ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240679

Individual Fairness with Group Constraints in Graph Neural Networks

Zichong Wang[a], David Ulloa[a], Tongjia Yu[b], Raju Rangaswami[a], Roland H. C. Yap[c] and Wenbin Zhang[a,*]

[a]Florida International University, Miami, FL, 33199
[b]Goldman Sachs, New York, NY, 10282
[c]National University of Singapore, Singapore, 117417

Abstract. Graph Neural Networks (GNNs) have demonstrated remarkable capabilities across various domains. Despite the successes of GNN deployment, their utilization often reflects societal biases, which critically hinder their adoption in high-stake decision-making scenarios such as online clinical diagnosis, financial crediting, etc. Numerous efforts have been made to develop fair GNNs but they typically concentrate on either individual or group fairness, overlooking the intricate interplay between the two, resulting in the enhancement of one, usually at the cost of the other. In addition, existing individual fairness approaches using a ranking perspective fail to identify discrimination in the ranking. This paper introduces two innovative notions dealing with individual graph fairness and group-aware individual graph fairness, aiming to more accurately measure individual and group biases. Our Group Equality Individual Fairness (GEIF) framework is designed to achieve individual fairness while equalizing the level of individual fairness among subgroups. Preliminary experiments on several real-world graph datasets demonstrate that GEIF outperforms state-of-the-art methods by a significant margin in terms of individual fairness, group fairness, and utility performance.

1 Introduction

Graph data is widely available in real-world scenarios, such as the financial markets [54], biological networks [49], and social networks [37]. Given that graph data contain not only node feature information but also structural details about the nodes and their neighbors, numerous graph mining algorithms have been developed over the past few decades to gain deep insights from graph data [29, 60, 62, 26]. Among these, GNNs stand out as a fundamental and widely adopted approach, demonstrating remarkable efficacy across a multitude of tasks and applications [12, 34, 9], which has led to GNNs playing an increasingly important role in a variety of high-stakes decision-making scenarios, such as credit scoring [31], recommendation [28], and healthcare [25]. However, the increased adoption of GNNs in high-risk decision-making scenarios is intensifying concerns about their fairness. For instance, serious ethical issues arise when a credit agency's scoring is influenced by racial information in a customer's social network, including the racial identities of their close contacts [48].

To address fairness in algorithmic decision-making, researchers have introduced various fairness concepts, generally categorized into two types: group fairness and individual fairness [21]. Group fairness [57, 58, 61] focuses on ensuring that models equitable treatment of different subgroups defined by *sensitive attributes* such as gender or race, *i.e.,* aiming to achieve statistical parity across these subgroups, ensuring that no group is disadvantaged by the model's outcomes. On the other hand, individual fairness [10, 27, 64] is able to enforce fairness at a finer granularity at the individual level compared to group fairness. Its objective is to ensure that similar individuals (based on their input features) receive similar predictions. To achieve individual fairness, existing works [41, 55, 56] on individual fairness employ a Lipschitz condition, which is parametrized by a Lipschitz constant. Specifically, it ensures that the output distance between any pair of individuals is proportionally less than or equal to their input distance multiplied by this scalar. Essentially, it guarantees that small differences in inputs do not lead to disproportionately large differences in outcomes. However, specifying a suitable Lipschitz constant to compare these distances accurately can be difficult due to the variation in distance metrics between the input and output spaces.

While existing fairness works [6, 8, 38, 45] in machine learning have shown effectiveness, they often treat individual and group fairness as separate goals. The drawback is that this can result in models that enhance individual fairness at the expense of group fairness [39]. For instance, consider a loan application scenario illustrated in Figure 1, where d_1 already has his loan approved and d_2, d_3, and d_4's loan decisions are to be decided by the AI algorithm. In the input space, the similarity ranking of applicants to d_1 is $d_2 > (d_3 = d_4)$. The Lipschitz condition for d_1 is $Dis(d_1, d_j) \leq 35$, where d_j denotes all the comparable neighbors and $Dis(\cdot)$ is a function that measure the distance of instances in the output space. Here, d_2, d_3, and d_4 satisfy the Lipschitz condition ($Dis \leq 35$). However, when the model concentrates exclusively on individual fairness, the optimization of the constraining scalar may be susceptible to the influence of sensitive attributes associated with the individuals involved. As a result, d_4 is further away from the reference individual d_1 in the output space compared to d_3 (30 v.s. 20). Although this result satisfies individual fairness, a larger outcome distance disparity could potentially put the applicant d_4 on the unfavorable side of the loan approval decision boundary, giving her a different outcome for her loan application, thereby introducing the group bias. We believe it is crucial to consider group fairness alongside individual fairness. By balancing the model's performance in terms of individual fairness across different subgroups, our model can effectively avoid favoring any particular subgroup.

Despite its fundamental importance, achieving individual fairness with group awareness in GNNs remains a largely unexplored area, presenting three distinct challenges: **i) Simultaneous Achievement**

* Corresponding author. Email: wenbin.zhang@fiu.edu

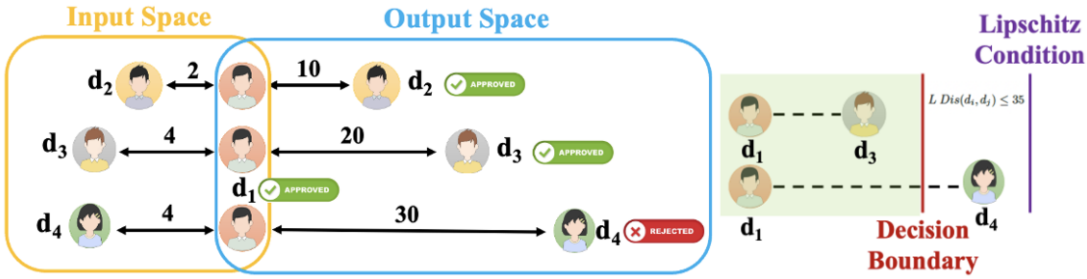

Figure 1: A toy example of the disparity of individual fairness between different subgroups in a loan approval system.

of Individual and Group Fairness. Achieving individual fairness can inadvertently result in disparate treatment of individuals from different subgroups. This introduces additional biases, thereby potentially undermining group fairness, which necessitates a nuanced approach that can mitigate biases at both individual and group levels. **ii) Identifying individuals who are truly discriminated against.** Individual fairness requires a careful examination of each individual, taking into account the interconnected nature of graph data. The non-independent and Identically Distributed nature of graph data further complicates the process of identifying similar individuals and accurately quantifying individual unfairness. **iii) Unbounding the Lipschitz Condition.** Most existing individual fairness works rely on the Lipschitz condition to align differences in input and output space metrics. However, this condition often limits their applicability in real-world scenarios. Moreover, the absolute distance comparison in the Lipschitz condition fails to calibrate the differences between different instances.

