], which promotes group fairness through learnable noise injection and knowledge sharing in a dual-model framework; **Themis** [Wang *et al.*, 2025d], which addresses fair graph learning without demographic information by inferring demographic proxies via a Bayesian variational autoencoder and enforcing fairness through disentangled latent; and **fairGNN-WOD** [Wang *et al.*, 2025e], which mitigate bias without demographics by leveraging disentanglement learning to separate sensitive and task-related information. For approaches not originally developed for graph-structured data, we adapt their implementations to operate with our GNN backbone following the authors' original designs.

5.3 Evaluation Metrics

We adopt two predictive performance metrics, namely AUC and F1-score, with higher scores indicating better predictive accuracy. For fairness evaluation, we utilize two commonly used metrics: Demographic Parity (Δ_{DP}) [Dwork *et al.*, 2012] and Equalized Odds (Δ_{EO}) [Hardt *et al.*, 2016], where values approaching zero represent greater fairness.

5.4 Experiment Results

Comparison Study. Table 1 presents a detailed comparison of DIFGL against six baseline methods on three benchmark datasets, evaluating both predictive performance metrics and fairness metrics. We first observed that DIFGL consistently achieves the best fairness outcomes across all datasets, reflecting lower values of fairness metrics compared to baseline methods. This improvement in fairness arises from DIFGL's explicit strategy of targeting the highest-loss nodes, identified via the LDI perspective, and incrementally adjusting their weights during training. In contrast, standard GNN methods, such as GCN and GIN, optimize only average predictive performance, often leaving persistent disparities among certain node subpopulations. Data-level approaches (*e.g.*, KSMOTE) attempt to mitigate disparity by altering the training data distribution, but this can cannot guarantee a balanced representation across different subgroups, nor can they ensure independence between demographic attributes and model decisions. Similarly, approaches based on demographic inference may suffer from inaccuracies in inferred demographic labels, preventing them from accurately enforcing fairness constraints at the group level. In addition, we also observed that DIFGL reliably maintains competitive predictive performance, closely matching or slightly trailing behind the top predictive backbones, while substantially improving fairness. This is due to DIFGL's incremental correction mechanism guided by stable node-level residual signals, which avoids drastic parameter updates and thus preserves predictive accuracy. Overall,

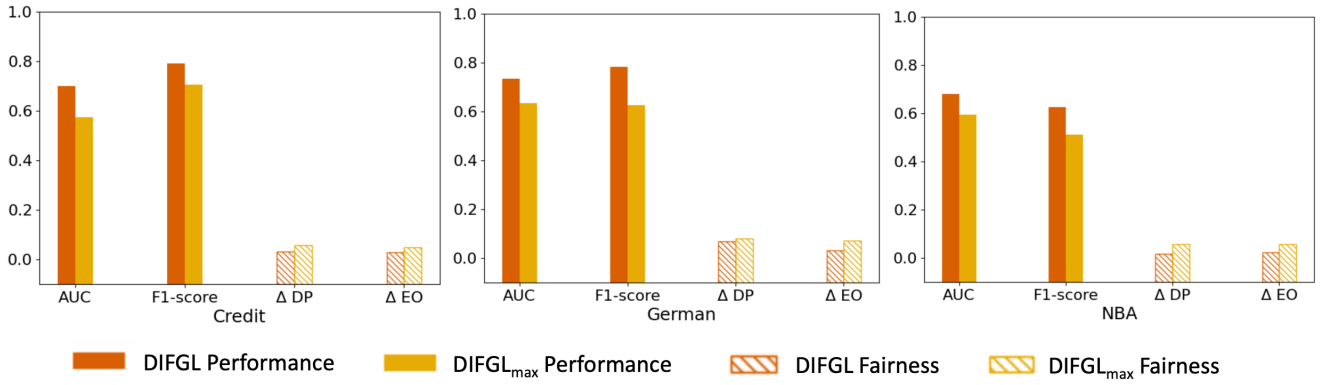

Figure 1: Ablation study results on the on the Credit, German and NBA dataset.

