GUIDE: Group Equality Informed Individual Fairness in Graph Neural Networks

Weihao Song
University of Virginia
ws5dw@virginia.edu

Yushun Dong
University of Virginia
yd6eb@virginia.edu

Ninghao Liu
University of Georgia
ninghao.liu@uga.edu

Jundong Li
University of Virginia
jundong@virginia.edu

ABSTRACT

Graph Neural Networks (GNNs) are playing increasingly important roles in critical decision-making scenarios due to their exceptional performance and end-to-end design. However, concerns have been raised that GNNs could make biased decisions against underprivileged groups or individuals. To remedy this issue, researchers have proposed various fairness notions including individual fairness that gives similar predictions to similar individuals. However, existing methods in individual fairness rely on Lipschitz condition: they only optimize overall individual fairness and disregard equality of individual fairness between groups. This leads to drastically different levels of individual fairness among groups. We tackle this problem by proposing a novel GNN framework GUIDE to achieve group equality informed individual fairness in GNNs. We aim to not only achieve individual fairness but also equalize the levels of individual fairness among groups. Specifically, our framework operates on the similarity matrix of individuals to learn personalized attention to achieve individual fairness without group level disparity. Comprehensive experiments on real-world datasets demonstrate that GUIDE obtains good balance of group equality informed individual fairness and model utility. The open-source implementation of GUIDE can be found here: https://github.com/mikesong724/GUIDE.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Applied computing** → **Law, social and behavioral sciences**.

KEYWORDS

individual fairness, graph neural networks

ACM Reference Format:
Weihao Song, Yushun Dong, Ninghao Liu, and Jundong Li. 2022. GUIDE: Group Equality Informed Individual Fairness in Graph Neural Networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3534678.3539346

KDD '22, August 14–18, 2022, Washington, DC, USA

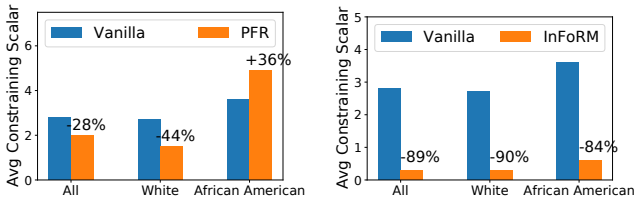
ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539346

1 INTRODUCTION

Graph data is ubiquitous across a number of high-impact domains, e.g., social networks [23], user-product graphs [29], and knowledge graphs [15], etc. In fact, graphs provide a useful abstraction to describe data and their inherent relations. To effectively gain deeper understanding from graph data, various algorithms have been proposed to tackle different graph mining tasks, including prediction [14], community detection [2], recommendation [11], and many more. Among them, Graph Neural Networks (GNNs) have attracted a surge of research interests in recent years due to their superior learning performances [12, 18, 24]. They are increasingly adopted in various tasks such as anomaly detection [26], social network recommendation [28], and graph classification [27]. Although GNNs have excelled in a diverse set of tasks, concerns have been raised that directly adopting GNNs could empirically result in ethical and fairness issues [1, 4, 6], such as racial or gender discrimination, which renders the adoption of GNNs in high-stake scenarios questionable. Generally, to analyze algorithmic fairness, researchers have developed multiple fairness notions [21], such as group fairness [13], which ensures equal outcome rates for members of different demographic subgroups; and individual fairness [9], which promotes treating similar individuals similarly. While group fairness has been widely studied in many works [19, 22], individual fairness still remains under-explored. Nevertheless, considering that individual fairness is able to enforce fairness at a finer granularity at the individual level compared to group fairness, it is a desirable notion of fairness to enforce in GNNs.

To model individual fairness in graphs, Lipschitz condition is the most commonly adopted mathematical foundation among existing works [9, 16, 20]. Specifically, for any pair of individuals (v_i, v_j), there is a constraining scalar $\epsilon_{i,j}$ such that the output distance between v_i and v_j is bounded by their input distance multiplied by this scalar. Here, the input distance between individuals could be given by domain experts or oracle similarity matrix [5, 16]. The largest constraining scalar across all pairs is named as the Lipschitz constant. Intuitively, if the Lipschitz constant is small, then the outcome distance is also constrained to be small for similar pairs (i.e., pairs with small input distances) in the dataset. This implies that *similar people are treated similarly*. To enforce individual fairness, a commonly adopted approach[1] in existing works [16, 20] is to

[1]Existing works use this loss form to optimize individual fairness: $\mathcal{L} = \sum_{v_i \in \mathcal{V}} \sum_{v_j \in \mathcal{V}} \|\mathbf{Z}[i,:] - \mathbf{Z}[j,:]\|_2^2 \mathbf{S}[i,j]$ and $\mathcal{L} \leq m\epsilon$, where $\mathcal{V}$ is the set of individuals, $\mathbf{Z}$ is the model output matrix, $\mathbf{S}$ is the pairwise similarity matrix of individuals in $\mathcal{V}$, m is the number of pairwise comparisons, ϵ is the average constraining scalar.



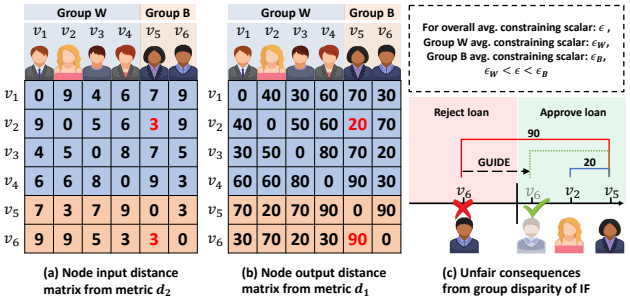
(a) GNN-PFR (b) GNN-InFoRM

Figure 1: Average constraining scalar optimization based on PFR and InFoRM. Percentages denote the optimization results compared with vanilla values. Group disparity is exacerbated because the White group receives better optimization results compared with the group of African American.

minimize the sum of output distance divided by input distance (or multiplied by similarity) for all pairs of individuals. Intuitively, this sum is bounded by the average constraining scalar ϵ (for all pairs of individuals in the dataset) times the total number of pairs m. Therefore, as this sum is minimized, the average constraining scalar ϵ is also minimized. Such technique has been empirically proved effective in forcing the Lipschitz constant to be as small as possible. However, for different individual pairs in the population, the optimization of the constraining scalar could be potentially influenced by the sensitive attributes of the involved individuals such as gender or race. For example, the optimization result of some privileged demographic subgroups could be significantly better than that of the disadvantaged groups. We empirically show the presence of such phenomenon in existing works below.

PFR [20] and InFoRM [16] are two representative works that optimize individual fairness in graphs. The effectiveness of these two approaches for the optimization of the average constraining scalar between the outcome and input distance for individual pairs has been empirically proved. Nevertheless, sensitive attributes (e.g., race) could severely affect the optimization results in both approaches. For example, empirical explorations are shown in Fig. 1, where PFR and InFoRM are adapted to GNNs for node classification task on Income dataset [8]. Generally, the overall average constraining scalar is optimized towards a smaller value, which indicates a smaller Lipschitz constant. However, such optimization effectiveness is largely attributed to the optimization of the constraining scalars for pairs that involve white individuals. For pairs that involve African Americans, they do not enjoy as much optimization in InFoRM, and their situation is even worse after the optimization of PFR.