In response to these challenges, this paper introduces two novel fairness metrics: **Individual Rearranged Graph Fairness (IRGF)** and **Individual-Group Graph Fairness** ($\mathrm{IG}^2\mathrm{F}$). Building on this, we proposed the **Group Equality Individual Fairness (GEIF)**, a novel framework for fair AI decision-making in graph-based models. *To the best of our knowledge, GEIF represents the first work that fundamentally analyzes individual unfairness and its variation across subgroups in graphs.* Specifically, we propose IRGF identifies individual unfairness by measuring the difference between individuals in the input and output spaces through a ranking perspective. This approach enables a direct measurement of individual unfairness while simultaneously unbounding the constraints of the traditional Lipschitz condition. We propose $\mathrm{IG}^2\mathrm{F}$ derived from the IRGF, which measures the individual unfairness of each person. $\mathrm{IG}^2\mathrm{F}$ allows us to quantify the disparity in fairness between subgroups, offering a comprehensive view of group unfairness. The GEIF framework incorporates these two innovative metrics to achieve a dual objective: ensuring overall individual fairness and establishing comparable levels of fairness between groups, all while maintaining the model's predictive utility. The key contributions of this paper can be summarized as follows:

- **Notions.** We propose innovative individual and group fairness metrics. Our individual fairness metric, IRGF, evaluates bias through a novel ranking perspective, enabling more precise individual bias assessment and quantification of associated group-level bias. This approach eliminates the dependence on the Lipschitz condition. Additionally, the $\mathrm{IG}^2\mathrm{F}$ metric, derived from IRGF, measures disparities in individual fairness across different subgroups.
- **Method.** We present a novel GEIF framework, GEIF, to mitigate individual biases while avoiding group unfairness. Utilizing our newly proposed fairness metrics, we develop specific loss functions for both individual and group fairness. Moreover, the framework allows for adjustable hyperparameters, enabling a balanced control between the model's utility and fairness.
- **Experiments.** We conduct extensive experiments on three real-world benchmark datasets. The results demonstrate the efficacy of our GEIF model which not only outperforms existing baselines in fairness but also achieves comparable prediction performance in downstream tasks.

2 Related Work

2.1 Graph Neural Networks

Graph neural networks (GNNs) have found widespread utility in various tasks involving graph-structured data, such as node classification [35, 18, 3], graph classification [33, 24], and link prediction [66, 50]. Their exceptional performance in these domains has broadened their applicability [13]. For example, financial institutions can employ GNNs to analyze customer transaction networks, helping to make informed credit decisions. However, using GNNs in such high-risk decision-making scenarios necessitates additional aspects to be addressed, one of them being fairness [52, 63, 59]. In the context of the financial scenario, decisions guided by GNNs need to be both accurate and fair, considering their far-reaching implications on an individual's financial status and future opportunities. As such, there's a growing emphasis in the research community on devising GNN models that factor in fairness when dealing with graph-related tasks [15, 7].

2.2 Fairness in Graph Learning

Recent years have seen growing attention in fairness within machine learning [32, 40, 42, 43, 47]. Typical notions of fairness in graphs are generally categorized into two types: group fairness and individual fairness [44]. Group fairness works [5, 14, 53] seeks statistical equality among subgroups defined by sensitive attribute(s). On the other hand, individual fairness works [30, 46, 65] aim to ensure that similar individuals receive similar treatment, promoting equitable treatment irrespective of sensitive attributes like race or gender. Nevertheless, considering that individual fairness is able to enforce fairness at a finer granularity at the individual level compared to group fairness, it relies on the Lipschitz condition. To this end, recent works [64] have attempted to measure individual fairness through a ranking perspective. However, these methods often inaccurately assess individual fairness loss. They quantify individual bias directly on the focused individual used to form the ranking, rather than individuals within the ranking. This approach overlooks those who experience a shift in their ranking position, which is a root cause of individual bias. In addition, a common limitation of these approaches is their failure to address the intrinsic connection between individual and group fairness. Enhancing one often comes at the cost of the other, thus failing to eliminate systematic bias at its root. Furthermore, traditional group fairness

metrics in graphs are unable to measure individual fairness disparities across different subgroups. Some recent work, such as GUIDE [32], has attempted to address both individual and group equity by minimizing losses associated with each. However, these approaches are often constrained by the Lipschitz condition and lack versatility across different datasets. Maxmin-Fair [11] employs a Min-Max algorithm [36], and aims to equalize the total group fairness loss experienced by each individual across multiple runs of the algorithm. Yet, in scenarios requiring only a single query, such as healthcare resource allocation, achieving both individual and group fairness simultaneously remains elusive.

In response to these challenges, our work differs from these cited works in that we not only optimize overall individual fairness but also explicitly equalize the levels of fairness across subgroups. We aim to ensure that sensitive attributes do not disproportionately affect an individual's fairness level compared to the similar individual. Additionally, our approach overcomes the limitations imposed by the Lipschitz condition, enhancing its applicability in a variety of real-world scenarios.

