# Fair Graph U-Net: A Fair Graph Learning Framework Integrating Group and Individual Awareness

**Zichong Wang**[1], **Zhibo Chu**[1], **Thang Viet Doan**[1], **Shaowei Wang**[2], **Yongkai Wu**[3], **Vasile Palade**[4], **Wenbin Zhang**[1*]

[1] Florida International University, Florida, U.S.A.
[2] University of Manitoba, Manitoba, Canada
[3] Clemson University, South Carolina, U.S.A.
[4] Coventry University, Coventry, U.K.

## Abstract

Learning high-level representations for graphs is crucial for tasks like node classification, where graph pooling aggregates node features to provide a holistic view that enhances predictive performance. Despite numerous methods that have been proposed in this promising and rapidly developing research field, most efforts to generalize the pooling operation to graphs are primarily performance-driven, with fairness issues largely overlooked: i) the process of graph pooling could exacerbate disparities in distribution among various subgroups; ii) the resultant graph structure augmentation may inadvertently strengthen intra-group connectivity, leading to unintended inter-group isolation. To this end, this paper extends the initial effort on fair graph pooling to the development of fair graph neural networks, while also providing a unified framework to collectively address group and individual graph fairness. Our experimental evaluations on multiple datasets demonstrate that the proposed method not only outperforms state-of-the-art baselines in terms of fairness but also achieves comparable predictive performance.

## Introduction

Graph Neural Networks (GNNs) have shown superior performance in modeling graph-structured data in various domains, such as social network analysis (Wang et al. 2023d), recommendation system (Ying et al. 2018a), and urban computing (Bao et al. 2017). These models, which are extensions of Convolutional Neural Networks (CNNs) (Lee, Lee, and Kang 2019), are specifically adapted to address the unique challenges of non-Euclidean data that lacks spatial locality and a fixed order (Gao and Ji 2019). A key application of GNNs is graph classification, which involves algorithmically identifying class labels for subgraphs (*i.e.*, the label of the central node). This task requires holistic graph-level representations that can handle graphs of varying sizes and topologies, underscoring the critical role of the pooling mechanism in the process. To this end, various graph pooling designs have been developed to create effective and meaningful graph representations. For example, SAGPool (Lee, Lee, and Kang 2019) utilizes a self-attention mechanism to determine which nodes to retain or discard. Similarly, Graph U-net (Gao and Ji 2019) selects the top $k$ most important nodes to form a more compact graph while preserving crucial information.

Despite their effectiveness, the existing pooling methods can inadvertently introduce biases against subgroups defined by *sensitive attributes* such as race, gender, and age. Specifically, pooling layers, driven by performance optimization, typically remove nodes to downsample representations, avoiding the introduction of noise into the computed features, especially since graphs are frequently represented sparsely. This process, however, often disproportionately affects nodes from various subgroups, leading to skewed representations in the downsampled subgraphs. As a result, the labels of these subgraphs, represented by their central node labels, may become excessively correlated with sensitive attributes, embedding biases directly into the model outputs and potentially resulting in discriminatory decisions against the center node. Such biases can severely limit the adoption of GNNs in high-stake decision-making scenarios, such as job screening (Wang et al. 2023c), healthcare (Zhang et al. 2023) and criminology (Zhang and Weiss 2022). For instance, in the study of interaction networks of complex diseases, researchers have identified specific subnetworks that are associated with the disease (Cho, Kim, and Przytycka 2012). In such cases, differences in the representation of various groups in subgraphs can lead to an overcorrelation of diseases with sensitive attributes, which can affect the subsequent allocation of social welfare healthcare resources. Moreover, another critical limitation of existing pooling-based graph methods is their neglect of graph structure bias (*i.e.*, excessively linking nodes with similar sensitive attribute values). When nodes are eliminated during pooling, associated edges are removed, potentially resulting in isolated nodes. To counteract this and maintain information propagation in subsequent layers, existing methods employ an edge generation mechanism that increases the connectivity of the graph (Gao and Ji 2019). However, this strategy can unintentionally introduce graph structural bias as nodes sharing sensitive attribute values are often more frequently connected within graphs.

On the other hand, a number of research approaches have been proposed to address bias and discrimination in graph learning (Wang et al. 2024c), generally categorized into group fairness (Mehrabi et al. 2021), aiming for

*Corresponding author.

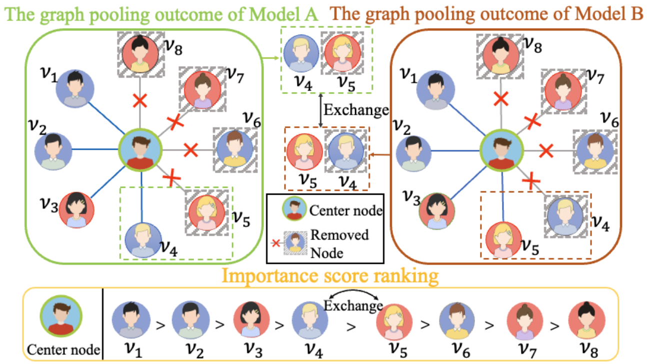


Figure 1: An illustrative toy example showcases individual unfairness while maintaining group fairness in pooling operation, with relevance scores decreasing from left to right.

statistical equality among subgroups, and individual fairness (Zhang 2024a; Zhang, Hernandez-Boussard, and Weiss 2023; Zhang 2024b), which ensures that similar individuals receive comparable treatment. However, harnessing their fair graph learning capability to capitalize on the strong graph learning capability of graph pooling is challenging due to the distinctive downsampling and upsampling procedures involved. In addition, although group fairness has been widely studied (Kleindessner et al. 2019; Wang et al. 2024b, 2023b), it often ignores the nuanced impacts on individual-level fairness within subgroups (Wang et al. 2024d). Specifically, existing group fairness works focusing on statistical equality between *deprived group* (*e.g.*, male) and *favored group* (*e.g.*, female) can inadvertently neglect the experiences of specific individuals within these groups (Zhang and Ntoutsi 2019; Wang and Zhang 2024; Chu, Wang, and Zhang 2024). In the scenario depicted in Figure 1, where gender is the sensitive attribute, nodes $v_1$ to $v_8$ are ranked in descending order by their importance scores to the central node of their subgraph, and graph pooling selects the top half to create a new graph. In this scenario, Model 'A' is optimized for performance, while Model 'B' also enforces group fairness (*i.e.*, ensuring that the parity in the ratio of males to females in the newly formed subgraph is maintained). However, while model 'B' achieves group fairness, it may disproportionately impact an individual, such as $v_4$, because he is the only one in a worse position in the prediction of model 'A' than in model 'B' despite other solutions being possible. In other words, the use of group fairness constraints can lead to a decrease in individual fairness, where certain individuals are unfairly disadvantaged.

