FairGMLP: Learning Individually Fair Node Representations via Knowledge Distillation

Seong Jin Ahn, Myoung Ho Kim

KAIST, School of Computing

sja1015@kaist.ac.kr, mhkim@kaist.ac.kr

Abstract

Nowadays, it has become important to mitigate unfairness in deep learning models. Individual fairness aims to ensure that similar individuals should be treated similarly. Graph neural networks (GNNs) are widely adopted in graph representation learning. GNNs aggregate node representations with message-passing mechanisms to learn neighborhood information. However, the individual fairness of GNNs can be limited through this process. On the other hand, MLPs can achieve a higher degree of individual fairness despite having a lower accuracy than GNNs in general. In this paper, we introduce a knowledge distillation framework that captures both accuracy and fairness. The proposed framework first learns the structural information with a GNN teacher, then distills the knowledge to a MLP student. Through extensive experiments, we demonstrate the effectiveness of the proposed method.

1. Introduction

Graph representation learning has attracted a wide range of interest due to its possibility of analyzing social networks, knowledge graphs, and biological networks [1,2]. It aims to learn low-dimensional vectors that preserve attributes and local structures of nodes. With the advent of deep learning, graph neural networks (GNNs) have surpassed other state-of-the-art graph representation learning methods in a variety of graph analysis tasks. They take advantage of message-passing architectures that aggregate neighborhood information to learn node representations. Recently, there has been growing concern in society that there is a lack of consideration for the fairness of GNNs.

One of the main concerns of GNNs is the tendency to lose individual fairness that is the requirement for similar individuals to be treated alike. They may fail to guarantee consistent predictions for individuals with similar characteristics because of their reliance on the message-passing. Figure 1 shows an example that describes the weakness of GNNs concerning individual fairness. Suppose there are two nodes v_1 and v_2 that have the same attribute values ($x_{v_1} = x_{v_2}$). As GNN aggregates neighborhood information, node representations z_{v_1} and z_{v_2} become different. There are some attempts to handle the limitation by introducing objective functions concerning individual fairness. Although they mitigate the problem to some extent, they cannot completely overcome the problem because of the dependence on message-passing mechanisms.

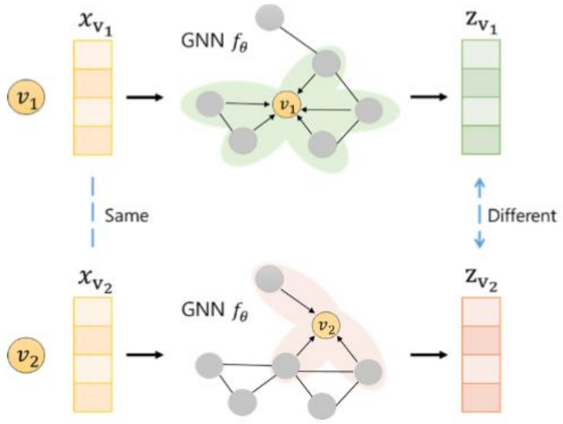


Figure 1. An example that illustrates the inherent weakness of GNNs in preserving individual fairness.

To address the issue, we introduce a framework called Fair Graph Multi-Layer-Perceptron (FairGMLP). Based on the MLP architecture, the framework encodes a representation of a node with its attributes. In this way, it achieves a higher level of individual fairness compared to GNNs. We enhance the power of the MLP in graph analysis by injecting the knowledge of a GNN teacher. Extensive experiments show that FairGMLP achieves the best utility and fairness results in almost

[1] This work was supported by the ITRC support program (IITP-2020-2020-0-01795) supervised by the IITP.

all combinations of datasets with various backbone models.

2. Background : Individual Fairness for GNNs

Individual fairness is a principle that similar individuals are to be treated similarly. With the advance of GNNs, there have been many efforts to enhance the individual fairness of GNNs. Some studies address the limitation by introducing objective functions with respect to individual fairness in GNNs. InFoRM designs a node pair distance-based fairness based on the Lipschitz condition. REDRESS enhances individual fairness through a ranking-based strategy, wherein fairness scores were assigned to nodes in the graph. However, they still rely heavily on message-passing mechanisms which limit the individual fairness of models.