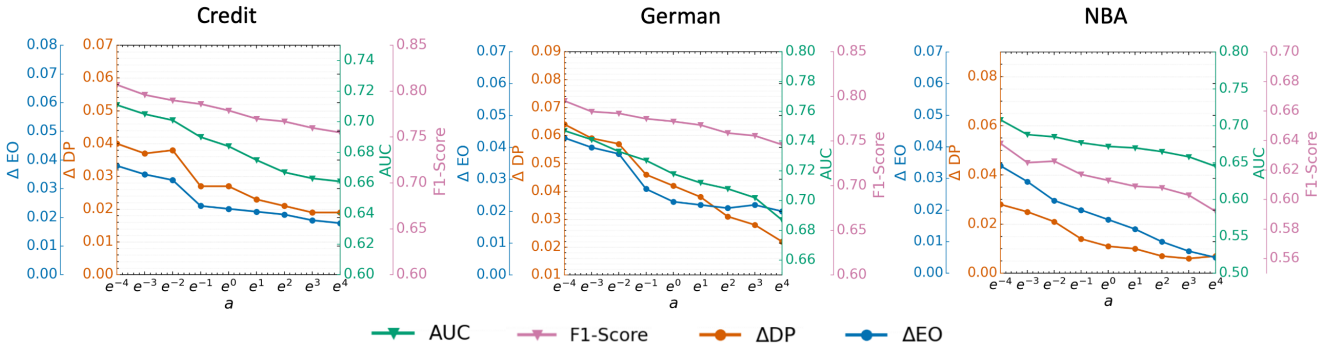

Figure 2: Parameter sensitivity analysis of hyperparameters on the Credit, German and NBA dataset.

DIFGL consistently achieves a balanced trade-off between fairness and predictive performance, demonstrating its effectiveness in practical scenarios where explicit demographic information is unavailable or restricted.

Ablation Study. We conduct an ablation study to gain deeper insights into the design and effectiveness of the LDI in DIFGL. Specifically, we compare DIFGL against a variant (denoted as $\mathrm{DIFGL}_{\max}$) that directly defines model bias as the maximum individual loss gap, representing a stricter, worst-case definition of latent disparity. Figure 1 shows the results, indicating that the $\mathrm{DIFGL}_{\max}$ variant yields both worse predictive performance and weaker fairness outcomes compared to the original DIFGL. This occurs because the max–min loss range used by $\mathrm{LDI}_{\max}$ is dominated by only a few extreme nodes, and as training progresses, the identities of these nodes (maximizer and minimizer) often change abruptly. Such instability generates unreliable update signals, causing the model to chase after outlier nodes rather than systematically enhancing outcomes for the genuinely least-advantaged subset. Moreover, reducing $\mathrm{LDI}_{\max}$ may involve overly suppressing the loss of the highest-loss node or artificially inflating the lowest-loss node, neither of which necessarily decreases the average disparity experienced by a meaningful fraction of the population. Hence, DIFGL with the original LDI effectively maintains a better balance between predictive utility and fairness compared to the $\mathrm{DIFGL}_{\max}$.

Parameters Sensitivity. We investigate DIFGL's sensitivity to the fairness–utility trade-off parameter a, varying it across the set $\{e^{-4}, e^{-3}, \cdots, e^{4}\}$. The results are presented in Figure 2. We observe that increasing a places greater emphasis on minimizing the LDI-based disparity, typically resulting in improved fairness metrics. However, this improvement often comes with a trade-off in utility, as the model focuses more heavily on mitigate bias, potentially diverting capacity from overall prediction accuracy. Conversely, smaller values of a push the training closer to standard, accuracy-driven learning, which generally produces higher predictive performance but offers weaker fairness.

6 Conclusion

Given the observed gap between the assumptions made by existing AI fairness methods regarding the availability of demographic information and the reality of practical constraints in real-world applications, we introduced a novel framework for fair graph learning without explicit or inferred demographic information. Furthermore, we developed a randomized optimization strategy explicitly targeting variance reduction in node-level losses, ensuring consistent compliance with fairness threshold. Extensive experiments confirmed our framework's effectiveness, showing significant fairness gains without major drops in predictive accuracy. Our work thus provides a practical solution for achieving fairness in graph learning scenarios constrained by privacy or regulatory limitations, paving the way for future research in fairness without demographic supervision.

References

[Andrus *et al.*, 2021] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 249–260, 2021.

[Arvey, 1979] Richard D Arvey. Unfair discrimination in the employment interview: Legal and psychological aspects. *Psychological bulletin*, 86(4):736, 1979.

[Ashurst and Weller, 2023] Carolyn Ashurst and Adrian Weller. 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, 2023.

[Asuncion and Newman, 2007] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

[Chen *et al.*, 2024] April Chen, Ryan A Rossi, Namyong Park, Puja Trivedi, Yu Wang, Tong Yu, Sungchul Kim, Franck Dernoncourt, and Nesreen K Ahmed. Fairness-aware graph neural networks: A survey. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–23, 2024.