To address these limitations, this paper pursues fair pooling with group and individual awareness for the development of fair GNNs, which is largely unexplored and poses distinct challenges: **i) Fairness-aware graph pooling.** Graph pooling involves unique downsampling and upsampling processes, and the bias during these processes cannot be directly constrained by the loss function alone. A unified design that incorporates fairness considerations is therefore necessary. **ii) Mitigating multiple sources of bias.** Graph data exhibits diverse biases stemming from various sources, including inherent biases in node attributes and biases embedded within the structural relationships. A holistic approach to address these biases simultaneously is thus needed. **iii) Jointly addressing individual and group graph fairness.** Existing fair graph research primarily focuses on either group or individual fairness, overlooking the interplay between them. Addressing both aspects simultaneously demands innovative approaches that can navigate the nuances associated with each type of fairness concurrently.

To this end, this paper introduces a novel framework, *Fair Graph U-net*, which, *to the best of our knowledge, is the first to explore fairness in graph pooling while reconciling both group and individual fairness*. Specifically, Fair Graph U-net ensures a consistent representation of diverse subgroups in the devised graph pooling process by equitably selecting a balanced number of nodes from various subgroups. Moreover, to counter potential individual-level unfairness (*i.e.*, individuals disproportionately bearing the losses due to the enforcement of group fairness constraints) that arises from enforcing group fairness, a probabilistic distribution of valid rankings is implemented. This implementation, inspired by distributive justice theory, ensures a uniform distribution of individual fairness loss, leading to consistent and fair treatment at the individual level. Furthermore, Fair Graph U-net incorporates fairness constraints when enhancing graph connectivity to avoid introducing graph structural bias. This is designed to prevent nodes from the same subgroups from being excessively connected, thereby averting introducing graph structure bias. The main contributions are:

- **Problem.** We investigate a novel problem in fair graph pooling to extend the powerful generation capabilities of convolutional pooling while addressing inherent biases present in the pooling process.
- **Method.** We propose a novel framework, Fair Graph U-net, designed to maintain a consistent representation for each subgroup and ensure that each individual bears a uniform group fairness loss, thus promoting individual fairness while enforcing group fairness. Additionally, our approach avoids introducing bias in the graph structure.
- **Evaluations.** Extensive empirical evaluations on three real-world graph datasets show that Fair Graph U-net surpasses existing baselines across multiple fairness metrics while also maintaining comparable prediction performance for downstream tasks.

## Related Work

**Graph Pooling.** Graph Neural Networks (GNNs) have emerged as a powerful tool for various tasks involving graph-structured data (Zhang et al. 2024). They have shown remarkable success in both node-level and graph-level tasks, including node classification (Kipf and Welling 2016; Veličković et al. 2017), graph classification (Sui et al. 2022; Gao and Ji 2019; Lee, Lee, and Kang 2019), and graph generation (Rahman et al. 2019; Zhao et al. 2022). Among the techniques central to the success of GNNs, graph pooling stands out as a crucial computational feature that significantly enhances performance at the graph level by providing a condensed, holistic graph-level representation (*i.e.*, center

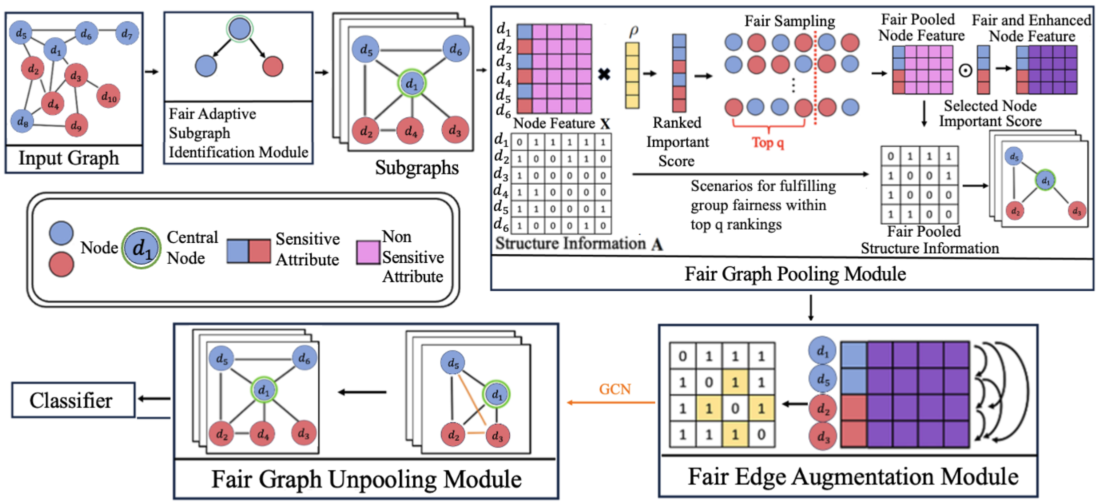


Figure 2: Overview of the proposed Fair Graph U-net framework.

node representation) (Yuan et al. 2022). Specifically, existing methods (Lee, Lee, and Kang 2019; Gao and Ji 2019; Zhang et al. 2018) achieve this by calculating ranking vectors to determine the importance of various nodes. By selecting the top $k$ important nodes to form a new graph, these methods retain essential information and reduce the graph's complexity. Graph pooling techniques can be broadly categorized into flat pooling and hierarchical pooling (Liu et al. 2022). Flat pooling directly generates a graph-level representation in one step, typically by taking the average or sum of all node embeddings (Duvenaud et al. 2015). Hierarchical pooling, on the other hand, gradually coarsens a graph into a smaller-sized graph through two main methods: node clustering pooling (Ying et al. 2018b) and node drop pooling (Gao and Ji 2019). These methods have enabled GNNs to achieve superior performance and have led to their widespread adoption in various domains (Hamilton 2020). For instance, in the financial sector, banks have leveraged GNNs to analyze transaction networks (*i.e.*, the transaction patterns of individual clients) for detecting fraudulent activities or unusual user behavior (Dai et al. 2022). This expanded application into critical decision-making systems demands that GNNs be not only effective but also non-discriminatory (Yuan et al. 2022). In this context, graph pooling, a powerful framework for downsampling and summarizing graph data, also requires attention to its potential biases but has largely been overlooked amidst the increasing focus on fair GNNs.