3 Preliminary

3.1 Notation

Consider an undirected graph denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$, where $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ represents the set of N nodes, $\mathcal{E} \subseteq \{\{v_i, v_j\} \mid v_i, v_j \in \mathcal{V}\}$ denotes 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. To encapsulate the graph structural information, we define the adjacency matrix $\mathbf{A}$ corresponding to $\mathcal{G}$, where the matrix element $\mathbf{A}_{i,j}$ is set to 1 if an edge exists between nodes v_i and v_j (*i.e.,* $\{v_i, v_j\} \in \mathcal{E}$), and is set to 0 otherwise. Meanwhile, following previous works, we focus on binary sensitive attributes and binary node classification tasks. Each node v_i has a sensitive attribute, we utilize $s_i \in \{0, 1\}$ to represent the sensitive attribute, where s_i indicates the group membership of the individual v_i: if s_i = 0, v_i belongs to the deprived group S_0 = $\{\forall\, v_i : v_i \in \mathcal{V} \wedge s_i = 0\}$; if s_i = 1 , v_i belongs to the favored group S_1 = $\{\forall\, v_i : v_i \in \mathcal{V} \wedge s_i = 1\}$. Note that $s_i \in \mathbf{X}$. N_{S_0} and N_{S_1} are the number of nodes in S_0 and S_1. In addition, we let $\mathcal{L}$ denote the set of labeled vertices $\{v_1, v_2, \ldots, v_{N_L}\}$ in the graph, where N_L is the number of labeled vertices, and let $Y = \{y_1, y_2, \ldots, y_{N_L}\}$ denote the corresponding set of ground-truth labels, with y_i representing the ground-truth label for vertex v_i. Additionally, let $\mathcal{U}$ represent the set of unlabeled vertices $\{v_{N_L+1}, v_{N_L+2}, \ldots, v_{N_L+N_U}\}$, where N_U is the number of unlabeled vertices. For any unlabeled vertex v_i, the predicted label is denoted as $\hat{y}_i$. It is important to note that the union of the labeled and unlabeled vertex sets equals the entire set of vertices in the graph, *i.e.,* $\mathcal{L} \bigcup \mathcal{U} = \mathcal{V}$.

3.2 Problem Definition

Given $\mathcal{G}$ with the sensitive attribute information S, node features $\mathbf{X}$, graph topology $\mathbf{A}$, and node labels $\mathbf{Y}$, our goal is to learn a node classifier, denoted as f_θ, which is parameterized by θ, trained based on $\mathcal{G}$ that can balance both individual fairness and group fairness.

4 Methodology

Existing fairness works often treat individual and group fairness as separate optimization goals, leading to enhanced individual fairness at the cost of group fairness. Instead, our proposed model aims to measure individual fairness by pinpointing the origins of individual bias, which in turn enables the quantification of associated group-level bias as a unified goal. Section 4.1 proposes a new metric, IRGF@k, to identify the origins of individual bias from a ranking perspective. The intuition is that similar individuals should hold similar positions in the corresponding ranking lists within the input space and output space, and these lists are formed through different reference individuals. This criterion allows for a direct evaluation of individuals who face unfair treatment and serves as a basis for assessing group-level bias. Then Section 4.2 introduces a unified metric, IG^{2}F, to measure the disparity in individual unfairness between subgroups, facilitating the quantification of group-aware individual fairness. Finally, Section 4.3 delves into the loss functions for each optimization objective and presents the objective function for optimizing our framework.

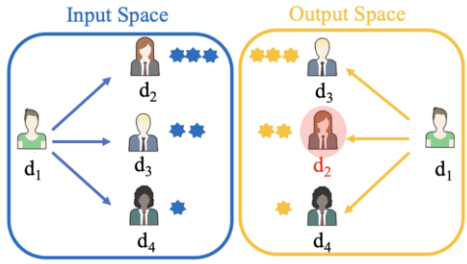

Figure 2: A toy example of quantifying and mitigating individual unfairness from a ranking perspective. The number of star(s) next to each individual represents the corresponding pairwise level of similarity to the reference individual.

4.1 Quantifying Individual Graph Unfairness

This section introduces a novel individual fairness metric, *Individual Rearranged Graph Fairness (IRGF)*, which measures individual fairness through a ranking perspective, offering a refined approach to evaluating bias by examining changes in individuals' positions within ranking lists, as opposed to focusing on the reference individual who forms the list. This strategy identifies individuals who face genuine discrimination while avoiding the limitations of the Lipschitz condition. Specifically, IRGF assesses the fairness of an individual d_i by evaluating his/her positional changes in the ranking lists established based on the reference individual d_r. To illustrate the overarching concept, take the example shown in Figure 2 with d_1 as the reference individual, the input space ranking list (d_2, d_3, d_4) and output space ranking list (d_3, d_2, d_4) are both arranged in descending order of similarity to d_1. To quantify the bias on d_2 from a ranking standpoint, we evaluate the ranking shift of d_2 in relation to the reference individual (*i.e.,* d_1), as well as the changes in relative positions between d_2 and their neighbors. Finally, the overall individual unfairness of d_2 is determined by calculating the average value of d_2's individual unfairness across all his/her associated ranking lists. Mathematically, the IRGF is defined as follows:

$$\mathrm{IRGF@k} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{1}{\mathrm{M}} \sum_{\mathrm{r}=1}^{\mathrm{M}} \frac{\mathrm{CNF}^{\mathrm{d_r}}_{\mathrm{Sim_{D'}}}(\mathrm{d_i})}{\mathrm{CNF}^{\mathrm{d_r}}_{\mathrm{Sim_D}}(\mathrm{d_i})} \tag{1}$$

where N represents the total number of individuals, M is the number of times d_i appears as one of the top-k neighbors of d_r, and $Sim_{(\cdot)}$ denotes the similarity matrix in the input space D (output space D'). Note that the focus on top-k neighbors aligns with the principle of individual fairness, which stipulates that only similar individuals

should be treated similarly. In addition, the *Cumulative Node Fairness* $\mathrm{CNF}^{\mathrm{d_r}}_{\mathrm{Sim}_{(\cdot)}}(\mathrm{d_i})$, motivated by learning to rank [4], is calculated for each focused individual d_i, with d_r as the reference individual, as follows:

$$\mathrm{CNF}^{\mathrm{d_r}}_{\mathrm{Sim}_{(\cdot)}}(\mathrm{d_i}) = \frac{Sim_D(d_i)}{\log_2(\mathrm{pos}(d_i)+1)} + \frac{1}{k-1}\sum_{\substack{\mathrm{pos}=1 \\ d_{l_{\mathrm{pos}}}\neq d_i}}^{k} \frac{Sim_D(d_{l_{\mathrm{pos}}}, d_r)}{\log_2(\mathrm{pos}+1)} \tag{2}$$