It is worth mentioning that the group disparity of individual fairness optimization discussed above could lead to discrimination in real-world decision-making scenarios. Here we utilize an illustrative toy example in Fig. 2 to scrutinize how such group disparity leads to discrimination. Assume that two races (group W for white and group B for black) are involved in a loan approval system. Given the input distance matrix for individual pairs (i.e., the matrix in Fig. 2(a)), assume that the outcome distance (i.e., distance in the matrix of Fig. 2(b)) is already optimized through an existing individual fairness enforcing approach. In this example, the average constraining scalar for individual pairs involving members of group W (blue entries) is at a significantly lower level compared with that of group B (orange entries), i.e., $\epsilon_W < \epsilon_B$. Assume that there are a black (v_6)



(a) Node input distance matrix from metric d_2 (b) Node output distance matrix from metric d_1 (c) Unfair consequences from group disparity of IF

Figure 2: A toy example on the disparity of individual fairness between different groups in a loan approval system. Since existing individual fairness approaches do not enforce equal average constraining scalars between groups, it is possible that Group B has a larger average constraining scalar than Group W. Then, when comparing to an individual v_5 who is already approved the loan, a member v_6 from Group B could have a larger outcome distance than a member v_2 from Group W (90 vs. 20) even though their input distances to v_5 are equal (both are 3). This difference in outcomes could place v_6 across the decision boundary, giving him rejection when he is equally similar to v_5 compared to v_2.

and a white (v_2) individual who have equal input distance of 3 compared to a third individual v_5 who is already approved for the loan. Since the input distances of (v_2, v_5) and (v_6, v_5) are small, both pairs should be considered as similar pairs, and thus they should receive similar outcomes in the loan approval system. However, since existing individual fairness promotion approaches do not enforce equal average constraining scalars between groups, Group B could have a higher average constraining scalar than Group W. As a result, it is possible that the black individual v_6 has larger outcome distance to v_5 than the white individual v_2 has (90 vs 20), even though they have the same input distances of 3 to v_5. A larger outcome distance could potentially put the black individual v_6 on the other side of the decision boundary in the loan approval system, giving him a different outcome for his loan application.

To properly handle the problems we mentioned above, in this paper, we study a novel problem of enforcing group equality informed individual fairness. Specifically, we first design a metric to capture the disparity of individual fairness in groups. Then a novel GNN framework named GUIDE (Group eqUality Informed indiviDual fairnEss) is proposed to achieve not only overall individual fairness but also similar levels of individual fairness between groups, and moreover maintain the utility of the prediction model. In GUIDE, there are a backbone GNN learning from node adjacency matrix and node feature matrix to extract informative node embeddings for downstream tasks, and an attention-based GNN learning from the node similarity matrix and node embeddings to produce the final outputs that achieve our aforementioned objectives. Our main contributions can be summarized as:

- **Problem Formulation.** We propose a novel metric to quantitatively measure the group disparity of individual fairness. Based on the proposed metric, we formulate a novel problem of promoting group equality informed individual fairness in Graph Neural Networks.

Table 1: Symbols.

Symbols	Definitions
$\mathcal{G}$	input graph
$\mathcal{V}$	set of all nodes in a graph
$\mathcal{V}_p$	p^{th} group
n	total number of nodes
d	total number of features
m	total number of nonzero similarities in $\mathbf{S}$
m_p	nonzero similarities for members in $\mathcal{V}_p$
ϵ	overall average constraining scalar
ϵ_p	average constraining scalar for $\mathcal{V}_p$
$\mathbf{A} \in \{0,1\}^{n\times n}$	adjacency matrix of graph $\mathcal{G}$
$\mathbf{X} \in \mathbb{R}^{n\times d}$	node features matrix of graph $\mathcal{G}$
$\mathbf{Z} \in \mathbb{R}^{n\times c}$	graph learning output matrix
$\mathbf{S} \in \mathbb{R}^{n\times n}$	pairwise similarity matrix
$\mathbf{L} \in \mathbb{R}^{n\times n}$	Laplacian of similarity matrix

- **Algorithm Design.** We propose a novel framework GUIDE to relieve the disparity of individual fairness in different groups, optimize overall individual fairness while preserving prediction performance in downstream task.
- **Experimental Evaluation.** We conduct comprehensive experiments on multiple real-world datasets and experimental results validate the superiority of our proposed framework.

2 PROBLEM FORMULATION

In this section, we introduce notations and preliminaries on existing individual fairness approaches on graph learning. Finally, we formulate our research problem.

2.1 Notations

In this paper, we use bold uppercase characters (e.g., $\mathbf{A}$) for matrices, bold lowercase characters (e.g., $\mathbf{a}$) for vectors, lowercase characters (e.g., a) for scalars, uppercase caligraphic characters (e.g., $\mathcal{V}$) for sets. Also, we represent the i-th row, j-th column, (i,j)-th entry of a matrix $\mathbf{A}$ as $\mathbf{A}[i,:]$, $\mathbf{A}[:,j]$ and $\mathbf{A}[i,j]$, respectively. Additionally, we use lowercase bold vectors with index to represent the row vector of a matrix (e.g., $\mathbf{z}_i = \mathbf{Z}[i,:]$). The trace of matrix $\mathbf{A}$ is $\mathrm{Tr}(\mathbf{A})$. The ℓ_2-norm of a vector $\mathbf{a} \in \mathbb{R}^d$ is $\|\mathbf{a}\|_2 = \sqrt{\mathbf{a}^\mathbf{T}\mathbf{a}}$. A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ consists of (1) $\mathcal{V}$: set of nodes ($|\mathcal{V}| = n$), (2) $\mathcal{E} \in \mathcal{V} \times \mathcal{V}$: set of edges, and (3) $\mathbf{X} \in \mathbb{R}^{n\times d}$: node attributes where $\mathbf{x}_i \in \mathbf{R}^d$ is the attribute vector for i-th node. We assume there is a sensitive attribute set $\mathcal{T}$ containing sensitive attributes for each individual and $\mathcal{T}$ yields G disjoint groups in $\mathcal{V}$. We use $\mathcal{V}_p$ to denote the set of individuals in group p. We use $\mathbf{A} \in \{0,1\}^{n\times n}$, $\mathbf{X} \in \mathbb{R}^{n\times d}$ for graph adjacency matrix and node feature matrix, respectively. We also assume there is a similarity matrix $\mathbf{S}$ which contains pairwise node similarities according to domain knowledge or human judgement. It is worth noting that $\mathbf{S}$ may not be equal to $\mathbf{A}$. The Laplacian $\mathbf{L}$ of the similarity matrix $\mathbf{S}$ is derived by subtracting $\mathbf{S}$ from the diagonal degree matrix of $\mathbf{S}$. Symbols are summarized in Table 1.

2.2 Preliminaries

In this subsection, we first introduce the Lipschitz condition, which is commonly utilized to formulate an objective function to optimize individual fairness in existing works [16, 20]. We present Lipschitz condition in Definition 1 below.

Definition 1. *Lipschitz condition is satisfied if*

$$D_1(f(v_i), f(v_j)) \leq L.D_2(v_i, v_j), \forall v_i, v_j \in \mathcal{V} \tag{1}$$

where v_i and v_j are two instances (nodes specifically in our case), $D_1(.,.)$ and $D_2(.,.)$ are distance metrics for outputs and inputs respectively, $f(\cdot)$ is a function, and $L > 0$ is the Lipschitz constant.