**Fairness in Graph.** Much progress has been made to address the discriminatory nature of GNNs (Mehrabi et al. 2021; Le Quy et al. 2022; Wang et al. 2025a), broadly categorized into group fairness (Dai and Wang 2021; Wang et al. 2025b; Zhu et al. 2024) and individual fairness (Kang et al. 2020; Wang et al. 2023a, 2024d). Those solutions often import regularization strategies from traditional Euclidean models to address bias in graph-structured data. However, these methods are predominantly developed for Graph Convolution Networks (GCNs) (Dong et al. 2023), it is still hard to understand their effect in conjunction with the graph pooling operation, which is crucial for leveraging the full capabilities of GNNs. Specifically, for pooling methods with graph topology, bias is encoded into the model throughout its unique downsampling and upsampling procedure. In addition, most of the existing fairness works prioritize a single fairness goal—be it individual or group fairness, and come at the cost of the other. For instance, FDGNN (Wang et al. 2024a) utilizes counterfactual samples to learn disentangled node representation to mitigate the multi-source biases and enhance group fairness. To jointly address the above challenges, this work aims to tackle the root causes of bias in graph pooling by i) mitigating multiple potential biases during the pooling process and ii) taking an additional step to ensure both group and individual fairness collectively.

## Notation

Assume the input graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ be an undirected attributed graph; let $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ denote the set of $n$ nodes, $\mathcal{E} \subseteq \{\{v_i, v_j\} \mid v_i, v_j \in \mathcal{V}\}$ denote the set of undirected edges with each edge represented as an unordered pair $\{v_i, v_j\}$, and $\mathbf{X}$ is a $n \times D$ ($n = |V|$) node feature matrix with the $i$-th row, $x_i \in \mathbb{R}^D$, containing the $D$-dimensional feature vector of node $v_i$. Each node $v_i$ has a binary sensitive attribute $s_i \in \{0, 1\}$, included in the feature set $\mathbf{X}$. The set of nodes belonging to the deprived group (*e.g.*, female) is denoted as $S_d = \{v_i | v_i \in V \wedge s_i = 0\}$, and the favored group (*e.g.*, male) is denoted as $S_f = \{v_i | v_i \in V \wedge s_i = 1\}$. Let $N_1$ and $N_2$ represent the number of nodes in $S_d$ and $S_f$, respectively. The graph structure can be denoted as the adjacency matrix $\mathbf{A} \in \{0,1\}^{n \times n}$, where $\mathbf{A}_{i,j} = 1$ if $(v_i, v_j) \in \mathcal{E}$, and $\mathbf{A}_{i,j} = 0$ otherwise. An edge $A_{i,j}$ is classified as an intra-edge if nodes $v_i$ and $v_j$ share the same sensitive attribute value, and as inter-edge if they differ. In addition, we use $y_i$ to denote the ground-truth label for node $v_i$, where $\hat{y}_i$ is the predict label of $v_i$.

## Methodology

### Fair Graph U-net: In a Nutshell

As depicted in Figure 2, the Fair Graph U-net comprises four integral components: **i) Fair Adaptive Subgraph Iden-**

**tification Module** identifies contextual subgraphs for each node, ensuring they encapsulate diverse and pertinent information from different subgroups. This establishes the basis for fair and effective feature extraction in subsequent processes. **ii) Fair Graph Pooling Module** focuses on extracting vital features from the subgraphs. Meanwhile, it tackles representational disparities among various subgroups and alleviates individual unfairness introduced while enforcing group fairness constraints. **iii) Fair Edge Augmentation Module** enhances the graph structure for each node, ensuring no nodes remain isolated post-pooling. It also incorporates a fair edge prediction mechanism that prevents the introduction of structural bias. **iv) Fair Graph Unpooling Module**, designed for upsampling the pooled graph, aims to recover the original graph fairly and accurately. The subsequent sections will delve into the specifics of each module and their collective role in Fair Graph U-net.

## Fair Adaptive Subgraph Identification Module

Real-world graphs are often large and contain significant noise (Pan et al. 2014), posing challenges in processing the entire graph to obtain a graph representation (*i.e.*, center node's representation). Furthermore, computing representations for large graphs can be prohibitively costly and often infeasible (Rossi, Zhou, and Ahmed 2017). To this end, existing methods typically employ a fixed distance (*e.g.*, 1-hop) to generate a subgraph for each node (Jiao et al. 2020), which is inspired by the local dependency assumption (Hamilton, Ying, and Leskovec 2017). This assumption suggests that nodes are significantly influenced by their immediate neighbors. However, this may result in important nodes not being included and introduces structural bias (Ma et al. 2022). Drawing upon previous work (Wang et al. 2024a), this module is introduced to extract contextual subgroups that contain significant information and are fairly represented. Specifically, a *related score* (RS) for each node pair is initially computed to overcome the limitations of generating subgraphs based on preset neighboring node distances, thereby capturing more relevant and contextual information from the graph. Mathematically, the RS measures the relevance of node $v_j$ to node $v_i$ using PageRank (Jeh and Widom 2003) is represented as $RS = \alpha(I - (1-\alpha)\widetilde{A})^{-1}\mathbf{1}$, where $\alpha$, within the interval $[0, 1]$, is the parameter that controls the probability of re-starting from the central node, $I$ is the identity matrix, $\mathbf{1}$ is a vector of all ones, and $\widetilde{A} = \mathbf{A}D^{-1}$ represents the transfer probability with $D$ being the diagonal matrix where $D_{i,i} = \sum_j A_{i,j}$.