where pos is the position of each individual in the ranking list, and the sequence $\{l_{\mathrm{pos}}\}_{\mathrm{pos}=1}^{k}$ represents the ordered list of individual indices from the similarity matrix $Sim(\cdot)$ for the reference individual d_r. Thus, the first tern in the Equation 2 assesses the bias related to the individual's positional change, while the second term evaluates the bias impact due to the positional changes of the individual's neighbors (*i.e.*, $d_{l_{\mathrm{pos}}} \neq d_i$). Note that both $\mathrm{CNF}_{\mathrm{Sim}_D(d_i)}$ and $\mathrm{CNF}_{\mathrm{Sim}_{D'}(d_i)}$ are computed using the similarity values from Sim_D (*i.e.,* input space), with the corresponding similarity being used only for deriving the ordering list l_{pos}, which directly quantifies individual bias while eliminating the necessity of specifying a Lipschitz constant. In addition, the input space oracle similarity matrix Sim_D is often problem-specific and given a priori [20, 19]; For the output space similarity matrix $Sim_{D'}$, the Gaussian Kernel function is applied to measure the output space similarity of each pair node as follows:

$$Sim_{D'}(v_i, v_j) = \exp\left(-\frac{D(f(v_i), f(v_j))^2}{2\sigma^2}\right) \tag{3}$$

where σ here is a scalar, and $D(\cdot)$ is a chosen distance metric, and $f(\cdot)$ represents a GNNs model.

Overall, our proposed IRGF@k values lie within the interval $[0, 1]$, which aligns with existing notions of individual fairness. Higher IRGF@k scores indicate greater consistency between the ranking lists encoded from the input and output spaces, which suggests a fairer model. In other words, if two individuals are ranked closely in the input space (*e.g.,* their personal circumstances), then they should also be ranked closely in the output space (*e.g.,* their application results). By scrutinizing the two ranking lists obtained in the input and output spaces, our approach can pinpoint the origins of individual bias while also becoming feasible to quantify associated unified group-level bias further.

4.2 Quantifying Individual-Group Graph Unfairness

A critical limitation in existing individual fairness works is their insufficient consideration of group fairness implications. Specifically, individual fairness constraints might differ across subgroups, and such disparities may result in individuals in poor groups being treated differently in terms of meeting individual fairness, thereby leading to increased biases against deprived groups. To address this, we introduce a novel unified metric that quantitatively measures this disparity, thereby bridging the gap between individual and group fairness. This metric leverages our previously introduced individual fairness metric IRGF@k to evaluate if subgroups are treated fairly by the model. Specifically, we evaluate the ranking consistency ratio of each individual d_i within a subgroup, thereby evaluating the model's disparity across each subgroup. This strategy enables us to scrutinize the fairness of treatment across subgroups, thereby offering a more refined and comprehensive fairness assessment. In other words, we consider group-aware individual fairness to be satisfied if every individual in a group is not significantly better than an individual in another group. Formally, *Group Graph Fairness* (G^2F) is defined as follows:

$$\mathrm{G^2F}_{\mathrm{D_{s_i}}} = \frac{1}{|\mathrm{D_{s_i}}|}\sum_{\mathrm{i}=1}^{|\mathrm{D_{s_i}}|}\frac{1}{\mathrm{M_{d_i}}}\sum_{\mathrm{r}=1}^{\mathrm{M_{d_i}}}\frac{\mathrm{CNF}_{\mathrm{Sim}_{D'}(\mathrm{d_i})}}{\mathrm{CNF}_{\mathrm{Sim}_{\mathrm{D}}(\mathrm{d_i})}} \tag{4}$$

where $|\cdot|$ represent the number of samples in the subgroup $\mathbf{D}_{s_i}$. Utilizing $\mathrm{G^2F}_{\mathrm{D_{s_i}}}$, *Individual-Group Graph Fairness (IG^2F)*, designed to identify and quantify disparities in individual fairness performance between distinct subgroups $\mathbf{D}_{s_i}$ and $\mathbf{D}_{s_j}$, is introduced in Equation 5:

$$\mathrm{IG^2F} = \max_{\forall\, \mathbf{D}_{s_i}, \mathbf{D}_{s_j} \in \mathbf{D}, \mathbf{D}_{s_i} \neq \mathbf{D}_{s_j}} \left\{ \left| \frac{\mathrm{Min}(\mathrm{G^2F}_{\mathbf{D}_{s_i}}, \mathrm{G^2F}_{\mathbf{D}_{s_j}})}{\mathrm{Max}(\mathrm{G^2F}_{\mathbf{D}_{s_i}}, \mathrm{G^2F}_{\mathbf{D}_{s_j}})} \right| \right\} \tag{5}$$

The $\mathrm{IG^2F}$ value, ranging from 0 to 1, provides a quantitative measure of fairness across subgroups. A value of 1 indicates equivalent levels of individual fairness among all subgroups, representing unbiased treatment. Conversely, a value of 0 suggests a significant disparity in treatment between subgroups, indicating biased outcomes. Thus, $\mathrm{IG^2F}$ effectively quantifies the maximal disparity in individual fairness performance among different subgroups, calculated by considering all possible subgroup combinations.