3. Approach

In this section, we present our knowledge distillation framework to overcome the limited individual fairness of GNNs. Our work consists of three components: (1) GNN teacher, (2) MLP student, and (3) Fairness Promoting Regularization.

3.1 GNN Teacher

An initial step of our work is to obtain the neighborhood information from a pre-trained teacher GNN as eq.(1). The teacher model employed in the framework can be an arbitrary GNN model, including GCN, GAT, and SGC.

$$z_v^{(Tea)} = \mathrm{GNN}(\mathrm{x}_\mathrm{v}, \{x_u | u \in N(v)\}) \quad (1)$$

3.2 MLP Student

Next, the framework encodes node attributes with a student MLP to generate node representations as eq.(2). As node attributes are the only matter, node representations of similar individuals are likely to be treated similarly. Consequently, the representations can achieve a higher level of individual fairness compared to those of the teacher.

$$z_v^{(Stu)} = \mathrm{MLP}(\mathrm{x}_\mathrm{v}, \{x_u | u \in N(v)\}) \quad (2)$$

where x_v is a node attribute vector of a node v.

Despite the potential of MLPs for improving fairness, they have difficulty in learning graph structures. With the success of knowledge distillation, our framework injects the local structure knowledge of the teacher into the fairer student as follows:

$$\mathcal{L}_{KD} = \sum_{v \in V} \mathrm{D}_{KL}(z_v || \tilde{z}_v), \qquad (3)$$

where $\mathrm{D}_{KL}(p||q)$ is a Kullback-Leibler divergence between p and q.

3.3 Fariness Promoting Regularization

We further encourage the framework to produce fairer results by incorporating a fairness loss. Our work tries to integrate the ranking-based fairness loss designed in REDRESS. The fairness minimize the difference between the feature similarity matrix S and the outcome similarity matrix $\hat{S}$ as:

$$\hat{P}_{j,m}(i) = \frac{1}{1 + e^{-\alpha(\hat{s}_{i,j} - \hat{s}_{i,m})}}, \qquad (4)$$

$$P_{j,m}(i) = \begin{cases} 1 & \text{if } s_{i,j} > s_{i,m} \\ 0.5 & \text{if } s_{i,j} = s_{i,m} \\ 0 & \text{if } s_{i,j} < s_{i,m} \end{cases} \qquad (5)$$

$$\mathcal{L}_{j,m}(i) = -P_{j,m} \log \hat{P}_{j,m} - (1 - P_{j,m})(1 - \log \hat{P}_{j,m}), \qquad (6)$$

$$\mathcal{L}_{fair}(i) = \sum_{j,m} \mathcal{L}_{j,m}(i) |\Delta z_{@k}|_{j,m}, \qquad (7)$$

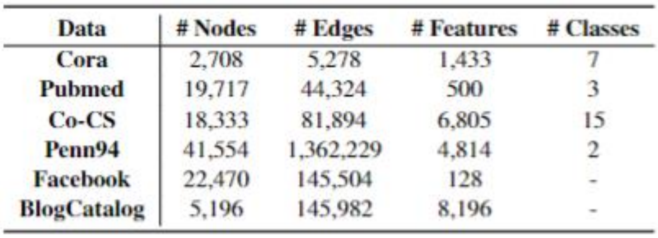
where <·,·> is a similarity function, $\mathrm{z}_{@k}$ is a similarity metric between two top-k ranking lists

The total loss is defined as the sum of the distillation loss and the fairness loss as follows:

$$\mathcal{L} = \mathcal{L}_{KD} + \gamma \sum_i \mathcal{L}_{fair}(i), \qquad (8)$$

where γ is a hyperparameter that controls the importance of individual fairness regularization term.