However, assigning equal transfer probabilities to each neighboring node can lead to biases. Specifically, nodes with the same sensitive attribute often have stronger connections within the graph (Jiang et al. 2022). This results in higher probability transitions to neighbors sharing the same attribute, thereby over-representation of those nodes in the subgraphs. To this end, a fairness constraint that adjusts the transfer probabilities to ensure equitable representation of different sensitive attribute groups is further introduced. Specifically, the constraint first categorizes neighbors based on their sensitive attributes and then equalizes the total selection probability for each group during node transition. This process ensures that the probability of moving to a node from a particular group is equal during the transition, eliminating the influence of higher inter-connectivity within the group. This is mathematically represented as:

$$\sum(P_{v_a}|\mathbf{A}_{i,a}=1, s_a \in S_d) = \sum(P_{v_b}|\mathbf{A}_{i,b}=1, s_b \in S_f) \qquad (1)$$

where $P_{v_a}$ and $P_{v_b}$ represent the transition probabilities to neighboring nodes belonging to the deprived and favored groups, respectively.

## Fair Graph Pooling Module

With the obtained fair subgraphs, the subsequent challenge is to derive fair graph representations without introducing bias, particularly during pooling operations. Existing graph pooling methods (Gao and Ji 2019; Lee, Lee, and Kang 2019) often overlook the issue of differences in subgroup representation that can arise from the pooling process. This results in biased pooled graphs, which in turn cause downstream tasks to inherit these biases. To this end, we introduce a novel fair graph pooling method that facilitates pooling for graph data while simultaneously addressing both group and individual fairness; Fair Graph U-net fairly chooses a subset of nodes to create a new, downsampled graph while ensuring equitable representation of different groups in the new graph with equalized individual loss.

To understand this process, let's first discuss the trainable projection vector $\rho$ (shown as the yellow rectangle in Figure 2) in Fair Graph U-net, which projects the $D$-dimensional node features matrix ($\mathbf{X}$) onto a 1-dimensional ($1D$) space, enabling the application of fair $k$-max pooling while maintaining consistent connectivity in the resulting pooled graph. Mathematically, this is defined as $\hat{\mathbf{X}} = \mathbf{X}\left(\frac{\rho}{\|\rho\|}\right)$, where $\hat{\mathbf{X}}$ represents the matrix of scalar projections of the node feature vectors in $\mathbf{X}$ onto the projection vector $\rho$. Each entry $\hat{x}_i \in \hat{\mathbf{X}}$ corresponds to the projected feature vector of a node in the graph $\mathcal{G}$, quantifying the retained importance of each node's features when projected onto the direction of $\rho$.

Building on this, nodes are fairly sampled for downsampling, aiming to retain as much information as possible from the original graph while ensuring a consistent representation of different subgroups in the downsampled graph. To achieve this, an equal number of nodes are selected from each subgroup to form the new graph, adhering to the principle of group fairness, which requires equitable treatment across groups. A straightforward way is to choose the top $\frac{q}{2}$ nodes with the largest scalar projection values on $\rho$ from each subgroup. However, this strategy is individual fairness unaware. For instance, as illustrated in Figure 1, achieving group fairness might result in a situation where only certain nodes, such as $v_4$, experience a notable change in their representation or outcome. This leads to a scenario where the burden of maintaining group fairness falls disproportionately on individuals like $v_4$. Such differential treatment, although in line with group fairness objectives, might not be perceived as fair by these affected individuals, thereby raising questions about the overall fairness of the outcome from an individual perspective.

To mitigate the individual unfairness introduced by group fairness constraints, we incorporate the Rawlsian difference principle (Rawls 1971), which aims for equity by optimizing the well-being of the most deprived individuals. This principle, drawn from Rawls' theories, ensures that in a state of equilibrium, all individuals maintain their status quo since the welfare of the least deprived cannot be further maximized, leading to fair outcomes for everyone. In Fair Graph U-net, the concept of welfare is quantified by the group fairness loss borne by individuals, which is evaluated through changes in their ranking positions. Essentially, the task involves finding a ranking that, while adhering to the group fairness constraint, maximizes the utility of the least-deprived individual. This objective is mathematically defined as follows:

$$\min \sum_{i=0}^{n} |(U(v_i, pos) - U(v_i, pos')) - \mu| \tag{2}$$

where $U(v_i, pos)$ represents the welfare (utility) of individual $v_i$ when placed at position $pos$ in a ranking, and $\mu$ denotes the average group loss borne by all individuals. The objective is to minimize the absolute difference between individual utility and the average group loss, aligning with the Rawlsian principle of optimizing the welfare of the least advantaged.

Building on this, we define a validity ranking $R_{valid}$ that ensures group fairness can be created. Assuming distinct utilities for each position ($U(v_i, pos)$) to avoid ties, $R_{valid}$ is defined as a set of rankings $r_i$ that satisfy group fairness, meaning the number of samples from different subgroups in the top $q$ positions is equal:

$$R_{valid} = \{\forall r_i \in R, |\{S_d \cap \text{top}_q(r_i)\}| = |\{S_f \cap \text{top}_k = q(r_i)\}|\} \tag{3}$$

We then induce a probability distribution $\mathbf{D}$ over $R_{valid}$ that the minimum expected value obtained by any individual $v_i$ is higher than the expected value that could be obtained by any single ranking $r_i$. However, since $R_{valid}$ may be exponentially large, direct enumeration is impractical. To this end, we initially assign equal probability to all valid rankings and incrementally adjust $\mathbf{D}$. We use a weighted optimization oracle to find new solutions that place more emphasis on less satisfied individuals. By progressively shifting the weight in $\mathbf{D}$, we will obtain probability distributions that satisfy both group fairness and individual fairness. Specifically, for each probability distribution, individuals are categorized into those that satisfy their expected satisfaction and those that do not. Through an iterative optimization process, we gradually adjust the distributions to ensure that as many individuals as possible meet their expected satisfaction while maintaining the state of satisfied individuals until all individuals are added to the set of satisfied individuals. Mathematically, it can be defined as follows:

