Interpreting Unfairness in Graph Neural Networks via Training Node Attribution

Yushun Dong\textsuperscript{1}, Song Wang\textsuperscript{1}, Jing Ma\textsuperscript{1}, Ninghao Liu\textsuperscript{2}, Jundong Li\textsuperscript{1}

\textsuperscript{1}University of Virginia
\textsuperscript{2}University of Georgia
\{yd6eb, sw3wv, jm3mr, jundong\}@virginia.edu, ninghao.liu@uga.edu

Abstract

Graph Neural Networks (GNNs) have emerged as the leading paradigm for solving graph analytical problems in various real-world applications. Nevertheless, GNNs could potentially render biased predictions towards certain demographic subgroups. Understanding how the bias in predictions arises is critical, as it guides the design of GNN debiasing mechanisms. However, most existing works overwhelmingly focus on GNN debiasing, but fall short on explaining how such bias is induced. In this paper, we study a novel problem of interpreting GNN unfairness through attributing it to the influence of training nodes. Specifically, we propose a novel strategy named Probabilistic Distribution Disparity (PDD) to measure the bias exhibited in GNNs, and develop an algorithm to efficiently estimate the influence of each training node on such bias. We verify the validity of PDD and the effectiveness of influence estimation through experiments on real-world datasets. Finally, we also demonstrate how the proposed framework could be used for debiasing GNNs. Open-source code can be found at https://github.com/yushundong/BIND.

Introduction

Graph data is pervasive among a plethora of realms, e.g., financial fraud detection (Wang et al. 2019; Pourhabibi et al. 2020; Cheng et al. 2020), social recommendation (Fan et al. 2019; Song et al. 2019; Guo and Wang 2020), and chemical reaction prediction (Do, Tran, and Venkatesh 2019; Shi et al. 2020; Kwon et al. 2022). As one of the state-of-the-art approaches to handle graph data, Graph Neural Networks (GNNs) have been attracting increasing attention (Kipf and Welling 2017; Hamilton, Ying, and Leskovec 2017; Veličković et al. 2017). Over the years, various graph analytical tasks have benefited from GNNs, where node classification is among the most widely studied ones (Kipf and Welling 2017; Wu et al. 2019, 2020). Nevertheless, in node classification, GNNs often yield results with discrimination towards specific demographic subgroups described by certain sensitive attributes (Dong et al. 2022a; Dai and Wang 2021a; Agarwal, Lakkaraju, and Zitnik 2021; Zhang et al. 2022b; Wang et al. 2022), such as gender, race, and religion. In many high-stake applications, critical decisions are made based on the classification results of GNNs (Shumovskaia et al. 2020), e.g., crime forecasting (Jin et al. 2020), and the exhibited bias (i.e., unfairness) is destructive for the involved individuals (Dong et al. 2022b,c; Song et al. 2022). To tackle this problem, there has been a line of works focusing on debiasing GNNs in node classification (Dong et al. 2022a; Dai and Wang 2021a; Agarwal, Lakkaraju, and Zitnik 2021; Dong et al. 2021; Loveland et al. 2022; Dai and Wang 2022). Their goal is to relieve the bias in GNN predictions on the test set and in this paper we refer to it as model bias.

In addition to debiasing GNNs, it is also critical to interpret how the model bias arises in GNNs. This is because such an understanding not only helps to determine whether a specific node should be involved in the training set, but also has much potential to guide the design of GNN debiasing methods (Dong et al. 2022a; Loveland et al. 2022; Li et al. 2021). Nevertheless, most existing GNN interpretation methods aim to understand how a prediction is made (Yuan et al. 2020b; Liu, Feng, and Hu 2022) instead of other aspects such as fairness. Consequently, although the graph data has been proved to be a significant source of model bias (Dong et al. 2022a; Li et al. 2021), existing works are unequipped to tackle this problem. In this paper, we aim to address this problem at the instance (node) level. Specifically, given a GNN trained for node classification, we aim to answer: “To what extent the GNN model bias is influenced by the existence of a specific training node in this graph?”