4.3 Mitigating Individual and Group-level Unfairness

This section proposes a novel GNN framework *Group Equality Individual Fairness (GEIF)* to mitigate individual- and group-level unfairness collectively, consisting of three distinct modules: i) the utility module, ii) the individual fairness module, and iii) the group-level fairness module. Specifically, the utility module aims to maximize the performance of the backbone GNN model in advancing downstream learning tasks. To achieve this, we adopt the cross-entropy loss to enforce the predictions $\hat{y}_i$ to be closer to the ground truth y_i, defined as:

$$\mathcal{L}_U = \frac{1}{|\mathcal{V}_L|}\sum_{v_i \in \mathcal{V}_L} -(y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)) \tag{6}$$

Second, we introduce an individual fairness module designed to mitigate individual unfairness. As we introduced in Section 4.1, we treat individual fairness as a ranking problem, thereby alleviating the dependence on non-trivial Lipschitz condition. As observed in Figure 2, while d_4 maintains consistent positions across input and output spaces, d_2 and d_3 swap their respective orderings, indicating inconsistency. We want a loss function to promote the first relative ordering while penalizing the latter. However, defining a loss function minimizing the difference between Sim_D and $Sim_{D'}$ with gradient-based optimization techniques has the problem that is non-differentiable due to ranking operations. To address this, we formulate the loss function as a probability function, enabling the application of gradient-based optimization techniques. To formulate it as a probability score between 0 and 1, we make use of the Sigmoid function, as shown as follows:

$$\hat{P}_{d_i,d_e} = \frac{1}{1+e^{-(\hat{Sim}_D(d_r,d_i) - \hat{Sim}_D(d_r,d_e))}} \tag{7}$$

Furthermore, for P_{d_i,d_e}, representing the known probability, we

use a priori information to determine if d_i is more similar to d_r than d_e when d_r as reference individual in the input space, defined as:

$$P_{d_i,d_e} = \begin{cases} 1, & \text{if } \mathrm{Sim_D}(\mathrm{d_r}, \mathrm{d_i}) > \mathrm{Sim_D}(\mathrm{d_r}, \mathrm{d_e}) \\ 0.5, & \text{if } \mathrm{Sim_D}(\mathrm{d_r}, \mathrm{d_i}) = \mathrm{Sim_D}(\mathrm{d_r}, \mathrm{d_e}) \\ 0, & \text{if } \mathrm{Sim_D}(\mathrm{d_r}, \mathrm{d_i}) < \mathrm{Sim_D}(\mathrm{d_r}, \mathrm{d_e}) \end{cases} \tag{8}$$

Building on this, to promote individual fairness via ranking optimization, it is necessary to quantify and minimize the difference between the predicted probability distribution and the known one. With this, the loss function $\mathcal{L}_{d_i}(\cdot)$ is defined as the cross-entropy loss that quantifies the difference between the predicted and actual probability distributions of order consistency:

$$\mathcal{L}_{d_i}^{\mathrm{TOP(d_i)}} = -\frac{1}{M}\sum_{j=1}^{M}\Big(P_{d_i,d_e}\log\hat{P}_{d_i,d_e} + (1-P_{d_i,d_e})\log(1-\hat{P}_{d_i,d_e})\Big) \tag{9}$$

where $\mathrm{TOP}(\cdot)$ operation selects all individuals d_r for whom d_i and d_e are among the top ranked k similar individuals in the input space. Finally, the total individual fair loss $\mathcal{L}_I$ is defined as follows:

$$\mathcal{L}_I = \sum_{i=1}^{N}\mathcal{L}_{d_i}^{\mathrm{TOP(d_i)}}(d_i, d_e) \tag{10}$$

Overall, this similarity-based relative ranking difference in the input and output space represents the ranking inconsistency loss and is quantified as $\mathcal{L}_{d_i}^{\mathrm{TOP(d_i)}}$. The overall individual loss function, $\mathcal{L}_I$, aggregates this inconsistency loss across all individuals.

Finally, based on the foundation laid by our individual fairness metric, we introduce a group-level fairness module designed to promote group equality of individual fairness. Specifically, we cast group fairness guarantee as the difference between the mean individual fairness loss of the different subgroups. In other words, this approach is to prevent any single group from disproportionately bearing the brunt of individual fairness loss. The loss function should thus promote consistency in the mean individual fairness loss of different groups. Motivated by this, we define a differentiable loss function with pairwise $\mathbf{D}_{s_i}$ and $\mathbf{D}_{s_j}$ for minimizing disparities in fairness across subgroups. The group fairness loss ($\mathcal{L}_G$) is defined as follows:

$$\mathcal{L}_G = \left(\frac{\mathrm{G^2F_{\mathbf{D}_{s_i}}}}{\mathrm{G^2F_{\mathbf{D}_{s_j}}}} - 1\right)^2 + \left(\frac{\mathrm{G^2F_{\mathbf{D}_{s_j}}}}{\mathrm{G^2F_{\mathbf{D}_{s_i}}}} - 1\right)^2 \tag{11}$$

where $\mathrm{G^2F_{\mathbf{D}_{s_i}}}$ and $\mathrm{G^2F_{\mathbf{D}_{s_j}}}$ are individual unfairness for subgroup $\mathbf{D}_{s_i}$ and $\mathbf{D}_{s_j}$. Each group's individual unfairness is computed with our proposed individual metrics according to Equation 1. Notably, this loss function is symmetric, ensuring that its application is consistent and fair regardless of the subgroup pairing order. By adopting this broader approach, our method effectively evaluates and addresses potential fairness issues within machine learning models, ensuring that individual fairness outcomes are more equitable across diverse subgroups.

In summary, there are three objectives in total for the optimization of GEIF, as depicted in Equation 12. The first term, $\mathcal{L}_U$, focuses on maintaining the utility of the GNNs model (*i.e.,* minimizing the prediction loss). For this purpose, we adopt the cross-entropy loss, which is widely utilized in node classification tasks. The next term, $\mathcal{L}_I$, targets minimizing the individual fairness loss. As discussed in section 4.1, this term enhances individual fairness by promoting consistent ordering in the input and output spaces, thereby reducing reliance on the Lipschitz condition. The final term, $\mathcal{L}_G$, aims to equalize levels of individual unfairness for all subgroups, thereby addressing group fairness.

$$\arg\min\ \mathcal{L}_{total} = \mathcal{L}_U + \alpha\mathcal{L}_I + \beta\mathcal{L}_G \tag{12}$$

where α and β are tunable hyperparameters that control the strength of the individual and group fairness constraints, respectively. These parameters allow for the flexible balancing of fairness objectives with the model's predictive accuracy. For training our proposed framework, gradient-based optimization techniques can be directly applied to minimize $\mathcal{L}_{total}$, ensuring an effective and efficient approach to achieving both utility and fairness in the model's outcomes.