$$\min \left( \max(\textstyle\sum_{i=0}^{n} w_{v_i} \cdot Val(D, v_i)) - \sum_{v_i \in V_{Sat}} \delta_{v_i} w_{v_i} \right) \tag{4}$$

where $w_{v_i}$ denotes the probability assigned to each validity ranking, $Val(\mathbf{D}, v_i)$ represents the relevant scores of individual $v_i$ under validity ranking $r_i$ with distribution $\mathbf{D}$, $\delta_{v_i}$ is satisfaction threshold for the individuals, and $V_{Sat}$ is a set of individual that satisfy their expected satisfaction. By formulating the problem this way, given a weighted optimization oracle, we establish the groundwork for creating a separation oracle for the dual problem. Given the presence of a separation oracle enables the polynomial-time resolution of a linear program through the ellipsoid algorithm (Grötschel, Lovász, and Schrijver 1981). Consequently, we can obtain probability distributions that satisfy both group fairness and individual fairness in polynomial time.

Overall, this module ensures consistent representation of different groups in the downsampled graph while ensuring that no individual is disproportionately disadvantaged when enforcing group fairness constraints.

## Fair Edge Augmentation Module

Existing graph pooling approaches sample crucial nodes to create a new graph for improved feature encoding. However, the removal of nodes in this process, along with their associated edges, may lead to isolated nodes in the pooled graph. This has the potential to impact the propagation of information in subsequent layers, particularly when utilizing GCN layers that depend on aggregating information from neighboring nodes. Moreover, straightforward methods to enhance graph connectivity, such as directly linking 2-hop neighbors, can inadvertently exacerbate biases present within the original graph structure, leading to graph structure bias. To address these challenges, this module aims to fairly increase connectivity in the pooled graph. Specifically, link predictors are trained using the edge distribution of the original data. These predictors are subsequently used to enhance the graph's connectivity, determining the likelihood of a link between any two nodes $v_i$ and $v_j$ based on their similarity, calculated through a weighted inner product. Mathematically, the predicted relation ($E_{(v_i,v_j)}$) between nodes $v_i$ and $v_j$ is define as follows:

$$E_{(v_i,v_j)} = \text{softmax}(\sigma(\overline{x_i} \cdot \mathbf{Z} \cdot \overline{x_j})) \tag{5}$$

where $\sigma$ represents the activation function, and $\overline{x_i} = x_i \odot \hat{x}_i \mathbf{1}^T$ denotes enhanced feature vector, while the parameter matrix $\mathbf{Z}$ captures their interaction dynamics. In addition, given that most connections in the input graph $\mathcal{G}$ are intra-group, there is a risk of the module reinforcing these connections, which can lead to overassociation of the graph structure with sensitive attributes and thereby intensifying graph structure bias. Therefore, we also assess the model performance difference between inter-group and intra-group connections when generating graph structure information. Overall, the fair edge augmentation module's loss function is represented as:

$$\mathcal{L}_A = \|\mathbf{E} - \mathbf{A}\|_F^2 + \|\mathbf{E}_{inter} - \mathbf{A}_{inter}\|_F^2 - \|\mathbf{E}_{intra} - \mathbf{A}_{intra}\|_F^2 \tag{6}$$

where $\mathbf{E}$ predicts the connections between nodes in $\mathcal{V}$, and $E_{inter}$ and $E_{intra}$ represent the predicted connections between inter-group and intra-group nodes, respectively, while $A_{inter}$ and $A_{intra}$ denote the actual connections.

### Fair Graph Unpooling Module

Aligned with the principles of encoder-decoder networks, this module executes the reverse operation of the fair graph pooling layer, efficiently restoring the graph to its initial structural integrity. To achieve this, the positions of nodes selected in the corresponding fair graph pooling layer are recorded, guiding the placement of nodes back to their original positions in the graph during the up-sampling phase. Mathematically represented as follows:

$$\mathbf{X}^{l+1} = \mathrm{Dis}(0_{N\times D}, \overline{\mathbf{X}}^{l}, \mathrm{pos}_{ori}) \tag{7}$$

where $\mathrm{pos}_{ori}$ contains indices of selected nodes in the corresponding fair graph pooling layer that reduces the graph size from $N$ nodes to $q$ nodes, and $0_{N\times D}$ represents an initially empty feature matrix for the restored graph. In addition, $\mathrm{Dis}(\cdot)$ distributes the row vectors from the pooled feature matrix $\overline{\mathbf{X}}^{l}$ into the $0_{N\times D}$ matrix according to their original indices stored in $\mathrm{pos}_{ori}$, and in $\mathbf{X}^{l+1}$, the row vectors corresponding to the indices in $\mathrm{pos}_{ori}$ are updated with the appropriate vectors from $\overline{\mathbf{X}}^{l}$, while all other vectors remain zero, effectively reconstructing the original feature matrix of the graph.

### Overall Objective Function

This section presents the overall objective function for optimizing Fair Graph U-net as depicted in Equation 8. This function consists of three parts and is governed by the tunable hyperparameters $\beta$ and $\gamma$, which are responsible for balancing the contributions of the various elements in the overall objective function. Specifically, the first term, $\mathcal{L}_U$, aims to minimize the prediction loss. The next term, $\mathcal{L}_F$ designed to mitigate differences in relevant scores between subgroups, and the last term, $\mathcal{L}_A$, aims to promote the fairness of the edge predictor and avoid introducing structural bias.

$$\begin{aligned} \arg\min\ \mathcal{L} &= \mathcal{L}_U + \beta\mathcal{L}_F + \gamma\mathcal{L}_A \\ &= \frac{1}{|\mathcal{V}_L|}\sum_{v_i\in\mathcal{V}_L} -(y_i\log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)) \\ &+ \beta\left(\frac{1}{|S_d|}\sum_{v_i\in S_d}\ell(\hat{y}_i, y_i) - \frac{1}{|S_f|}\sum_{v_j\in S_f}\ell(\hat{y}_j, y_j)\right) + \gamma\mathcal{L}_A \end{aligned} \tag{8}$$

## Experiment

### Experiment Setting

**Datasets and Baselines** Experiments are conducted on three commonly used datasets, including German dataset (Asuncion and Newman 2007), Credit dataset (Yeh and Lien 2009), and Bail Dataset (Agarwal, Lakkaraju, and Zitnik 2021). In addition, we compare Fair Graph U-net with the following baseline methods, including, GCN (Kipf and Welling 2016), Graph U-net (Gao and Ji 2019), SAGpool (Lee, Lee, and Kang 2019), FairGNN (Dai and Wang 2021), Graphair (Ling et al. 2023) and FairAGG (Zhu et al. 2024).