Nevertheless, answering the above question is technically challenging. Essentially, there are three main challenges: (1) Influence Quantification. To depict the influence of each training node on the model bias of GNNs, the first and foremost challenge is to design a principled fairness metric. A straightforward approach is to directly employ traditional fairness metrics (e.g., \(\Delta_{SP}\) for Statistical Parity (Dwork et al. 2012) and \(\Delta_{EO}\) for Equal Opportunity (Hardt, Price, and Srebro 2016a)). However, these metrics are not applicable in our task. The reason is that most of them are computed based on the predicted labels, while a single training node can barely twist these predicted labels on test data (Zhang et al. 2022a; Sun et al. 2020). Consequently, the influence of a single training node on the model bias would be hard to capture. (2) Computation Efficiency. To compute the influence of each training node on the model bias, a natural way is to re-train the GNN on a new graph with this specific training node being deleted and observe how the exhibited model bias changes. However, such a re-training process is prohibitively expensive. (3) Non-I.I.D. Characterization. Graph data goes against the widely
We first present the notations used in this paper. Then, we where

\[ \text{PDD} \]

Therefore, PDD is with finer granularity and is more suitable

vectors, and \( x \) as

the number of classes. We use \( Y \) and \( S \) to denote the ground

truth label and the sensitive attribute for nodes, respectively.

For an \( L \)-layer GNN, we denote the optimized parameters (i.e.,

\( \text{f}_{G} \)) as

\( \text{GNN model}. \) We denote the optimized parameters (i.e.,

\( \text{AGG} \left( z_{i}^{(l)} , h \left( \{ z_{j}^{(l)} : v_{j} \in N(v_{i}) \} \right) \right) . \) (1)

Here \( z_{i}^{(l)} \) is the embedding of node \( i \) at the \( l \)-th layer: \( N(v_{i}) \) is the set of one-hop neighbors around \( v_{i} \); \( h(\cdot) \) is a function with learnable parameters; \( \text{AGG}(\cdot) \) and \( \sigma(\cdot) \) denote the aggregation function (e.g., mean operator) and activation function (e.g., ReLU), respectively. Later on, a loss function \( \text{L}_{y_{i}} \) (e.g., cross-entropy loss) defined on the set of training nodes \( Y_{i} \) is employed for GNN training.

Probabilistic Distribution Disparity

Traditional bias metrics such as \( \Delta_{SP} \) for statistical parity and \( \Delta_{EO} \) for equal opportunity are computed on the predicted class labels. However, a single training node can hardly twist these predicted labels (Zhang et al. 2022a; Sun et al. 2020). Hence the node-level contribution to model bias can barely be captured by traditional bias metrics. To capture the influence of a single training node on model bias, we propose Probabilistic Distribution Disparity (PDD) as a novel bias quantification strategy. PDD can be instantiated with different fairness notions to depict the model bias from different perspectives. Specifically, we assume the population is divided into multiple subgroups in GNNs. To handle the second challenge, we propose an estimation algorithm for the node influence on model bias, which avoids the re-training process and thus achieves better efficiency. To tackle the third challenge, we also characterize the dependency between nodes based on the analysis of the training loss for GNNs. Finally, we introduce how to characterize the dependency between nodes in influence estimation, which tackles the challenge of Non-I.I.D. Characterization.

GANs in Node Classification

In the node classification task, GNNs take the input graph \( G \) and output a probabilistic output matrix \( \hat{Y} \), where the \( i \)-th row in \( \hat{Y} \) is \( \hat{y}_{i} \), i.e., the probabilistic prediction of a node’s membership over all possible classes. Usually, there are multiple layers in GNNs, where the formulation of the \( l \)-th layer can be summarized as:

Preliminaries

We first present the notations used in this paper. Then, we define the problem of interpreting GNN unfairness through quantifying the influence of each specific training node.