5 Experiments

5.1 Datasets

Four real-world datasets are utilized to evaluate the effectiveness of our framework: i) **Facebook Dataset** [22]: This dataset represents the Ego Network within Facebook, where nodes are Facebook users, and edges denote friendships. The classification task involves predicting whether users belong to the same social circle based on their network connections. ii) **German Dataset** [2]: Comprising credit information from a German bank, this dataset represents clients as nodes and their credit account similarities as edges. The sensitive attribute is gender, with the classification objective being to differentiate between good and bad credit risks. iii) **Credit Dataset** [51]: This dataset contains individuals' default payment information. Nodes represent individuals, and edges reflect similarities in expenditure and payment patterns. The sensitive attribute here is age, with the goal to predict whether individuals prefer credit card payments as their default mode. iv) **Bail Dataset** [1]: This dataset presents data related to defendants who were granted bail in U.S. state courts. In this context, each node corresponds to a defendant, while an edge connecting two nodes signifies similarities in their criminal records and demographic details. The sensitive attribute is the race of the defendants, and the objective is to classify them into suitable or unsuitable categories for bail. Table 1 summarizes detailed statistics of these datasets.

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

Dataset	Facebook	German	Credit	Bail
Vertices	1,034	1,000	30,000	18,876
Edges	26,749	21,742	137,377	311,870
Feature Dimension	224	27	13	18
Average Degree	51.7	44.5	10	34
Sensitive Attribute	Gender	Gender	Age	Race

5.2 Evaluation Protocol

Our evaluation encompasses a range of fairness and performance metrics to provide a thorough assessment. To gauge fairness, the proposed graph fairness metrics IRGF@10 and $\mathrm{IG^2F}$ are employed. Note that existing, commonly used fairness metrics are not applicable, as they typically do not account for the inherent implications between individual and group fairness. In evaluating performance, we align with prior works [64, 7, 23] and consider two widely-used node classification

performance metrics: *i.e.,* accuracy and F1-Score. For all of them, higher values correspond to better performance. To demonstrate the generalization of GEIF, we build $\text{Sim}_{D'}$ by calculating the cosine similarity between node representations.

5.3 Baselines

To benchmark the performance, GEIF is compared against six state-of-the-art methods: GCN [18], FairGNN [7], PFR [20], InFoRM [16], REDRESS [9], and GUIDE [32]. Specifically, GCN serves as a performance-driven vanilla baseline, which leverages spatial graph convolutions for neighbor representation aggregation. The remaining methods focus on various aspects of fairness: FairGNN employs adversarial learning to enforce group fairness in node classifications; PFR focuses on learning fair node embeddings as a preprocessing step, thereby satisfying individual fairness in downstream tasks; InFoRM applies a Lipschitz condition to formulate an individual fairness loss in graphs; REDRESS aims to achieve individual fairness from a ranking perspective; and GUIDE targets both individual and group fairness, also based on the Lipschitz condition.

5.4 Implementation Details

Our proposed model is designed to be flexible and not constrained to a specific GNNs architecture. For the purposes of our experiments, we employ the GCN as the GNN backbone. The number of GCN layers is 2, and we set the hidden size as 16. The activation function is ReLU. We use the Adam optimizer [17] to train the classification model with 1×10^{-4} learning rate and 1×10^{-4} weight decay. We conducted all experiments 10 times and reported average results. For fairness and to achieve optimal performance across all models tested, we tune the hyperparameters based on each method's performance on the validation set.

5.5 Experiment results

Benchmark Performance. The effectiveness of the GEIF framework is evaluated and summarized in Table 2. From the perspective of model utility, GEIF demonstrates competitive performance and achieves the highest accuracy on the Credit dataset. In terms of F1-score, GEIF ranks highest on the Bail dataset and second highest on the Credit dataset. This notable improvement, particularly in deprived subgroups, is attributed to GEIF's comprehensive integration of fairness considerations into the model. From the perspective of fairness, GEIF outperforms all the baseline methods. It efficiently identifies and mitigates individual unfairness by analyzing ranked lists from both input and output spaces. Furthermore, GEIF effectively reduces the performance disparities across different subgroups, leading to a significant enhancement in group fairness compared to existing individual fairness baselines. From the perspective of balancing the model utility and fairness, across four datasets and four metrics, it achieves top rankings in most categories, illustrating its superiority in managing the trade-off between accuracy and fairness. Overall, the integration of both individual and group fairness considerations allows GEIF to outperform models focused on only one aspect of fairness, illustrating its comprehensive advantage in fostering fairness.

Parameter Studies. The proposed GEIF framework has two critical hyperparameters: α is key in optimizing individual fairness, while β focuses on balancing individual-group fairness objectives. To explore their impacts on performance and fairness, hyperparameter sensitivity experiments are conducted. These experiments varied α and β within the set $\{1e^{-3}, 1e^{-2}, 1e^{-1}, 1e^{0}, 1e^{1}, 1e^{2}, 1e^{3}\}$. Taking the Facebook dataset as an example, Figure 3 illustrates the outcomes of these experiments. As one can see, increments in α and β tend to be inversely proportional to GEIF's performance. Specifically, increasing α enhances the model's individual fairness but can lead to a noticeable decline in overall performance. Conversely, elevating β has a positive effect on equalizing the variances in individual fairness performance between different subgroups. This is because a higher β value allows the model to more effectively balance fairness across these subgroups, addressing disparities more holistically.

Table 2: Performance and fairness results for GEIF and baselines. The darkest cells indicate the top rank, while lighter cells represent the second rank.