Existing works that promote individual fairness in graph learning are commonly based on Lipschitz condition. The intuition is that individual pairs with higher similarity should be constrained to achieve smaller output distance. To achieve such property, the loss function in existing works is usually formulated as below:

$$\mathcal{L}_{\text{ifair}} = \frac{\sum_{v_i\in\mathcal{V}}\sum_{v_j\in\mathcal{V}} \|\mathbf{Z}[i,:] - \mathbf{Z}[j,:]\|_2^2\, \mathbf{S}[i,j]}{2} = \mathrm{Tr}(\mathbf{Z}^T\mathbf{L}\mathbf{Z}), \tag{2}$$

and this loss satisfies

$$2\mathcal{L}_{\text{ifair}} \leq m\epsilon, \tag{3}$$

where ϵ is the average constraining scalar for outcome distance given their similarities for all individual pairs, m is the number of nonzero elements in $\mathbf{S}$, $\mathbf{Z}$ is the graph learning output matrix (e.g., node embeddings for GNNs), $\mathbf{S}$ is the pairwise node similarity matrix, and $\mathbf{L}$ is the Laplacian of $\mathbf{S}$. Intuitively, the average constraining scalar ϵ is minimized accordingly as $\mathcal{L}_{\text{ifair}}$ is minimized. As such, similar individuals will receive smaller output distances on average, hence the overall individaul fairness level is improved. Note this loss is a relaxed form of Lipschitz condition (ϵ as the average constraining scalar instead of Lipschitz constant L as the absolute constraining scalar for all pairs), where $D_1(f(v_i), f(v_j))$ is $\|\mathbf{Z}[i,:] - \mathbf{Z}[j,:]\|_2^2$ and $D_2(v_i, v_j)$ is $\frac{1}{\mathbf{S}[i,j]}$.

As discussed in Section 1, minimizing $\mathcal{L}_{\text{ifair}}$ according to Eq. (2) only minimizes the average constraining scalar for all pairs of individuals. Nevertheless, the problem of how to reduce the disparity of the constraining scalars across different groups is ignored. As a result, when only $\mathcal{L}_{\text{ifair}}$ is minimized, one group can have dramatically lower constraining scalar (i.e., much higher level of individual fairness[2]) than the other group. Here we aim to properly handle such undesired fairness disparity between groups. In other words, we aim to achieve balanced levels of individual fairness across different groups. Specifically, let the level of individual fairness for a group $\mathcal{V}_p$ be U_p, we can formally define *Group Equality of Individual Fairness* as

Definition 2. *Group Equality Informed Individual Fairness is satisfied if the levels of individual unfairness for all groups are equal, i.e. for G disjoint groups in $\mathcal{V}$ ($\bigcup_{i=1}^{G}\mathcal{V}_i = \mathcal{V}$), $U_1 = U_2 = ... = U_G$.*

However, it would be difficult to achieve *Group Equality of Individual Fairness* in application scenarios. In this regard, a straightforward alternative goal is to satisfy it as much as possible. More specifically, when overall individual fairness level is maximized and the difference of individual fairness level across different groups is minimized, we deem that *group equality informed individual fairness* is achieved. Following such intuition, we formally formulate our research problem as follows.

[2]We associate high level of individual fairness with low level of individual unfairness in this paper.

Problem 1. ***Promoting Group Equality Informed Individual Fairness in GNNs.*** *Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$, ground truth labels $\mathbf{Y}$ for a given prediction task[3], a symmetric similarity matrix $\mathbf{S}$ for nodes in $\mathcal{V}$, G disjoint groups classified by sensitive attributes (i.e., $\bigcup_{i=1}^{G} \mathcal{V}_i = \mathcal{V}$), our goal is to learn an output $\mathbf{Z}$ (e.g., node embeddings) satisfying: (1) overall individual fairness level is maximized; (2) difference of the individual fairness level between different groups is minimized.*

To properly handle Problem 1, it is necessary to quantitatively measure both the overall individual fairness level and the difference of the individual fairness level between groups. Generally, the overall individual fairness level can be measured by metrics such as Eq. (2) [16]. In the following section, we define a metric to measure the difference of individual fairness level in different groups.

3 MEASURING GROUP DISPARITY

In this section, we develop a metric which can be utilized to measure the level of individual unfairness in one demographic subgroup. Subsequently, based on this group-level individual unfairness metric, we formally define a metric to quantify the disparity of individual unfairness between groups.

3.1 Individual Unfairness of a Group

We first describe the criterion of individual fairness for a group. The intuition of individual fairness is that individual pairs with higher similarity should be constrained with smaller output distance. Existing works evaluate this on all pairwise comparisons of individuals from the entire population $\mathcal{V}$, i.e. from $\mathcal{V} \times \mathcal{V}$. Then, to determine if a group $\mathcal{V}_p$ has individual fairness, we could evaluate the pairwise comparisons of individuals from this group against the entire population, i.e., from $\mathcal{V}_p \times \mathcal{V}$. The intuition is that when deciding for a group whether its members are treated fairly with respect to individual fairness, we should evaluate both intra-group and inter-group pairwise comparisons for completeness. Then, we can define a metric to quantitatively measure the level of individual (un)fairness for one group. In a similar fashion as Eq. (2), we define a group-level average constraining scalar ϵ_p for pairwise comparisons between members of $\mathcal{V}_p$ and all individuals in $\mathcal{V}$. Specifically, following Eq. (2), for a group $\mathcal{V}_p \subseteq \mathcal{V}$, we define a metric U_p to measure its group level of individual unfairness.

$$U_p = \frac{\sum_{v_i \in \mathcal{V}_p} \sum_{v_j \in \mathcal{V}} \|\mathbf{Z}[i,:] - \mathbf{Z}[j,:]\|_2^2 \, \mathbf{S}[i,j]}{m_p} \leq \epsilon_p, \quad (4)$$

where m_p is the number of nonzero pairwise similarities for members of $\mathcal{V}_p$ against all individuals in $\mathcal{V}$, and ϵ_p is the average constraining scalar for group $\mathcal{V}_p$. After introducing the metric for group-level individual fairness, we define a metric for measuring the group disparity of individual fairness in next subsection.

3.2 Group Disparity of Individual Fairness

In this subsection, we introduce a metric to measure differences of individual unfairness between groups based on the metric proposed above. We name it as Group Disparity of Individual fairness (GDIF).

[3]Without loss of generality, we take the widely studied node classification task as the downstream task in this paper.

It should be noted that *Group Equality of Individual Fairness* could be hard to achieve in application scenarios. Thus, to tackle Problem 1, we develop a quantitative metric for the disparity of individual fairness for different groups. Note U_p represents the *level of individual unfairness* of group $\mathcal{V}_p$. It is calculated by taking the average of pairwise constraining scalars (output distance divided by input distance or multiplied by similarity) for people in group $\mathcal{V}_p$ and people in the total population $\mathcal{V}$. If it is different between two groups $\mathcal{V}_p$ and $\mathcal{V}_q$ and let $U_p < U_q$, then people from $\mathcal{V}_p$ will on average have smaller output distances when they are compared to similar individuals than people from $\mathcal{V}_q$ have. The larger output distances for people from group $\mathcal{V}_q$ against their similar counterparts could potentially negatively affect them as illustrated in Fig. 2. Thus, the level of individual fairness should be equal for all groups such that there are no preferential outcomes for any group. To quantify such disparity between groups, we first define a metric below to measure the GDIF between two groups $\mathcal{V}_p$ and $\mathcal{V}_q$:

$$GDIF_{p,q} = max\left(\frac{U_p}{U_q}, \frac{U_q}{U_p}\right). \quad (5)$$

Here $GDIF_{p,q} \geq 1$. For two groups, $GDIF_{p,q} = 1$ means they are with equal level of individual fairness. Then, we can extend the group disparity of individual fairness for all groups as the sum of pairwise $GDIF_{p,q}$ for all combinations of pairwise groups,

$$GDIF = \sum_{p,q}^{1 \leq p < q \leq G} GDIF_{p,q}, \quad (6)$$

where G is the total number of groups. After defining quantifiable metrics for group disparity of individual fairness, we introduce our framework to tackle Problem 1 in next section.