Table 1. Statistics of Datasets



Data	# Nodes	# Edges	# Features	# Classes
Cora	2,708	5,278	1,433	7
Pubmed	19,717	44,324	500	3
Co-CS	18,333	81,894	6,805	15
Penn94	41,554	1,362,229	4,814	2
Facebook	22,470	145,504	128	-
BlogCatalog	5,196	145,982	8,196	-

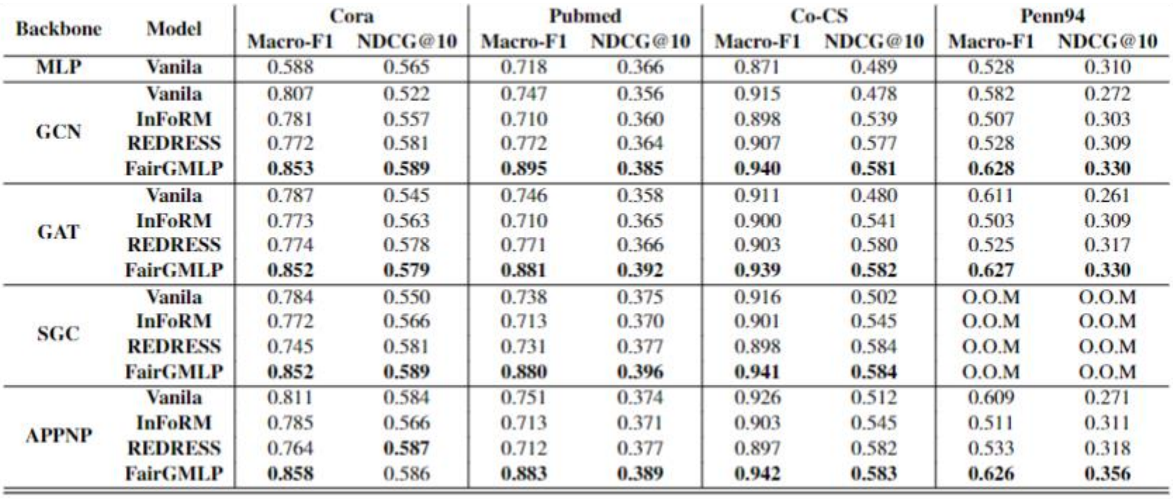
4. Evaluation

4.1 Datasets

To verify the effectiveness of our method in analyzing graphs with different characteristics, we carry out node classification and link prediction. For node classification, we use two citatio n networks (Cora and Pubmed), a co-authorship network (Co-CS), and a gender prediction network (Penn94). For link prediction, we use two social networks Facebook and BlogCatalog. The detailed statistics of these datasets are shown in Table 1.

4.2 Metrics

Table 2. Node Classification Results



Backbone	Model	Cora		Pubmed		Co-CS		Penn94	
		Macro-F1	NDCG@10	Macro-F1	NDCG@10	Macro-F1	NDCG@10	Macro-F1	NDCG@10
MLP	**Vanila**	0.588	0.565	0.718	0.366	0.871	0.489	0.528	0.310
GCN	**Vanila**	0.807	0.522	0.747	0.356	0.915	0.478	0.582	0.272
	InFoRM	0.781	0.557	0.710	0.360	0.898	0.539	0.507	0.303
	REDRESS	0.772	0.581	0.772	0.364	0.907	0.577	0.528	0.309
	FairGMLP	**0.853**	**0.589**	**0.895**	**0.385**	**0.940**	**0.581**	**0.628**	**0.330**
GAT	**Vanila**	0.787	0.545	0.746	0.358	0.911	0.480	0.611	0.261
	InFoRM	0.773	0.563	0.710	0.365	0.900	0.541	0.503	0.309
	REDRESS	0.774	0.578	0.771	0.366	0.903	0.580	0.525	0.317
	FairGMLP	**0.852**	**0.579**	**0.881**	**0.392**	**0.939**	**0.582**	**0.627**	**0.330**
SGC	**Vanila**	0.784	0.550	0.738	0.375	0.916	0.502	O.O.M	O.O.M
	InFoRM	0.772	0.566	0.713	0.370	0.901	0.545	O.O.M	O.O.M
	REDRESS	0.745	0.581	0.731	0.377	0.898	0.584	O.O.M	O.O.M
	FairGMLP	**0.852**	**0.589**	**0.880**	**0.396**	**0.941**	**0.584**	O.O.M	O.O.M
APPNP	**Vanila**	0.811	0.584	0.751	0.374	0.926	0.512	0.609	0.271
	InFoRM	0.785	0.566	0.713	0.371	0.903	0.545	0.511	0.311
	REDRESS	0.764	**0.587**	0.712	0.377	0.897	0.582	0.533	0.318
	FairGMLP	**0.858**	0.586	**0.883**	**0.389**	**0.942**	**0.583**	**0.626**	**0.356**