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

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

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

[Jiang *et al.*, 2024] Zhimeng Jiang, Xiaotian Han, Chao Fan, Zirui Liu, Na Zou, Ali Mostafavi, and Xia Hu. Chasing fairness in graphs: A gnn architecture perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21214–21222, 2024.

[Kenfack *et al.*, 2024] Patrik Joslin Kenfack, Samira Ebrahimi Kahou, and Ulrich Aïvodji. A survey on fairness without demographics. *Transactions on Machine Learning Research*, 2024.

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

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

[Rawls, 1971] A Rawls. Theories of social justice, 1971.

[Voigt and Von dem Bussche, 2017] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[Wang and Zhang, 2024] Zichong Wang and Wenbin Zhang. Group fairness with individual and censorship constraints. In *Proceedings of the 27th European Conference on Artificial Intelligence*, pages 906–913. IOS Press, 2024.

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

[Wang *et al.*, 2023a] Zichong Wang, Giri Narasimhan, Xin Yao, and Wenbin Zhang. Mitigating multisource biases in graph neural networks via real counterfactual samples. In *ICDM 2023*, 2023.

[Wang *et al.*, 2023b] Zichong Wang, Nripsuta Saxena, Tongjia Yu, Sneha Karki, Tyler Zetty, Israat Haque, Shan Zhou, Dukka Kc, Ian Stockwell, Xuyu Wang, et al. Preventing discriminatory decision-making in evolving data streams. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 149–159, 2023.

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

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

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

[Wang *et al.*, 2025a] Xuemin Wang, Tianlong Gu, Xuguang Bao, and Liang Chang. Towards fair graph neural networks via graph counterfactual without sensitive attributes. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pages 265–277. IEEE, 2025.

[Wang *et al.*, 2025b] Zichong Wang, Zhibo Chu, Thang Viet Doan, Shiwen Ni, Min Yang, and Wenbin Zhang. History, development, and principles of large language models: an introductory survey. *AI and Ethics*, 5(3):1955–1971, 2025.

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

[Wang *et al.*, 2025d] Zichong Wang, Nhat Hoang, Xingyu Zhang, Kevin Bello, Xiangliang Zhang, Sundararaja Sitharama Iyengar, and Wenbin Zhang. Towards fair graph learning without demographic information. In *The 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pages 2107–2115, 2025.

[Wang *et al.*, 2025e] Zichong Wang, Fang Liu, Shimei Pan, Jun Liu, Fahad Saeed, Meikang Qiu, and Wenbin Zhang. fairgnn-wod: Fair graph learning without complete demographics. pages 556–564, 8 2025. Main Track.

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

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

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

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

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

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

[Wang *et al.*, 2026c] Zichong Wang, Zhipeng Yin, Zhong Chen, Jack Yang, Jun Liu, and Wenbin Zhang. Learning counterfactual fairness from authentic generation. In *Proceedings of the 35th International Joint Conference on Artificial Intelligence*, 2026.

[Wang *et al.*, 2026d] Zichong Wang, Zhipeng Yin, and Wenbin Zhang. A unified framework for fair graph generation: Theoretical guarantees and empirical advances. *Advances in Neural Information Processing Systems*, 38:64883–64909, 2026.

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

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

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

[Yin *et al.*, 2025] Zhipeng Yin, Zichong Wang, Avash Palikhe, Zhen Liu, Jun Liu, and Wenbin Zhang. Amcr: A framework for assessing and mitigating copyright risks in generative models. *28th European Conference on Artificial Intelligence*, 2025.

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

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

[Zhang and Weiss, 2021] Wenbin Zhang and Jeremy C Weiss. Fair decision-making under uncertainty. In *2021 IEEE international conference on data mining (ICDM)*, pages 886–895. IEEE, 2021.

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

[Zhang *et al.*, 2021] Wenbin Zhang, Liming Zhang, Dieter Pfoser, and Liang Zhao. Disentangled dynamic graph deep generation. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 738–746. SIAM, 2021.

[Zhang *et al.*, 2023] Wenbin Zhang, Tina Hernandez-Boussard, and Jeremy Weiss. Censored fairness through awareness. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 14611–14619, 2023.

[Zhang *et al.*, 2025] Wenbin Zhang, Shuigeng Zhou, Toby Walsh, and Jeremy C Weiss. Fairness amidst non-iid graph data: A literature review. *AI Magazine*, 46(1):e12212, 2025.

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

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