4 PROPOSED FRAMEWORK

In this section, we propose a novel GNN framework—GUIDE to solve Problem 1. Specifically, we first give an overview of the proposed framework GUIDE. We then present the detailed operations of GUIDE. Finally, we introduce the optimization objectives of GUIDE to tackle Problem 1.

4.1 Framework Overview

The overview of our proposed framework is presented in Fig 3. First, informative node embeddings are initialized with a backbone GNN to benefit the node classification task. Then, for each node, since its neighbors on the similarity matrix could (1) have different similarities compared to this node and (2) have different group memberships, its neighbors could have different influences on the GDIF metric through influencing the output of this node. Thus, to better capture the neighbor-specific information, we propose to learn personalized weights for each node with respect to its neighbors on the similarity matrix. The attention mechanism can enable the framework to learn such personalized aggregation weights given the node features and pairwise similarity values. Next, the embeddings are aggregated for similar individual pairs with the learned attention weights to derive final outputs. The final outputs for nodes should satisfy group equality informed individual fairness as much as possible while maintaining node classification utility performance. In next subsection, we introduce more in details.

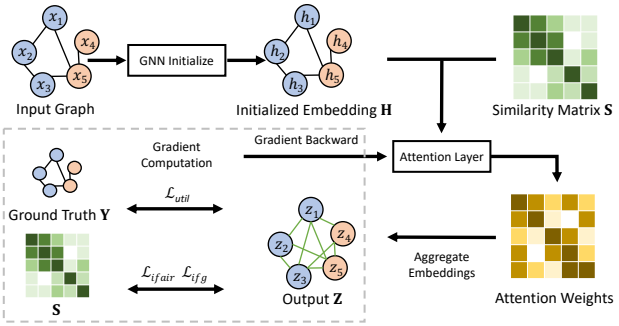


Figure 3: The overall framework of GUIDE.

4.2 Workflow of The Proposed Framework

The proposed framework generally needs to seek for a balance between three different objectives: (1) maximize model utility for the node classification task, (2) maximize overall individual fairness level, and (3) minimize GDIF. We discuss how we improve the our framework's performances on these three objectives in model workflow below through two main steps.

4.2.1 Achieving the Goal of Utility. First, to improve utility performance on the node classification task, we learn informative node embeddings with the graph adjacency matrix $\mathbf{A}$ and the node feature matrix $\mathbf{X}$ as inputs to a backbone GNN model. To ensure the embeddings encode critical information for the node classification task, we utilize the cross-entropy loss as the objective function during training. We extract the hidden layer representation $\mathbf{H} \in \mathbb{R}^{n \times h}$ from this GNN backbone as node embedding inputs for the aggregation operation in the next step.

4.2.2 Achieving the Goal of Fairness. Our fairness objectives are two-fold: maximizing overall individual fairness across the whole population and minimizing group disparity of individual fairness across different groups. To maximize the overall individual fairness, similar individuals should have similar outputs. To better capture the pairwise similarity information, we use a GNN to aggregate node embeddings based on the similarity matrix. In other words, we use the node similarity matrix $\mathbf{S}$ as a weighted input adjacency matrix and node embeddings $\mathbf{H}$ as the input feature matrix in a GNN to perform message passing. In this way, we can better capture the pairwise similarity values and encourage similar individuals to have similar outputs to improve overall individual fairness.

To minimize GDIF across different groups, as we mentioned in section 4.1, the neighbors of each node on the similarity matrix could (1) have different similarities compared to this node and (2) have different group memberships. Thus, these neighbors could have different influences on the GDIF metric. This effect indicates that in this GNN, we should learn personalized aggregation weights for each node and its neighbors to better capture the neighbor-specific information. However, the aggregation weights in many GNN frameworks are fixed weights [18, 30, 31] and cannot capture the different influences from neighbors. Such limitation could potentially affect our ability to achieve the optimal performances of the fairness objectives. Thus, to promote group equality informed individual fairness with fine-granularity, we adopt the attention mechanism in GAT [25] to learn personalized attention weights for each node and its neighbors. Specifically, we treat the similarity values as base values for learning aggregation weights so the similarity of output representations are more aligned with the input node similarities, which implicitly improves individual fairness. The pairwise aggregation weights are computed as follows:

$$\lambda_{i,j} = \frac{\exp(\phi(\mathbf{a}^T[\mathbf{W}\mathbf{h}_i || \mathbf{W}\mathbf{h}_j])\mathbf{S}[i,j])}{\Sigma_{j \in \mathcal{N}_i} \exp(\phi(\mathbf{a}^T[\mathbf{W}\mathbf{h}_i || \mathbf{W}\mathbf{h}_j])\mathbf{S}[i,j])}, \tag{7}$$

where $\lambda_{i,j}$ is the attention from node i to node j, $\mathbf{W} \in \mathbb{R}^{c \times h}$ is the weight matrix, $\mathbf{a} \in \mathbb{R}^{2c}$ is the attention weight vector, $\mathbf{h}_i \in \mathbb{R}^h$ is the input embedding of node i, $[\cdot||\cdot]$ is concatenation of two vectors, $\mathcal{N}_i$ is the neighborhood set of node i, and ϕ is an activation function. And the operation for node aggregation is

$$\mathbf{z}_i = \sigma(\Sigma_{j \in \mathcal{N}_i} \lambda_{i,j} \mathbf{W}\mathbf{h}_j), \tag{8}$$

where $\mathbf{z}_i \in \mathbb{R}^c$ is the output embedding of node i and σ is an activation function. The attention weights are learned by optimizing the total objective function which we introduce in next subsection.

4.3 Objective Function Formulation

In this subsection, we summarize the loss functions for each optimization objective and present the total objective function for optimizing our framework. First, to maintain the utility of the GNN model (i.e., to achieve accurate node classification accuracy), we adopt the previously mentioned cross entropy loss as the first objective function term, which is widely adopted in node classification tasks. Specifically, it is formulated as:

$$\mathcal{L}_{\text{util}} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{K} \mathbf{Y}_{ij} \log \hat{\mathbf{Y}}_{ij}, \tag{9}$$

where $\mathbf{Y}_{ij}$ indicates the true label (in K class) of i^{th} node and $\hat{\mathbf{Y}}_{ij}$ is the prediction of i^{th} node. Second, we utilize $\mathcal{L}_{\text{ifair}}$ for individual fairness optimization. As we introduced in section 2.2, this loss function is the sum of pairwise output distances multiplied by pairwise similarities for all individuals. Minimizing it will improve the overall individual fairness in the framework. Third, we define a loss function $\mathcal{L}_{\text{ifg}}$ for the GDIF objective in order to promote group equality of individual fairness. Specifically, we aim to equalize levels of individual unfairness for all groups. To do so, we define a differentiable loss with pairwise U_p and U_q for minimizing GDIF:

$$\mathcal{L}_{\text{ifg}} = \sum_{p,q}^{1 \le p < q \le G} \left(\frac{U_p}{U_q} - 1\right)^2 + \left(\frac{U_q}{U_p} - 1\right)^2, \tag{10}$$

where U_p and U_q are individual unfairness for group $\mathcal{V}_p$ and group $\mathcal{V}_q$ respectively. Each group's individual unfairness is computed with the aggregated embeddings $\mathbf{Z}$ according to Eq. (4). Note that this loss function is symmetrical to any given two groups such that it is the same regardless of the order.

In summary, there are three objectives in total for the optimization of GUIDE: utility objective from Eq. (9), overall individual fairness from Eq. (2) and group equality informed individual fairness from Eq. (10). The total loss function is a weighted sum of losses of each objective, weighted by hyperparameters α and β:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{util}} + \alpha \mathcal{L}_{\text{ifair}} + \beta \mathcal{L}_{\text{ifg}}. \tag{11}$$

Table 2: Statistics of the used datasets.