**Evaluation Metrics.** We evaluate the model performance from two perspectives: classification performance and fairness. For classification performance, we adopt the AUC and F1-score to evaluate node classification performance. Higher values of these metrics indicate better classification performance. For fairness performance, we follow existing work on fair GNNs to use two widely applied evaluations, *i.e.*, Statistical Parity Differences (SPD) (Le Quy et al. 2022) and Equal Opportunity Differences (EOD) (Hardt, Price, and Srebro 2016). A model is considered better in fairness when it achieves lower values of SPD and EOD. Additionally, we introduce the Maximum Individual Unfairness (MIU) as a measure of the individual's unfairness due to group fairness constraints. This is defined as $Max$ $(U(v_i, pos)$ - $U(v_i, pos'))$, where $pos$ and $pos'$ represent an individual's position before and after adjustments imposed by group fairness constraints.

| Dataset | Methods | Accuracy | F1-Score | SPD | EOD | MIU |
|---|---|---|---|---|---|---|
| German | GCN | 0.65 | 0.78 | 0.36 | 0.31 | - |
| | Graph U-net | 0.73 | 0.82 | 0.26 | 0.21 | - |
| | SAGPool | 0.71 | 0.81 | 0.18 | 0.16 | - |
| | FairGNN | 0.67 | 0.82 | 0.09 | 0.05 | 0.13 |
| | Graphair | 0.55 | 0.81 | 0.06 | 0.04 | 0.21 |
| | FairAGG | 0.68 | 0.78 | 0.06 | 0.04 | 0.18 |
| | Fair Graph U-net | 0.72 | 0.81 | 0.05 | 0.04 | 0.05 |
| Credit | GCN | 0.68 | 0.83 | 0.11 | 0.10 | - |
| | Graph U-net | 0.76 | 0.81 | 0.18 | 0.15 | - |
| | SAGPool | 0.69 | 0.71 | 0.15 | 0.12 | - |
| | FairGNN | 0.68 | 0.78 | 0.12 | 0.10 | 0.11 |
| | Graphair | 0.57 | 0.73 | 0.08 | 0.09 | 0.13 |
| | FairAGG | 0.65 | 0.77 | 0.07 | 0.06 | 0.18 |
| | Fair Graph U-net | 0.75 | 0.82 | 0.06 | 0.06 | 0.06 |
| Bail | GCN | 0.83 | 0.78 | 0.09 | 0.04 | - |
| | Graph U-net | 0.81 | 0.77 | 0.16 | 0.17 | - |
| | SAGPool | 0.82 | 0.75 | 0.14 | 0.13 | - |
| | FairGNN | 0.81 | 0.77 | 0.07 | 0.05 | 0.08 |
| | Graphair | 0.68 | 0.76 | 0.05 | 0.05 | 0.15 |
| | FairAGG | 0.75 | 0.74 | 0.06 | 0.04 | 0.12 |
| | Fair Graph U-net | 0.81 | 0.76 | 0.04 | 0.02 | 0.03 |

Table 1: Predictive and fairness performance for Fair Graph U-net and baselines across three datasets (the darker cells show the top rank and the lighter cells show the second rank).

### Experiment Results

**Performance Comparison.** We compare the performance of Fair Graph U-net with six baselines on the three datasets. To ensure a fair comparison, we select the optimal hyperparameter configurations for all methods based on their validation set performance. Note that GCN, Graph U-net, and SAGPool do not incorporate group fairness constraints, rendering their MIU scores inapplicable. Each experiment is repeated 10 times, and the average performance is reported, with the results presented in Table 1. As one can see, Fair Graph U-net outperforms all baseline methods across most evaluation metrics. Specifically, Fair Graph U-net demonstrates superior fairness performance (*i.e.*, SPD, EOD, and MIU) as evidenced by the significant margin overall baseline methods across all datasets. This superior performance in fairness can be primarily attributed to two reasons: i) Fair Graph U-net effectively reduces the overcorrelation with sensitive attributes in new graph representations by balancing the representation of different subgroups during the graph pooling process, and ii) it minimizes the influence of structural biases in subgraphs and edge augmentations, thereby reducing their impact on prediction outcomes. In addition to its fairness advantages, Fair Graph U-net also demonstrates exceptional utility performance, surpassing other methods in most cases. This indicates that Fair Graph U-net not only addresses fairness concerns effectively but also excels in capturing the underlying latent factors of data, which enhances the model's predictive performance.

| Datasets | German | Credit | Bail | German | Credit | Bail |
|---|---|---|---|---|---|---|
| Fairness Metrics | SPD | | | MIU | | |
| Fair Graph U-net-G | 0.03 | 0.07 | 0.03 | 0.11 | 0.13 | 0.06 |
| Fair Graph U-net | 0.05 | 0.06 | 0.04 | 0.05 | 0.06 | 0.03 |
| Utility Metrics | Accuracy | | | F1-score | | |
| Fair Graph U-net-G | 0.69 | 0.72 | 0.78 | 0.79 | 0.79 | 0.73 |
| Fair Graph U-net | 0.72 | 0.74 | 0.81 | 0.81 | 0.82 | 0.76 |

Table 2: Ablation study results for Fair Graph U-net and Fair Graph U-net-G.

**Individual Fairness Enhancement.** To evaluate the effectiveness of our balanced individual and group fairness strategy, we compared it with strategies focused only on group fairness. For this comparison, we introduced a variant of our model, Fair Graph U-net-G, which emphasizes only group fairness by ensuring an equal representation of different subgroups among the top $q$ individuals. As shown in Table 2, Fair Graph U-net-G, which lacks the individual fairness component, exhibited a notable reduction in individual fairness results compared to the original Fair Graph U-net model. Additionally, there was a slight decline in overall performance. This decline underscores the limitations of focusing solely on group fairness: it may inadvertently reduce the diversity of samples in the reduced graph, thereby diminishing the model's generalization capabilities.