Notations. In this paper, matrices, vectors, and scalars are represented with bold uppercase letters (e.g., \( A \)), bold lowercase letters (e.g., \( x \)), and normal lowercase letters (e.g., \( n \)), respectively. We denote an input graph as \( G = \{ V, E, \hat{X} \} \), where \( V = \{ v_{1}, ..., v_{n} \} \) denotes the node set, \( E \subseteq V \times V \) represents the edge set, \( \hat{X} = \{ x_{1}, ..., x_{n} \} \) is the node attribute vectors, and \( x_{i} \) (\( 1 \leq i \leq n \)) represents the attribute vector of node \( v_{i} \). We denote \( G' \) as the new graph with node \( v_{i} \) being deleted from \( G \). Additionally, we employ \( V' \) (\( V' \subseteq V \)) to represent the training node set, where \( |V'| = m \). The nodes in graph \( G \) are mapped to the output space with a trained GNN \( f_{W} \), where \( W \) represents the learnable parameters of the GNN model. We denote the optimized parameters (i.e., the parameters after training) as \( W \). In node classification, the probabilistic classification output for the \( n \) nodes is denoted as \( \hat{Y} = \{ \hat{y}_{1}, ..., \hat{y}_{n} \} \), where \( \hat{y}_{i} \in \mathbb{R}^{c} \), and \( c \) is the number of classes. We use \( Y \) and \( S \) to denote the ground truth label and the sensitive attribute for nodes, respectively. For an \( L \)-layer GNN \( f_{W} \), we define the subgraph up to \( L \) hops away centered on \( v_{i} \) as its computation graph (denoted as \( G'_{i} = \{ V_{i}, E_{i}, \hat{X}_{i} \} \)). Here \( V_{i}, E_{i} \), and \( \hat{X}_{i} \) denote the set of nodes, edges, and node attributes in \( G_{i} \), respectively. It is

worth noting that existing works have proven that \( G_{i} \) fully determines the information \( f_{W} \) utilizes to make the prediction of \( v_{i} \) (Ying et al. 2019). For node \( v_{i} \), we use \( V'_{i} \) to indicate the intersection between \( V_{i} \) and \( V' \), i.e., \( V'_{i} = V_{i} \cap V' \), which is the set of training nodes in \( G_{i} \).

Problem Statement. The problem of interpreting GNN unfairness is formally defined as follows.

Problem 1. GNN Unfairness Interpretation. Given the graph data \( G \) and a GNN model \( f_{W} \) trained based on \( G \), we define the problem of interpreting GNN unfairness as to quantify the influence of each training node to the unfairness exhibited in GNN predictions on the test set.

Methodology

In this section, we first briefly introduce GNNs for the node classification task. Then, to tackle the challenge of Influence Quantification, we propose Probabilistic Distribution Disparity (PDD) to measure model bias and define node influence on the bias in a trained GNN. Furthermore, to tackle the challenge of Computation Efficiency, we design an algorithm to estimate the node influence on the model bias. Finally, we introduce how to characterize the dependency between nodes in influence estimation, which tackles the challenge of Non-I.I.D. Characterization.

GANs in Node Classification
interest in different sensitive subgroups. Compared with traditional fairness metrics, continuous changes brought by each specific training node are reflected in the measured distributions, and Wasserstein distance is theoretically more sensitive to the change of the measured distributions over other commonly used distribution distance metrics (Arjovsky, Chintala, and Bottou 2017). In addition, we note that the variable of interest depends on the chosen fairness notion in applications, and a larger value of PDD indicates a higher level of model bias. We introduce two instantiations of PDD based on two traditional fairness notions, including Statistical Parity (Dwork et al. 2012) and Equal Opportunity (Hardt, Price, and Srebro 2016a). Both notions are based on binary classification tasks and binary sensitive attributes (generalizations to non-binary cases can be found in Appendix A1). For example, Statistical Parity requires the probability of positive predictions to be the same across two sensitive subgroups, where the variable of interest is the GNN probabilistic output \( \hat{y} \).