Dataset	Credit	Income	Pokec-n
# of nodes	30, 000	14, 821	66, 569
# of node attributes	13	14	266
# of edges in $\mathbf{A}$	304, 754	100, 483	1, 100, 663
# of edges in $\mathbf{S}$	1, 687, 444	1, 997, 641	32, 837, 463
Group ratio	11.2	3.16	21.0
Group avg degree ratio	12.6	2.8	58.8
Sensitive Attribute	age	race	age

To summarize the GUIDE framework, a backbone GNN utilizes the adjacency matrix $\mathbf{A}$ and feature matrix $\mathbf{X}$ to initialize node embeddings $\mathbf{H}$. Then, the embeddings are passed into a GNN layer operating on the similarity matrix $\mathbf{S}$ to learn individually personalized attention weights for node aggregation given the total objective function. The total objective function encompasses three loss functions corresponding to the three objectives in Problem 1: (1) utility maximization, (2) individual fairness maximization over the whole population, and (3) group disparity of individual fairness minimization for all groups in the population.

5 EXPERIMENTS

In this section, we conduct extensive experiments on real-world datasets to validate the effectiveness of GUIDE. Specifically, we aim to answer the following research questions:

- **RQ1:** How well can GUIDE balance utility, individual fairness, and group equality of individual fairness objectives compared to baselines?
- **RQ2:** How well does the attention mechanism help us achieve optimization of the fairness objectives: (1) minimizing overall individual unfairness and (2) minimizing GDIF?

5.1 Datasets

We employ three different real-world network datasets from different application domains: Credit, Income, and Pokec-n. We introduce the details about the three datasets as below.

Credit: The Credit graph dataset is constructed on 30,000 individuals. They are connected based on features such as spending and payment habits [32]. Our goal is to predict if an individual will default on credit card payment and the sensitive attribute is age.

Income: The Income graph dataset is constructed on 14,821 individuals who are sampled from the *Adult Data Set* [8]. The sensitive attribute used here is race and individuals are connected based on their features. The prediction task is to determine if a person's income is over $50K a year or not.

Pokec-n: Pokec-n is a sampled dataset from Slovakia's most popular social network [23]. The dataset contains 66,569 individuals and they are connected by friend relationships. Here, we treat age as the sensitive attribute and our goal is to predict the working field of users in this social network.

5.2 Experimental Settings

Metrics. We use the row-wise cosine similarity of the adjacency matrix $\mathbf{A}$ to instantiate the similarity matrix $\mathbf{S}$, i.e., the (i, j)-th entry in $\mathbf{S}$ represents the cosine similarity between the i-th row and the j-th row of the adjacency matrix $\mathbf{A}$. This is aligned with the similarity metric on individual fairness in existing works [16]. The utility performance (i.e., node classification accuracy) of GUIDE on node classification task is evaluated with the widely adopted metric AUCROC (AUC). Additionally, we also evaluate the performance of GUIDE on individual fairness metrics: overall individual (un)fairness (IF) [16, 20] and the proposed group disparity of individual fairness (GDIF).

GNN backbones. All baselines and GUIDE can use arbitrary GNN backbones. In order to present extensive comparison, we conduct experiments on three different GNN backbones: GCN [18], GIN [30], and JumpingKnowledge [31]. All experiments are performed five times and the average results are reported along with the standard deviation. The experiment results are shown in Table 3.

Baselines. To validate the effectiveness of the proposed framework, we conduct experiments based on the following baseline models:

- **FairGNN** [4] uses adversarial learning such that GNNs make fair node classifications that satisfy group fairness. We directly apply it to various GNN backbones and analyze if it optimizes our defined GDIF.
- **NIFTY** [1] optimizes counterfactual fairness and stability by perturbing attributes, using Lipschitz constant to normalize layer weights and training with contrastive learning. We directly adopt it for various GNN backbones.
- **PFR** [20] learns fair node embeddings as preprocessing step satisfying individual fairness in downstream tasks. The learned embeddings are used as inputs for GNN backbones.
- **InFoRM** [16] formulates individual fairness loss in a graph based on Lipschitz condition. We add the proposed individual fairness loss to GNN backbone training.

5.3 Effectiveness of GUIDE

We aim to answer **RQ1** in ths subsection. Experiment results are presented in Table 3. Our goal is to (1) minimize overall individual unfairness such that similar individuals can have similar outputs, (2) minimize GDIF such that there is less disparity of individual fairness for different groups, and (3) maintain good utility performance for the node classifcaition task.

- First, from the perspective of fairness promotion, GUIDE achieves lowest individual unfairness and lowest GDIF compared to all baseline models, indicating its superior performance in achieving both fairness objectives across different datasets and GNN backbones.
- Second, we observe GUIDE maintains relatively comparable utility performances compared to vanilla and baseline models. This illustrates that GUIDE can effectively perform the given supervised task of node classification.
- Third, from the perspective of balancing utility and fairness objectives, GUIDE achieves the best performances in both minimizing overall individual unfairness and minimizing GDIF while scoring comparably in utility performance with vanilla and baseline models. Thus, we claim it achieves a good trade-off in balancing the utility and fairness objectives.

5.4 Effect of Personalized Weights

To answer **RQ2**, we analyze the advantage of the adopted personalized aggregation weights over fixed aggregation weights towards

Table 3: Experiment results on Credit, Income and Pokec-n datasets. Model indicates the debiasing algorithm and Vanilla represents no debiasing is performed. ↑ denotes the larger, the better; ↓ means the opposite. Best performances are in bold. Individual (un)fairness numbers are reported in thousands. All entries are averages and standard deviations.