Dataset	Methods	Accuracy	F1-score	IRGF@10	IG^2F
Facebook	GCN	0.78	0.77	0.36	0.33
	PFR	0.73	0.73	0.43	0.29
	InFoRM	0.74	0.75	0.64	0.26
	REDRESS	0.77	0.80	0.51	0.31
	FairGNN	0.82	0.78	0.38	0.35
	GUIDE	0.79	0.79	0.61	0.47
	GEIF	0.79	0.78	0.71	0.67
German	GCN	0.66	0.76	0.42	0.37
	PFR	0.66	0.74	0.47	0.26
	InFoRM	0.62	0.71	0.65	0.28
	REDRESS	0.68	0.78	0.56	0.28
	FairGNN	0.63	0.69	0.39	0.47
	GUIDE	0.67	0.77	0.63	0.53
	GEIF	0.66	0.76	0.69	0.61
Credit	GCN	0.61	0.72	0.37	0.32
	PFR	0.64	0.76	0.51	0.25
	InFoRM	0.68	0.77	0.67	0.27
	REDRESS	0.65	0.74	0.51	0.28
	FairGNN	0.63	0.78	0.42	0.44
	GUIDE	0.68	0.81	0.65	0.49
	GEIF	0.69	0.78	0.71	0.63
Bail	GCN	0.82	0.78	0.34	0.37
	PFR	0.77	0.71	0.47	0.27
	InFoRM	0.80	0.80	0.57	0.20
	REDRESS	0.78	0.71	0.53	0.24
	FairGNN	0.81	0.77	0.36	0.52
	GUIDE	0.79	0.73	0.61	0.57
	GEIF	0.81	0.81	0.68	0.65

Effect of Different Number of Neighbors k Values. In the GEIF framework, the number of neighbors k plays a crucial role in determining both model fairness and predictive performance. To explore this, we conducted experiments with various k values in the set $\{4, 7, 10, 15, 20, 30, 50\}$, keeping other training factors constant. As illustrated in Figure 4, our findings reveal that as k increases, there is an enhancement in the model's performance on IRGF@k and IG^2F metrics, indicating more effective optimization for both individual and group fairness. However, the model's predictive accuracy remains largely unaltered when k is modest indicating an optimal balance between fairness and performance within this range. Conversely, when k values are significantly increased, there is a trend towards reduced precision due to the introduction of more noise from a larger number of samples per comparison, which can diminish the weights assigned to accurate labels and lead to ambiguous categorizations. Overall, a k value of 10 achieves the best measure of fairness and performance.

Ablation Studies. To further assess the design of GEIF, ablation studies are conducted by varying the loss function components of the model. Specifically, the GEIF variant, GEIF-, created by setting β to zero, eliminates the group fairness loss ($\mathcal{L}_{\text{G}}$) impact, enabling assessment when solely focusing on individual fairness. The ablation study results, shown in Table 3, indicate that while the GEIF- variant slightly improves individual fairness metrics over the complete GEIF model, it leads to a noticeable decrease in both individual-group fairness and overall predictive performance. This decline is attributable

Figure 3: Exploring hyperparameters study results in the Facebook dataset

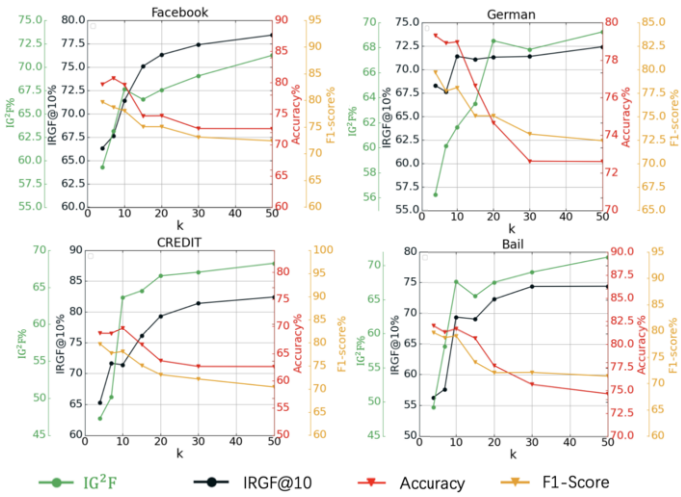

Figure 4: Exploring the choice of k-value effect model performance and fairness.

to GEIF- not accounting for the variations in individual fairness constraints across different subgroups, leading to diminished fairness and effectiveness, particularly for deprived groups. This ablation study confirms the efficacy of GEIF's design, which incorporates individual fairness with group awareness, and highlights the significance of considering group disparities in individual fairness.

Table 3: Ablation study results for GEIF and GEIF-.

Dataset	IRGF@10		IG^2F	
	GEIF-	GEIF	GEIF-	GEIF
Facebook	0.73	0.71	0.26	0.67
German	0.71	0.69	0.33	0.61
Credit	0.73	0.71	0.27	0.63
Bail	0.70	0.68	0.28	0.65
Dataset	Accuracy		F1-score	
	GEIF-	GEIF	GEIF-	GEIF
Facebook	0.80	0.79	0.76	0.78
German	0.65	0.66	0.73	0.76
Credit	0.69	0.69	0.75	0.78
Bail	0.80	0.81	0.77	0.81

6 Conclusion

This paper takes the first major step towards exploring the intricate interaction between individual fairness and group fairness in the graphs. Specifically, two novel fairness concepts designed to quantify individual unfairness and associated disparities across subgroups are introduced. Utilizing these concepts, a unified debiasing algorithm is developed to mitigate individual unfairness and associated group biases collectively. Experimental findings demonstrate the proposed method's superiority in achieving graph fairness compared to current state-of-the-art approaches. Future directions include expanding the proposed methodology to various other graph mining models.

Acknowledgement

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

References

[1] C. Agarwal, H. Lakkaraju, and M. Zitnik. Towards a unified framework for fair and stable graph representation learning. In *Uncertainty in Artificial Intelligence*, pages 2114–2124. PMLR, 2021.
[2] A. Asuncion and D. Newman. Uci machine learning repository, 2007.
[3] S. Bhagat, G. Cormode, and S. Muthukrishnan. Node classification in social networks. *arXiv preprint arXiv:1101.3291*, 2011.
[4] C. Burges, R. Ragno, and Q. Le. Learning to rank with nonsmooth cost functions. *Advances in neural information processing systems*, 19, 2006.
[5] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.
[6] Z. Chu, Z. Wang, and W. Zhang. Fairness in large language models: A taxonomic survey. *ACM SIGKDD Explorations Newsletter, 2024*, pages 34–48, 2024.
[7] E. Dai and S. 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.
[8] T. V. Doan, Z. Chu, Z. Wang, and W. Zhang. Fairness definitions in language models explained. *arXiv preprint arXiv:2407.18454*, 2024.
[9] Y. Dong, J. Kang, H. Tong, and J. Li. Individual fairness for graph neural networks: A ranking based approach. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 300–310, 2021.
[10] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.
[11] D. García-Soriano and F. Bonchi. Maxmin-fair ranking: individual fairness under group-fairness constraints. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 436–446, 2021.
[12] W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.
[13] W. L. Hamilton. Graph representation learning. *Synthesis Lectures on Artifical Intelligence and Machine Learning*, 14(3):1–159, 2020.
[14] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.
[15] Z. Jiang, X. Han, C. Fan, Z. Liu, N. Zou, A. Mostafavi, and X. Hu. Fmp: Toward fair graph message passing against topology bias. *arXiv preprint arXiv:2202.04187*, 2022.
[16] J. Kang, J. He, R. Maciejewski, and H. Tong. Inform: Individual fairness on graph mining. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 379–389, 2020.
[17] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
[18] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[19] P. Lahoti, K. P. Gummadi, and G. Weikum. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 ieee 35th international conference on data engineering (icde)*, pages 1334–1345. IEEE, 2019.