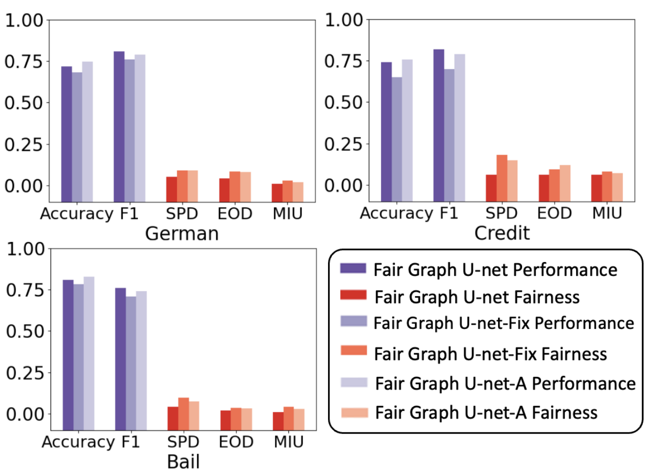


Figure 3: Ablation study results for Fair Graph U-net, Fair Graph U-net-Fix, and Fair Graph U-net-A.

**Ablation Study.** We conduct ablation studies to dissect the contributions of each component within Fair Graph U-net on improving fairness. Specifically, we evaluate Fair Graph U-net without the fair adaptive subgraph identifying module and without the fair edge augmentation module, denoted as Fair Graph U-net-Fix and Fair Graph U-net-A, respectively. Figure 3 presents the results of these studies. Compared to the full Fair Graph U-net, Fair Graph U-net-Fix exhibits reduced fairness and performance. This is attributed to its approach of statically determining subgraphs based on a fixed 2-hop distance from each node. This strategy overlooks potentially significant neighbors and exacerbates discrepancies in group representation within subgraphs, leading to a downgrade of the training quality of the model.

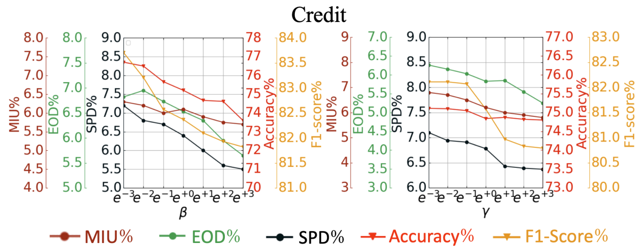


Figure 4: Exploring hyperparameters study results in the Credit dataset.

On the other hand, Fair Graph U-net-A employs a straightforward strategy of linking each node to its neighbors within a 2-hop radius without considering fairness, leading to diminished fairness and performance relative to the complete Fair Graph U-net model. The decrease observed with Fair Graph U-net-A underscores the vital role of the fair edge augmentation module in mitigating biases associated with the graph's structure. This module's absence in Fair Graph U-net-A results in an overassociation of the reduced graph with sensitive attributes, highlighting the importance of this component in balancing fairness and performance effectively.

**Parameters Sensitivity.** We explore the sensitivity of Fair Graph U-net with respect to two key hyperparameters: $\beta$, which is tuned to optimize relevant score fairness, and $\gamma$, aimed at enhancing graph structure. We assess the impact of these hyperparameters on the model's performance and fairness by conducting experiments where $\beta$ and $\gamma$ are varied across the set $\{1e^{-3}, 1e^{-2}, 1e^{-1}, 1e^{0}, 1e^{1}, 1e^{2}, 1e^{3}\}$. Using the Credit dataset as a case study, the outcomes are depicted in Figure 4. The results indicate a clear trend: increases in $\beta$ and $\gamma$ generally lead to a decline in Fair Graph U-net's performance metrics. Specifically, a higher $\beta$ improves the model's capability to achieve group fairness by giving greater emphasis to samples from underrepresented subgroups. Meanwhile, an elevated $\gamma$ enhances the model's effectiveness in counteracting graph structure bias, thereby elevating the overall fairness.

## Conclusion

This paper introduces Fair Graph U-net, a novel approach that extends the powerful generalization capability of convolutional pooling to graphs while addressing its inherent bias toward fair GNNs. Additionally, this paper takes one step further to innovatively navigate the nuances associated with group and individual graph fairness, unifying these two separate yet interconnected objectives in existing works. Extensive experimentation with real-world datasets demonstrates the superior capability of the proposed framework.

## Acknowledgements

This work was supported in part by the National Science Foundation (NSF) under Grant No. 2245895.

## References

Agarwal, C.; Lakkaraju, H.; and Zitnik, M. 2021. Towards a unified framework for fair and stable graph representation learning. In *Uncertainty in Artificial Intelligence*, 2114–2124. PMLR.

Asuncion, A.; and Newman, D. 2007. UCI machine learning repository.

Bao, J.; He, T.; Ruan, S.; Li, Y.; and Zheng, Y. 2017. Planning bike lanes based on sharing-bikes' trajectories. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 1377–1386.

Cho, D.-Y.; Kim, Y.-A.; and Przytycka, T. M. 2012. Chapter 5: Network biology approach to complex diseases. *PLoS computational biology*, 8(12): e1002820.

Chu, Z.; Wang, Z.; and Zhang, W. 2024. Fairness in Large Language Models: A Taxonomic Survey. *ACM SIGKDD Explorations Newsletter, 2024*, 34–48.

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

Dai, E.; Zhao, T.; Zhu, H.; Xu, J.; Guo, Z.; Liu, H.; Tang, J.; and Wang, S. 2022. A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability. *arXiv preprint arXiv:2204.08570*.

Dong, Y.; Ma, J.; Wang, S.; Chen, C.; and Li, J. 2023. Fairness in graph mining: A survey. *IEEE Transactions on Knowledge and Data Engineering*.

Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; and Adams, R. P. 2015. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28.

Gao, H.; and Ji, S. 2019. Graph u-nets. In *international conference on machine learning*, 2083–2092. PMLR.

Grötschel, M.; Lovász, L.; and Schrijver, A. 1981. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1: 169–197.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Hamilton, W. L. 2020. *Graph representation learning*. Morgan & Claypool Publishers.