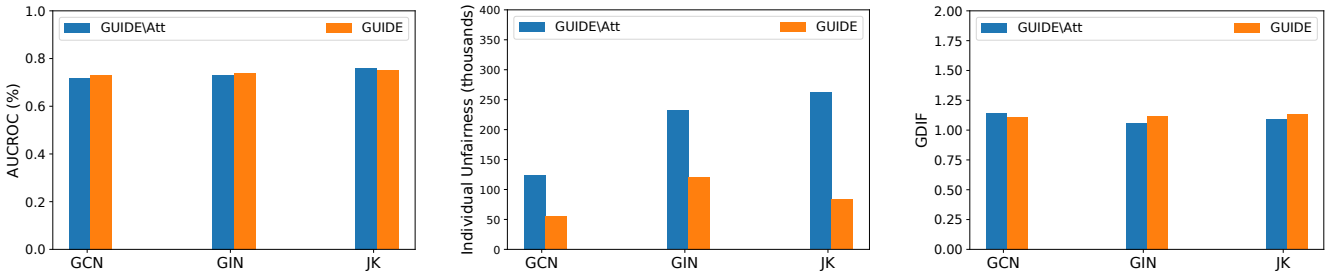
Credit									
Model	**AUC(↑)**	**IF(↓)**	**GDIF(↓)**	**AUC(↑)**	**IF(↓)**	**GDIF(↓)**	**AUC(↑)**	**IF(↓)**	**GDIF(↓)**
	GCN			GIN			Jumping Knowledge		
Vanilla	0.68±0.04	39.02±3.78	1.32±0.07	**0.71±0.00**	120.02±15.42	1.75±0.21	0.64±0.11	31.06±13.90	1.32±0.06
FairGNN	0.68±0.01	23.33±12.59	1.33±0.10	0.68±0.02	77.32±48.47	2.18±0.19	0.66±0.02	2.61±1.92	1.52±0.42
NIFTY	**0.69±0.00**	30.80±1.39	1.24±0.02	0.70±0.01	56.43±37.85	1.63±0.27	**0.69±0.00**	26.44±2.39	1.24±0.03
PFR	0.64±0.13	36.58±6.91	1.41±0.08	0.71±0.01	162.58±103.87	2.40±1.23	0.67±0.05	36.30±18.22	1.35±0.03
InFoRM	0.68±0.00	2.41±0.00	1.46±0.00	0.69±0.02	2.94±0.28	1.76±0.17	0.67±0.05	5.66±5.31	1.47±0.16
GUIDE	0.68±0.00	**1.93±0.11**	**1.00±0.00**	0.68±0.00	**2.43±0.02**	**1.00±0.00**	0.68±0.00	**2.34±0.11**	**1.00±0.00**
Income									
	GCN			GIN			Jumping Knowledge		
Vanilla	0.77±0.00	369.11±0.03	1.29±0.00	**0.81±0.01**	2815.59±1047.33	1.87±0.48	**0.80±0.00**	488.73±166.83	1.18±0.16
FairGNN	0.76±0.00	249.73±87.53	1.17±0.04	0.79±0.00	1367.93±875.64	3.30±1.18	0.77±0.00	219.30±42.92	1.30±0.12
NIFTY	0.73±0.00	42.14±5.83	1.38±0.04	0.79±0.01	608.98±314.83	1.17±0.26	0.73±0.02	48.25±10.48	1.39±0.09
PFR	0.75±0.00	245.97±0.58	1.32±0.00	0.79±0.00	2202.64±445.24	2.36±1.17	0.73±0.13	327.57±155.49	1.12±0.23
InFoRM	**0.78±0.00**	195.61±0.01	1.36±0.00	0.80±0.01	308.45±13.92	1.62±0.30	0.79±0.00	192.58±12.87	1.35±0.11
GUIDE	0.73±0.01	**33.19±10.17**	**1.00±0.00**	0.74±0.02	**83.88±20.29**	**1.00±0.00**	0.74±0.01	**42.49±21.93**	**1.00±0.00**
Pokec-n									
	GCN			GIN			Jumping Knowledge		
Vanilla	**0.77±0.00**	951.72±37.28	6.90±0.12	**0.76±0.01**	4496.47±1535.62	8.35±1.24	**0.79±0.00**	1631.27±93.94	8.47±0.45
FairGNN	0.69±0.03	363.73±78.38	6.21±1.28	0.69±0.01	416.28±402.83	4.84±2.94	0.70±0.00	807.97±281.26	11.68±2.89
NIFTY	0.74±0.00	85.25±10.55	5.06±0.29	**0.76±0.01**	2777.36±346.29	9.28±0.28	0.73±0.01	477.31±165.68	8.20±1.33
PFR	0.53±0.00	98.25±9.44	15.84±0.03	0.60±0.01	628.27±85.89	6.20±0.79	0.68±0.00	729.77±74.62	15.66±5.47
InFoRM	**0.77±0.00**	230.45±6.13	6.62±0.10	0.75±0.01	271.65±30.63	6.83±1.34	0.78±0.01	315.27±25.21	6.80±0.54
GUIDE	0.73±0.02	**55.05±30.87**	**1.11±0.03**	0.74±0.01	**120.65±17.33**	**1.12±0.03**	0.75±0.02	**83.09±18.70**	**1.13±0.02**

achieving the fairness objectives. Intuitively, personalized aggregation weights learned with attention in GUIDE should be better at capturing neighbor-specific information for each node in the similarity matrix. This effect should allow the model to perform targeted optimization on node pairs so it should have better performances in the fairness objectives. To analyze the effect of personalized aggregation weights, we use aggregation mechanism with fixed aggregation weights in place of GUIDE's attention aggregation mechanism. Specifically, we use the aggregation mechanism of three popular GNN frameworks: GCN [18], GIN [30] and JumpingKnowledge [31] to evaluate the performance differences against GUIDE. They are trained with the same total loss function as GUIDE. We name this variant GUIDE\Att here for reference. We explore the utility performance (measured by AUCROC), and fairness performances in overall individual unfairness, and group disparity of individual fairness. Specifically, GUIDE should have better fairness objectives performances with comparable utility performance. We present results on Pokec-n with GCN, GIN, and JumpingKnowledge GNN backbones in Fig 4. Similar observations can also be found on other datasets. From Fig (4a) and Fig (4c), we observe that both GUIDE and GUIDE\Att achieve similar levels of utility and GDIF performances. However, Fig (4b) shows that GUIDE has lower overall individual unfairness. This may indicate that personalized aggregation weights from GUIDE can indeed provide better tradeoff of the optimized objectives: achieving lower overall individual unfairness while minimizing GDIF and maintaining utility performance than the variant using fixed aggregation weights. We hypothesize that the better tradeoff derives from targeted optimization of node pairs. Specifically, the aggregation weights for each node are personalized with respect to the optimization objectives so these weights can be optimized on a case-by-case basis for each node pair. In next subsection, we analyze the personalized aggregation weights further to see if there is any relationship between them and each node pairs' influences on the objectives such as GDIF.

5.5 Aggregation Weights Analysis

In this subsection, we verify if the learned pairwise attention weights from GUIDE are personalized adaptively for each node pair to benefit the corresponding optimized objectives.

Intuitively, model with personalized aggregation weights instead of fixed aggregation weights can better capture the neighbor specific influences on optimized objectives such as GDIF. Here we aim to analyze if GUIDE induce more adjustments on the attention weights of node pairs that have large influences on GDIF than other node pairs when GDIF is added to the total loss. To verify if such targeted optimization of GDIF takes effect, we conduct correlation test between node pairs' GDIF influence and their corresponding attention weights changes $\Delta\lambda$ from optimizing the GDIF objective. We first approximate the influence on GDIF from different pairs of nodes by leave-one-out calculations. We obtain model outputs from GUIDE trained with only utility and individual fairness objectives (i.e. trained with $\beta = 0$) and calculate a benchmark GDIF denoted as $GDIF_{\text{benchmark}}$. Next, for a specific node pair (i, j), we leave them out of the GDIF calculation, i.e., $\left\|z_i - z_j\right\|_2^2 \mathbf{S}[i, j]$ and



(a) AUCROC (b) Individual (Un)fairness (c) Group Disparity of Individual Fairness

Figure 4: Performance results of GUIDE and its variant GUIDE\Att. (a) Node classification performance comparison between GUIDE and GUIDE\Att; (b) Individual fairness promotion comparison between GUIDE and GUIDE\Att; (c) GDIF optimization comparison between GUIDE and GUIDE\Att.

Table 4: Two-tailed correlation test of node pairs' GDIF influence and attention weight changes.

Dataset	Credit	Income
Correlation of $C_{i,j}$ and $\Delta\lambda_{i,j}$	0.036	0.044
T-statistics	46.638	62.904
Number of samples	1, 687, 444	1, 997, 641

$\left\|z_j - z_i\right\|_2^2 \mathrm{S}[j,i]$ are removed from the calculation of individual unfairness of corresponding groups $(U_1 ... U_G)$. We denote this GDIF as $GDIF_{-(i,j)}$. Finally, we define the influence of node pair (i, j) as:

$$C_{i,j} = C_{j,i} = |GDIF_{-(i,j)} - GDIF_{\text{benchmark}}|. \tag{12}$$

We then calculate the absolute value of change in attention weights between GUIDE trained with and without GDIF objective ($\beta = 0$ and $\beta \neq 0$) to represent how attention weights changed when GDIF objective is optimized.

$$\Delta\lambda_{i,j} = |\lambda_{i,j}^{\beta\neq 0} - \lambda_{i,j}^{\beta=0}|. \tag{13}$$

We propose the null hypothesis that there is no linear relationship between changes in attention weights and node pairs' influences on GDIF when GDIF is optimized. Hence, the null hypothesis is correlation ρ between $C_{i,j}$ and $\Delta\lambda_{i,j}$ is 0 ($H_0 : \rho = 0$) and the alternative hypothesis that ρ is not 0 ($H_a : \rho \neq 0$). We also set null hypothesis rejection threshold of p-value as 0.01. The two-tailed correlation test results for Credit and Income datasets are listed in Table 4. We observe positive correlations between influence on GDIF and magnitude of change in attention weight for node pairs. We also observe very significant t-stats. The p-values are significantly below 0.01. Hence, it is indicative that we can reject the null hypothesis and claim GUIDE can yield personalized attention weights for node pairs with respect to their influences on GDIF.