To evaluate our model in terms of utility, we use the Macro-F1 score for the node classification task and Area-Under-Curve (AUC) for the link prediction task. In addition, we use NDCG@k with k=10 for both tasks in order to evaluate the fairness. The higher value of Macro-F1 and AUC means better performance and the higher value of NDCG@10 means better individual fairness.

4.3 Baselines

We compare two state-of-the-art baselines (InFoRM [3] and REDRESS [4]) with FairGMLP in terms of their performances and individual fairness.

4.4 Results

4.4.1 Node Classification

We present the results of our experiments for the node classification task in Table 2. Here, "O.O.M" represents having an out-of-memory issue, and boldface indicates the highest score. Our framework outperforms existing state-of-the-art models for all four GNN backbone models. It highlights the effectiveness of our approach in achieving superior performance in terms of utility and fairness compared to existing methods. Note that the fairness scores of a vanilla MLP are better than those of vanilla GNNs. In addition, FairGMLPs have far higher F1 scores compared to vanilla MLPs.

This implies that MLPs can enhance individual fairness, and also the distilling structure knowledge can enhance node representation utility.

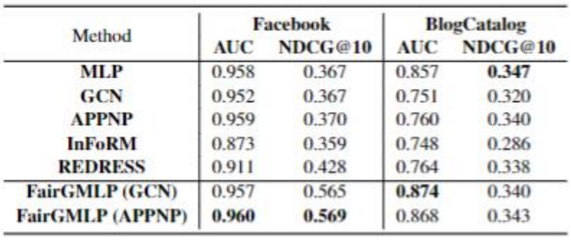
4.4.2 Link Prediction

Table 3 presents quantitative results for link prediction. The boldface indicates the highest score.

Generally, FairGMLPs provide higher AUC and NDCG scores compared to existing state-of-the-art models. We also observe that when using GCN and APPNP teachers, the MLP students consistently improve in both utility and fairness scores across all cases.

These results indicate that our approach can be used to predict the relationship between high performance and fairness among individuals.

Table 3 Link Prediction Results



Method	Facebook		BlogCatalog	
	AUC	NDCG@10	AUC	NDCG@10
MLP	0.958	0.367	0.857	**0.347**
GCN	0.952	0.367	0.751	0.320
APPNP	0.959	0.370	0.760	0.340
InFoRM	0.873	0.359	0.748	0.286
REDRESS	0.911	0.428	0.764	0.338
FairGMLP (GCN)	0.957	0.565	**0.874**	0.340
FairGMLP (APPNP)	**0.960**	**0.569**	0.868	0.343

5. Conclusion

In this paper, we introduce a knowledge distillation framework that injects the knowledge of a GNN teacher into a MLP student. The framework outperforms existing graph representation learning models in terms of utility and individual fairness. In the future, we will examine fairness in graphs directly related to real-world situations, such as heterogeneous graphs, knowledge graphs, and temporal graphs.

Reference

[1] Wu, Felix, et al. "Simplifying graph convolutional networks." *International conference on machine learning*. PMLR, 2019.

[2] Ahn, Seong Jin, and MyoungHo Kim. "Variational graph normalized autoencoders." *Proceedings of the 30th ACM international conference on information & knowledge management*. 2021.

[3] Kang, Jian, et al. "Inform: Individual fairness on graph mining." *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 2020.

[4] Dong, Yushun, et al. "Individual fairness for graph neural networks: A ranking based approach." *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021.