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

Jeh, G.; and Widom, J. 2003. Scaling personalized web search. In *Proceedings of the 12th international conference on World Wide Web*, 271–279.

Jiang, Z.; Han, X.; Fan, C.; Liu, Z.; Zou, N.; Mostafavi, A.; and Hu, X. 2022. Fmp: Toward fair graph message passing against topology bias. *arXiv preprint arXiv:2202.04187*.

Jiao, Y.; Xiong, Y.; Zhang, J.; Zhang, Y.; Zhang, T.; and Zhu, Y. 2020. Sub-graph contrast for scalable self-supervised graph representation learning. In *2020 IEEE international conference on data mining (ICDM)*, 222–231. IEEE.

Kang, J.; He, J.; Maciejewski, R.; and Tong, H. 2020. Inform: Individual fairness on graph mining. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 379–389.

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

Kleindessner, M.; Samadi, S.; Awasthi, P.; and Morgenstern, J. 2019. Guarantees for spectral clustering with fairness constraints. In *International Conference on Machine Learning*, 3458–3467. PMLR.

Le Quy, T.; Roy, A.; Iosifidis, V.; Zhang, W.; and Ntoutsi, E. 2022. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3): e1452.

Lee, J.; Lee, I.; and Kang, J. 2019. Self-attention graph pooling. In *International conference on machine learning*, 3734–3743. PMLR.

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, C.; Zhan, Y.; Wu, J.; Li, C.; Du, B.; Hu, W.; Liu, T.; and Tao, D. 2022. Graph pooling for graph neural networks: Progress, challenges, and opportunities. *arXiv preprint arXiv:2204.07321*.

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

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

Pan, S.; Wu, J.; Zhu, X.; and Zhang, C. 2014. Graph ensemble boosting for imbalanced noisy graph stream classification. *IEEE transactions on cybernetics*, 45(5): 954–968.

Rahman, T.; Surma, B.; Backes, M.; and Zhang, Y. 2019. Fairwalk: Towards fair graph embedding.

Rawls, A. 1971. Theories of social justice.

Rossi, R. A.; Zhou, R.; and Ahmed, N. K. 2017. Estimation of graphlet statistics. *arXiv preprint arXiv:1701.01772*.

Sui, Y.; Wang, X.; Wu, J.; Lin, M.; He, X.; and Chua, T.-S. 2022. Causal attention for interpretable and generalizable graph classification. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1696–1705.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wang, X.; Gu, T.; Bao, X.; Chang, L.; and Li, L. 2023a. Individual fairness for local private graph neural network. *Knowledge-Based Systems*, 268: 110490.

Wang, Z.; Chu, Z.; Blanco, R.; Chen, Z.; Chen, S.-C.; and Zhang, W. 2024a. Advancing Graph Counterfactual Fairness through Fair Representation Learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 40–58. Springer Nature Switzerland.

Wang, Z.; Dzuong, J.; Yuan, X.; Chen, Z.; Wu, Y.; Yao, X.; and Zhang, W. 2024b. Individual Fairness with Group Awareness Under Uncertainty. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 89–106. Springer Nature Switzerland.

Wang, Z.; Hoang, N.; Zhang, X.; Bello, K.; Zhang, X.; Iyengar, S. S.; and Zhang, W. 2025a. Towards Fair Graph Learning without Demographic Information. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*.

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

Wang, Z.; Qiu, M.; Chen, M.; Salem, M. B.; Yao, X.; and Zhang, W. 2024c. Toward Fair Graph Neural Networks via Real Counterfactual Samples. *Knowledge and Information Systems*, 1–25.

Wang, Z.; Saxena, N.; Yu, T.; Karki, S.; Zetty, T.; Haque, I.; Zhou, S.; Kc, D.; Stockwell, I.; Bifet, A.; et al. 2023c. 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. 2024d. 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. 2023d. FG$^2$AN: Fairness-Aware Graph Generative Adversarial Networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 259–275. Springer Nature Switzerland.

Wang, Z.; Yin, Z.; Zhang, Y.; Yang, L.; Zhang, T.; Pissinou, N.; Cai, Y.; Hu, S.; Li, Y.; Zhao, L.; and Zhang, W. 2025b. FG-SMOTE: Towards Fair Node Classification with Graph Neural Network. *ACM SIGKDD Explorations Newsletter*.

Wang, Z.; and Zhang, W. 2024. Group Fairness with Individual and Censorship Constraints. In *27th European Conference on Artificial Intelligence*.

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.

Ying, R.; He, R.; Chen, K.; Eksombatchai, P.; Hamilton, W. L.; and Leskovec, J. 2018a. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 974–983.

Ying, Z.; You, J.; Morris, C.; Ren, X.; Hamilton, W.; and Leskovec, J. 2018b. Hierarchical graph representation learning with differentiable pooling. *Advances in neural information processing systems*, 31.

Yuan, H.; Yu, H.; Gui, S.; and Ji, S. 2022. Explainability in graph neural networks: A taxonomic survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(5): 5782–5799.

Zhang, M.; Cui, Z.; Neumann, M.; and Chen, Y. 2018. An end-to-end deep learning architecture for graph classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

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

Zhang, W. 2024b. Fairness with Censorship: Bridging the Gap between Fairness Research and Real-World Deployment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 22685–22685.

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

Zhang, W.; and Ntoutsi, E. 2019. FAHT: An Adaptive Fairness-aware Decision Tree Classifier. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization.

Zhang, W.; Wang, Z.; Kim, J.; Cheng, C.; Oommen, T.; Ravikumar, P.; and Weiss, J. 2023. Individual Fairness under Uncertainty. In *26th European Conference on Artificial Intelligence*, 3042–3049.

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

Zhang, W.; Weiss, J. C.; Zhou, S.; and Walsh, T. 2024. Fairness amidst non-iid graph data: A literature review. *AI Magazine*, 2.

Zhao, T.; Liu, G.; Wang, D.; Yu, W.; and Jiang, M. 2022. Learning from counterfactual links for link prediction. In *International Conference on Machine Learning*, 26911–26926. PMLR.

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*.