6 RELATED WORK

Algorithmic fairness. Researchers have formulated a variety of algorithmic fairness notions and they can be broadly categorized as group fairness, counterfactual fairness, and individual fairness. *Group fairness* is defined as enforcing equal outcome statistics such as true positives across different groups. Zafar et al. [33] propose *demographic parity* which requires equal likelihood of positive outcome regardless of group membership. Hardt et al. [13] present *equal opportunity* which argues people from different groups should have equal true positive rates. Both works formalize the fairness notions as optimization constraints in model utility maximization. *Counterfactual fairness* promotes fixed model outcomes for individuals regardless of what their sensitive attributes are in reality or counterfactual scenarios. Agarwal et al. [1] perturb node features and flip node sensitive attributes to arrive three different model outputs and minimize the triplet similarity distance to achieve counterfactul fairness. For *Individual Fairness*, Dwork et al. [9] propose individual fairness which requires *treating similar individuals similarly*. They formulate it as an optimization problem involving pairwise individual similarity and Lipschitz condition. Lahoti et al. [20] treat individual fairness as a low-rank representation learning problem by minimizing output distances multiplied by individual similarity. García-Soriano et al. [10] minimize the amount of individual unfairness after enforcing group fairness by optimizing a max-min ranking problem. Majority of these models rely on Lipschitz condition and we have found this formulation could result in different levels of individual fairness for different groups which leads to discrimination against certain demographic subgroups. To our best knowledge, we are the first to investigate this issue in individual fairness and provide a viable solution.

Fairness in graph mining. As graph mining models are increasingly adopted for many learning tasks, numerous solutions have been proposed to mitigate potential unfairness in graph mining algorithms [7]. For group fairness, Rahman et al. [22] propose the notion of equality of representation which extends statistical parity to the node2vec model. Bose et al. [3] propose a compositional adversarial method to remove the influence of sensitive attributes in learned embeddings. Dai et al. [4] develop a similar adversarial framework but debiasing is performed in end-to-end GNN predictions. For individual fairness, Kang et al. [16] optimize individual fairness by deriving an individual fairness loss on graph datasets and reduce it before, during, and after training of the graph mining model. Dong et al. [5] treat optimization of individual fairness in GNNs as a ranking problem which bypasses the limitation of Lipschitz condition. Our approach differs from these cited works in that we not only optimize overall individual fairness but also explicitly equalize the levels of fairness across groups such that sensitive attributes such as race or age do not affect the level of individual fairness one experiences when compared to similar individuals.

7 CONCLUSION

Graph Neural Networks have shown superior performances in a variety of tasks and are increasingly adopted in high-stake decision-making systems. However, there has been heightened concerns that

GNNs could generate unfair decisions for underprivileged groups or individuals without fairness constraints. Out of various proposed algorithmic fairness notions on GNNs, individual fairness has finer granularity on the individual level and promotes *treating similar individuals similarly*. However, in our analysis of several works on individual fairness, we have found that their formulation from Lipschitz condition could lead to different levels of individual fairness for different groups, thus creating discrimination on the group level. We tackle this problem by developing a novel GNN framework: GUIDE which incorporates an attention based GNN that learns individually personalized attention weights for achieving group equality informed individual fairness. We conduct extensive experiments on real-world datasets to demonstrate the effectiveness of our proposed framework and the results show GUIDE substantially remove group disparity of individual fairness, achieve overall individual fairness, and maintain utility performance.

8 ACKNOWLEDGEMENTS

This material is supported by the Cisco Faculty Research Award. We also thank the anonymous reviewers for their feedback.

REFERENCES

[1] Chirag Agarwal, Himabindu Lakkaraju, and Marinka Zitnik. 2021. Towards a Unified Framework for Fair and Stable Graph Representation Learning. *CoRR* abs/2102.13186 (2021). arXiv:2102.13186 https://arxiv.org/abs/2102.13186

[2] Punam Bedi and Chhavi Sharma. 2016. Community detection in social networks. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 6, 3 (2016), 115–135.

[3] Avishek Joey Bose and William L. Hamilton. 2019. Compositional Fairness Constraints for Graph Embeddings. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (Proceedings of Machine Learning Research, Vol. 97)*. PMLR, 715–724.

[4] Enyan Dai and Suhang Wang. 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* (Virtual Event, Israel) *(WSDM '21)*. Association for Computing Machinery, New York, NY, USA, 680–688. https://doi.org/10.1145/3437963.3441752

[5] Yushun Dong, Jian Kang, Hanghang Tong, and Jundong Li. 2021. Individual Fairness for Graph Neural Networks: A Ranking based Approach. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 300–310.

[6] Yushun Dong, Ninghao Liu, Brian Jalaian, and Jundong Li. 2022. Edits: Modeling and mitigating data bias for graph neural networks. In *Proceedings of the ACM Web Conference 2022*. 1259–1269.

[7] Yushun Dong, Jing Ma, Chen Chen, and Jundong Li. 2022. Fairness in Graph Mining: A Survey. *arXiv preprint arXiv:2204.09888* (2022).

[8] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[9] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*. ACM, 214–226.

[10] David García-Soriano and Francesco Bonchi. 2021. Maxmin-Fair Ranking: Individual Fairness under Group-Fairness Constraints. *CoRR* abs/2106.08652 (2021). https://arxiv.org/abs/2106.08652

[11] Lin Gong, Lu Lin, Weihao Song, and Hongning Wang. 2020. JNET: Learning User Representations via Joint Network Embedding and Topic Embedding. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*. ACM, 205–213.

[12] William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. 1024–1034.

[13] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. 3315–3323.

[14] Mohammad Al Hasan and Mohammed J. Zaki. 2011. A Survey of Link Prediction in Social Networks. In *Social Network Data Analytics*, Charu C. Aggarwal (Ed.). Springer, 243–275.

[15] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2020. A Survey on Knowledge Graphs: Representation, Acquisition and Applications. *CoRR* abs/2002.00388 (2020). https://arxiv.org/abs/2002.00388

[16] Jian Kang, Jingrui He, Ross Maciejewski, and Hanghang Tong. 2020. InFoRM: Individual Fairness on Graph Mining. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*. ACM, 379–389.

[17] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[18] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

[19] Matthäus Kleindessner, Samira Samadi, Pranjal Awasthi, and Jamie Morgenstern. 2019. Guarantees for Spectral Clustering with Fairness Constraints. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (Proceedings of Machine Learning Research, Vol. 97)*. PMLR, 3458–3467.

[20] Preethi Lahoti, Krishna P. Gummadi, and Gerhard Weikum. 2019. Operationalizing Individual Fairness with Pairwise Fair Representations. *Proc. VLDB Endow.* 13, 4 (2019), 506–518.

[21] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6, Article 115 (July 2021), 35 pages. https://doi.org/10.1145/3457607

[22] Tahleen A. Rahman, Bartlomiej Surma, Michael Backes, and Yang Zhang. 2019. Fairwalk: Towards Fair Graph Embedding. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. ijcai.org, 3289–3295.