[20] P. Lahoti, K. P. Gummadi, and G. Weikum. Operationalizing individual fairness with pairwise fair representations. *arXiv preprint arXiv:1907.01439*, 2019.

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

[22] J. Leskovec and J. Mcauley. Learning to discover social circles in ego networks. *Advances in neural information processing systems*, 25, 2012.

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

[24] C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan, and M. Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 4602–4609, 2019.

[25] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

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

[27] N. A. Saxena, W. Zhang, and C. Shahabi. Missed opportunities in fair ai. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, pages 961–964. SIAM, 2023.

[28] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*, pages 1670–1679. PMLR, 2016.

[29] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–93, 2008.

[30] S. Sharifi-Malvajerdi, M. Kearns, and A. Roth. Average individual fairness: Algorithms, generalization and experiments. *Advances in neural information processing systems*, 32, 2019.

[31] V. Shumovskaia, K. Fedyanin, I. Sukharev, D. Berestnev, and M. Panov. Linking bank clients using graph neural networks powered by rich transactional data: Extended abstract. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 787–788, 2020.

[32] W. Song, Y. Dong, N. Liu, and J. Li. Guide: Group equality informed individual fairness in graph neural networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1625–1634, 2022.

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

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

[35] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, and P. Lio. ', and yoshua bengio. *Graph attention networks. ArXiv, abs/1710.10903*, 2(6): 13, 2018.

[36] J. Von Neumann and O. Morgenstern. Theory of games and economic behavior, 2nd rev. 1947.

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

[38] Y. Wang, Y. Zhao, Y. Dong, H. Chen, J. Li, and T. Derr. Improving fairness in graph neural networks via mitigating sensitive attribute leakage. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1938–1948, 2022.

[39] Z. Wang and W. Zhang. Group fairness with individual and censorship constraints. In *27th European Conference on Artificial Intelligence*. 2024.

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

[41] Z. Wang, N. Saxena, T. Yu, S. Karki, T. Zetty, I. Haque, S. Zhou, D. Kc, I. Stockwell, A. Bifet, et al. Preventing discriminatory decision-making in evolving data streams. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023.

[42] Z. Wang, C. Wallace, A. Bifet, X. Yao, and W. Zhang. Fg^2an: Fairness-aware graph generative adversarial networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 259–275. Springer Nature Switzerland, 2023.

[43] Z. Wang, Y. Zhou, M. Qiu, I. Haque, L. Brown, Y. He, J. Wang, D. Lo, and W. Zhang. Towards fair machine learning software: Understanding and addressing model bias through counterfactual thinking. *arXiv preprint arXiv:2302.08018*, 2023.

[44] Z. Wang, Z. Chu, R. Blanco, Z. Chen, S.-C. Chen, and W. Zhang. Advancing graph counterfactual fairness through fair representation learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2024.

[45] Z. Wang, Z. Chu, T. V. Doan, S. Ni, M. Yang, and W. Zhang. History, development, and principles of large language models-an introductory survey. *arXiv preprint arXiv:2402.06853*, 2024.

[46] Z. Wang, J. Dzuong, X. Yuan, Z. Chen, Y. Wu, X. Yao, and W. Zhang. Individual fairness with group awareness under uncertainty. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2024.

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

[48] Y. Wei, P. Yildirim, C. Van den Bulte, and C. Dellarocas. Credit scoring with social network data. *Marketing Science*, 35(2):234–258, 2016.

[49] J. Wu, X. Wang, F. Feng, X. He, L. Chen, J. Lian, and X. Xie. Self-supervised graph learning for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 726–735, 2021.

[50] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

[51] I.-C. Yeh and C.-h. 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.

[52] H. Yuan, H. Yu, S. Gui, and S. Ji. Explainability in graph neural networks: A taxonomic survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[53] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

[54] S. Zhang, D. Zhou, M. Y. Yildirim, S. Alcorn, J. He, H. Davulcu, and H. Tong. Hidden: hierarchical dense subgraph detection with application to financial fraud detection. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 570–578. SIAM, 2017.

[55] W. 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.

[56] W. Zhang. Ai fairness in practice: Paradigm, challenges, and prospects. *Ai Magazine*, 2024.

[57] W. Zhang and E. Ntoutsi. Faht: an adaptive fairness-aware decision tree classifier. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1480–1486, 2019.

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

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

[60] W. Zhang and J. C. Weiss. Fairness with censorship and group constraints. *Knowledge and Information Systems*, pages 1–24, 2023.

[61] W. Zhang, A. Bifet, X. Zhang, J. C. Weiss, and W. Nejdl. Farf: A fair and adaptive random forests classifier. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 245–256. Springer, 2021.

[62] W. Zhang, L. Zhang, D. Pfoser, and L. Zhao. Disentangled dynamic graph deep generation. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 738–746. SIAM, 2021.

[63] W. Zhang, S. Pan, S. Zhou, T. Walsh, and J. C. Weiss. Fairness amidst non-iid graph data: Current achievements and future directions. *arXiv preprint arXiv:2202.07170*, 2022.

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

[65] W. Zhang, Z. Wang, J. Kim, C. Cheng, T. Oommen, P. Ravikumar, and J. Weiss. Individual fairness under uncertainty. In *26th European Conference on Artificial Intelligence*, pages 3042–3049, 2023.

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