It is noteworthy that PDD is a function of \( \hat{y} \) and \( x \), which are functions of \( y \) and \( v \) according to the chain rule. Formally, the optimal parameters \( \hat{y} \) minimize the objective function \( L_{\hat{y}}(G, \hat{y}) \) of the node classification task, so that:

\[
\hat{y} = \arg \min_{y} L_{\hat{y}}(G, y) = \arg \min_{y} \frac{1}{m} \sum_{i=1}^{m} L_{y_i}(G_i, y) .
\]

Here \( L_{y_i}(G_i, y) \) denotes the loss term associated with node \( v_i \); \( G_i \) is the computation graph of \( v_i \); \( m \) is the total number of training nodes. If a training node \( v_i \) is deleted from \( G \), the loss function will change and thus leads to a different \( \hat{y} \). We take \( v_i \) as an example to analyze the influence on \( \hat{y} \) after deleting a training node from \( G \). Traditionally, the existence of node \( v_i \) is considered as a binary state, which is either one (if \( v_i \) exists in \( G \)) or zero (otherwise). But in our analysis, we treat it as a continuous variable to depict the intermediate states of the existence of \( v_i \). Suppose that the existence of \( v_i \) is down-weighted in the training of a GNN on \( G \). This operation leads to two changes in the loss function: (1) the loss term associated with node \( v_i \), i.e., \( L_{v_i}(G, \hat{y}) \), is down-weighted; (2) the loss terms associated with other training nodes in the computation graph of \( v_i \) would also be influenced.

The reason is that these nodes could be affected by the information from node \( v_i \) during the message passing in GNNs (Kipf and Welling 2017; Ying et al. 2019). Based on the above analysis, we define \( \hat{y}_{v_i} \) as the optimal parameter that minimizes the loss function when node \( v_i \) is down-weighted as follows:

\[
\hat{y}_{v_i} = \arg \min_{y} L_{\hat{y}}(G, y)
\]

\[
= \hat{y} - \epsilon \left( L_{v_i}(G_i, \hat{y}) + L_{v_i}(G_i, \hat{y}) \right),
\]

where \( \epsilon \in [0,1/m] \) controls the scale of down-weighting. In an illustration in Fig. 1 shows how down-weighting \( v_i \) affects the loss values of training nodes in its computation graph. To formally characterize how node \( v_i \) influences \( \hat{y} \), we have Theorem 1 as follows (see proofs in Appendix C).

**Theorem 1.** According to the optimization objective of \( \hat{y}_{v_i} \) in Eq. (4), we have

\[
\frac{d\hat{y}_{v_i}}{d\epsilon} \bigg|_{\epsilon=0} = \left( \frac{\partial L_{\hat{y}}(G, \hat{y})}{\partial \hat{y}^2} \right)^{-1} \left( \frac{\partial L_{v_i}(G_i, \hat{y})}{\partial \hat{y}} + \frac{\partial L_{v_i}(G_i, \hat{y})}{\partial \hat{y}} \right).
\]

Then, we characterize the influence of down-weighting node \( v_i \) on the value of PDD. We present Corollary 1 based on the chain rule as follows (see the proofs in Appendix C).

**Corollary 1.** Define the derivative of \( \Gamma \) w.r.t. \( \epsilon \) at \( \epsilon = 0 \) as \( I_\Gamma(v_i) \). According to Theorem 1, we have

\[
I_\Gamma(v_i) = \left( \frac{\partial I_\Gamma}{\partial \hat{y}} \right) \bigg|_{\epsilon=0} \frac{d\hat{y}_{v_i}}{d\epsilon} \bigg|_{\epsilon=0} .
\]

With Corollary 1, we can now estimate the value change of \( \Gamma \) when node \( v_i \) is down-weighted via

\[
\Gamma_{v_i} = \Gamma_{0,v_i} \approx -\epsilon \cdot I_\Gamma(v_i) + o(\epsilon) \approx -\epsilon \cdot I_\Gamma(v_i).
\]
Figure 1: An illustration of how down-weighting node \( v_i \) influences the loss values of the training nodes in \( G_i \) (including \( v_i, v_j, v_j, v_k, v_l, \) and \( v_m \)). Scenarios from \( \epsilon = 0 \) to \( \epsilon = 1/m \) are presented.