[23] L. Takac and Michal Zábovský. 2012. Data analysis in public social networks. *International Scientific Conference and International Workshop Present Day Trends of Innovations* (01 2012), 1–6.

[24] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

[25] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=rJXMpikCZ

[26] Daixin Wang, Yuan Qi, Jianbin Lin, Peng Cui, Quanhui Jia, Zhen Wang, Yanming Fang, Quan Yu, Jun Zhou, and Shuang Yang. 2019. A Semi-Supervised Graph Attentive Network for Financial Fraud Detection. In *2019 IEEE International Conference on Data Mining, ICDM 2019, Beijing, China, November 8-11, 2019*. IEEE, 598–607.

[27] Yiwei Wang, Wei Wang, Yuxuan Liang, Yujun Cai, and Bryan Hooi. 2021. CurGraph: Curriculum Learning for Graph Classification. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*. ACM / IW3C2, 1238–1248. https://doi.org/10.1145/3442381.3450025

[28] Qitian Wu, Hengrui Zhang, Xiaofeng Gao, Peng He, Paul Weng, Han Gao, and Guihai Chen. 2019. Dual Graph Attention Networks for Deep Latent Representation of Multifaceted Social Effects in Recommender Systems. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*. ACM, 2091–2102. https://doi.org/10.1145/3308558.3313442

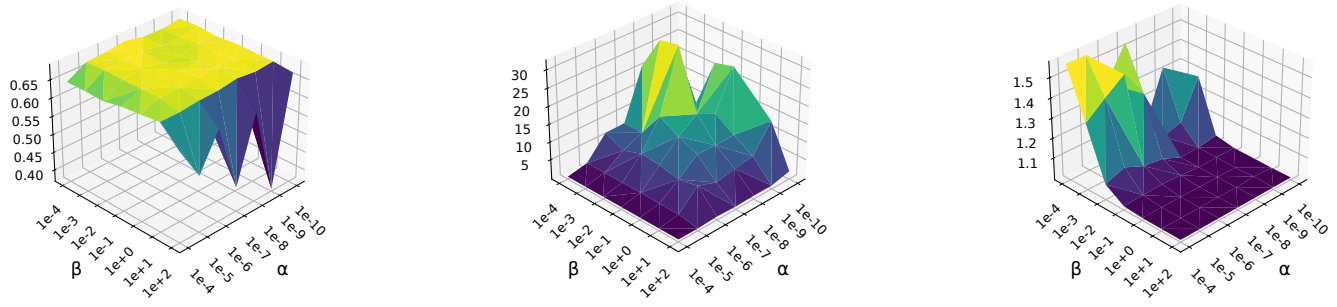
[29] Shiwen Wu, Wentao Zhang, Fei Sun, and Bin Cui. 2020. Graph Neural Networks in Recommender Systems: A Survey. *CoRR* abs/2011.02260 (2020). https://arxiv.org/abs/2011.02260

[30] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How Powerful are Graph Neural Networks?. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

[31] Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. 2018. Representation Learning on Graphs with Jumping Knowledge Networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*. PMLR, 5449–5458.

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

[33] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA (Proceedings of Machine Learning Research, Vol. 54)*. PMLR, 962–970.



(a) AUCROC (b) Individual (Un)fairness (c) Group Disparity of Individual Fairness

Figure 5: Performance results of GUIDE for Credit dataset on GCN with varying hyperparameters α for the overall individual unfairness objective and β for the GDIF objective.

A APPENDIX

A.1 Reproducibility

In this section, we present the details of the implementation of GUIDE and other baselines presented in Section 5.2

Experiment settings on datasets. For all three datasets, we randomly shuffle the nodes and take 25% of the labeled nodes as validation set and 25% of the labeled nodes as test set. For size of the training set, we use 6,000 labeled nodes (25%) for Credit, 3,000 labeled nodes (20%) for Income, and 4,398 labeled (6%) for Pokec-n. For Pokec-n, the edges are provided as the friendship linkages whereas for the other two dataset, there are no given edges. So we construct edges based on feature similarity. Specifically for a given pair of nodes, if the Euclidean distances of their features meet some threshold, we consider them connected.

Training settings. All models including baselines are trained with Adam optimizer [17] with learning rate as 1e-3 and weight decay as 1e-5. All models are trained with hidden dimension as 16 and number of epochs is 3,000. The best model for each framework/backbone combination is saved based on validation performance and is applied to the test set for results shown in Table 3.

Implementation details and hyperparameters. The proposed framework GUIDE is implemented in PyTorch and the code is available here: https://github.com/mikesong724/GUIDE. Mode details on model implementations and hyperparameters are listed below:

- **GUIDE.** We use $\alpha = 5e-6, \beta = 1$ for Credit, $\alpha = 1e-7, \beta = 0.25$ for Income, and $\alpha = 2.5e-7, \beta = 0.05$ for Pokec-n.
- **FairGNN.** We use $\alpha = 4, \beta = 1000$ for Credit, $\alpha = 4, \beta = 10$ for Income, and $\alpha = 4, \beta = 100$ for Pokec-n.
- **NIFTY.** We use $\lambda = 0.5$ across all datasets.
- **PFR.** We use the debiased embeddings from PFR as inputs to GNN backbones. PFR utilizes two relationship matrices: W_X for feature similarities derived from k-nearest-neighbor and W_F from human judgement for pairwise similarities. In order to compare it with other baselines, we use **A** and **S** for them respectively. For hyperparameter we use $\lambda = 0.5$ for Credit and Income, and $\lambda = 0.25$ for Pokec-n.
- **InFoRM.** InFoRM has three debiasing steps: the preprocessing, inprocessing and postprocessing. We adopt the main individual fairness loss term from their paper and add it to vanilla GNN backbone with hyperparameter α to vary its weight. We uses $\alpha = 5e-6$ for Credit, $\alpha = 1e-7$ for Income and Pokec-n.

Packages required for implementations. The main packages and their versions are provided below for our implementations.

- Python==3.7.11
- PyTorch==1.10.0
- CUDAtoolkit==11.1.1
- torch-scatter==2.0.9
- torch-sparse==0.6.13
- torch-geometric==2.0.1
- NetworkX==2.6.3
- NumPy==1.21.6
- SciPy==1.7.3
- AIF360 == 0.3.0

A.2 Hyperparameter Sensitivity

For the proposed GUIDE framework, there are two main hyperparameters α and β to optimize the overall individual unfairness objective and group disparity of individual fairness (GDIF) objective respectively. To study their effects on GUIDE's performance on each of the three objectives, we conduct hyperparameter sensitivity experiments in this subsection and vary α among {1e-10, 1e-9, 1e-8, 1e-7, 1e-6, 1e-5, 1e-4} and β among {1e-4, 1e-3, 1e-2, 1e-1, 1, 1e+1, 1e+2}. The results are presented in Figure 5.

- From Figure 5a, we observe that for $\alpha < 1e-5$ and $\beta < 1e+1$ the node classification task performance is not deteriorated. When α and β become larger that these two thresholds, the classification performance decreases. Interestingly, β seems to have a more abrupt effect than α.
- From Figure 5b, we observe that increasing α and β both decrease the overall individual unfairness while increasing α has a stronger effect in reducing individual unfairness as it specifically minimizes this objective.
- From Figure 5c, we clearly observe that when β is small, GDIF stays elevated no matter how we vary α. In fact, when $\beta = 1e-4$, the GDIF increases as α increases. It demonstrates that existing method for individual fairness optimization disregards GDIF objective. As we increase β, GDIF is successfully reduced and group equality of individual fairness ($GDIF = 1$) for the two groups is achieved.