Algorithm 1: Node Influence on Model Bias Estimation

| Line | Description |
|------|-------------|
| 1    | Initialize \( \mathcal{I}_T = \emptyset \); |
| 2    | Compute \( \{ \frac{\partial L}{\partial W} : v_i \in \mathcal{V}' \} \) based on \( f_W \); |
| 3    | while \( v_i \in \mathcal{V}' \) do |
| 4    | Compute \( \Delta \mathcal{I}_T(v_i) \) according to Eq. (5) and (8); |
| 5    | Compute \( \Delta \mathcal{I}_T(v_i) \) according to Eq. (6); |
| 6    | Compute \( \Gamma_{v_i}^\prime \) according to Eq. (7); |
| 7    | Append element \( \Gamma_{v_i}^\prime \) onto \( \mathcal{I}_T \); |
| 8    | end while |
| 9    | return \( \mathcal{I}_T \); |

Non-I.I.D. Characterization

Generally, there are two types of dependencies between a training node \( v_i \) and other nodes in its computation graph \( G_i \), namely its dependency on other training nodes and its dependency on test nodes. The dependency between training nodes directly influences \( W \) during GNN training, and thus influences the probabilistic outcome of all test nodes. Hence it is critical to properly characterize the dependency between \( v_i \) and other training nodes. Specifically, we aim to characterize how the loss summation of all training nodes in \( G_i \) changes due to the existence of \( v_i \). We denote the training nodes other than node \( v_i \) in \( G_i \) as \( \mathcal{V}_i^\prime \setminus \{ v_i \} \). For any node \( v_j \in \mathcal{V}_i^\prime \setminus \{ v_i \} \), we denote \( G_{j, i} \) as the computation graph of node \( v_j \) with node \( v_i \) being deleted.

\[
L_{v_j}(G_{j, i}, W) = \sum_{v_j \in \mathcal{V}_{i}^\prime \setminus \{ v_i \}} (L_{v_j}(G_{j, i}, W) - L_{v_j}(G_{j, i}, W)).
\] (8)

The first term represents the summation of loss for nodes in \( \mathcal{V}_i^\prime \setminus \{ v_i \} \) on \( G_i \), and the second term denotes the summation of loss for these nodes on \( G_{j, i} \). In this regard, \( L_{v_j}(G_{j, i}, W) \) generally depicts to what extent the loss summation changes for nodes in \( \mathcal{V}_i^\prime \setminus \{ v_i \} \) on graph \( G \) compared with \( G_{j, i} \). If \( v_i \) is down-weighted by a certain degree, the change of the loss summation for nodes in \( \mathcal{V}_i^\prime \setminus \{ v_i \} \) can be depicted by a linearly re-scaled \( L_{v_j}(G_{j, i}, W) \), as described in Eq. (4).

Additionally, there could also be dependencies between \( v_i \) and test nodes in \( G_i \), as \( v_i \) can influence the representations of its neighboring test nodes due to the information propagation mechanism in GNNs during inference. Such a dependency could also influence the value of PDD when \( v_i \) is deleted from \( G \). Correspondingly, we introduce the characterization of the dependency between \( v_i \) and test nodes. Specifically, we present an upper bound to depict the normalized change magnitude of the neighboring test nodes’ representations when a training node \( v_i \) is deleted. Here the analysis is based on the prevalent GCN model (Kipf and Welling 2017), and can be easily generalized to other GNNs. Following widely adopted assumptions in (Huang and Zitnik 2020; Xu et al., 2018), we have Proposition 1 (see the proofs in Appendix C).

Proposition 1. Denote the representations of node \( v_j \) \( (v_j \in \mathcal{V}_i^\prime \setminus \mathcal{V}_i^\prime) \) based on \( G \) and \( G_{j, i} \) as \( z_j \) and \( z_j^* \), respectively. Define \( \delta_j^{(j, i)} \) and \( \phi_j^{(j, i)} \) as the distance from \( v_j \) to \( v_i \) and the number of all possible paths from \( v_j \) to \( v_i \), respectively. Define the set of geometric mean node degrees of \( q_j^{(j, i)} \) paths as \( \mathcal{D} = \{ d_j^{(j, i)}, ..., d_j^{(j, i)} \} \). Define \( q_j^{(j, i)} \) as the minimum value of \( \mathcal{D} \). Assume the norms of all node representations are the same. We then have \( ||z_j^* - z_j||_2 \leq q_j^{(j, i)} \) / \( q_j^{(j, i)} \).
Complexity Analysis

To better understand the computational cost, here we analyze the time complexity of estimating $\Delta \Gamma$ according to Algorithm 1. We denote the number of parameters in $W$ and the average number of training nodes in the computation graph of an arbitrary training node as $t$ and $\bar{r}$, respectively. For each node $v_t$, the time complexity to compute $\partial L_{v_t}(G, W) / \partial W$ and $\partial L_{v_t}(G, W) / \partial W$ is $O(t)$ and $O(\bar{r} t)$, respectively. Hence the time complexity is $O(n \bar{r} t)$ to traverse all training nodes. For the Hessian matrix inverse, we employ a widely-used estimation approach (see Appendix A) with linear time complexity w.r.t $t$. Thus the time complexity of Eq. (5) and (8) is $O(n \bar{r} t)$. Additionally, the time complexity of Eq. (6) and (7) is $O(n m t)$ and $O(m)$, respectively. To summarize, the time complexity of Algorithm 1 is $O(n \bar{r} t)$.

Experiments

We aim to answer the following research questions in experiments. **RQ1**: How efficient is BIND in estimating the influence of training nodes on the mode bias? **RQ2**: How well can BIND estimate the influence of training nodes on the model bias? **RQ3**: How well can we debias GNNs via deleting harmful training nodes based on our estimation? More details of experimental settings, supplementary experiments, and further analysis are in Appendix B.

Experimental Setup

**Downstream Task & Datasets.** Here the downstream task is node classification. Four real-world datasets are adopted in our experiments, including Income, Recidivism, Pokec-z, and Pokec-n. Specifically, Income is collected from Adult Data Set (Dua and Graff 2017). Each individual is represented by a node, and we establish connections (i.e., edges) between individuals following a similar criterion adopted in (Agarwal, Lakkaraju, and Zitnik 2021). The sensitive attribute is race, and the task is to classify whether a defendant is on bail or not. Recidivism is collected from (Jordan and Freiburger 2015). A node represents a defendant released on bail, and defendants are connected based on their similarity. The sensitive attribute is race, and the task is to classify whether a defendant is on bail or not. Pokec-z and Pokec-n are collected from Pokec, which is a popular social network in Slovakia (Takac and Zabovsky 2012). In both datasets, each user is a node, and each edge stands for the friendship relation between two users. Here the locating region of users is the sensitive attribute. The task is to classify the working field of users. More details are in Appendix B.

**Baselines & GNN Backbones.** We compare our method with three state-of-the-art GNN debiasing baselines, namely FairGNN (Dai and Wang 2021a), NIFTY (Agarwal, Lakkaraju, and Zitnik 2021), and EDITS (Dong et al. 2022a). To perform GNN debiasing, FairGNN employs adversarial training to filter out the information of sensitive attributes from node embeddings; NIFTY maximizes the agreement between the predictions based on perturbed sensitive attributes and unperturbed ones; EDITS pre-processes the input graph data to be less biased via attribute and structural debiasing. We mainly present the results of using GCN (Kipf and Welling 2017) as the backbone GNN model, while experiments with other GNNs are discussed in Appendix B.

**Evaluation Metrics.** First, we employ running speedup factors to evaluate efficiency. Second, we use the widely adopted Pearson Correlation (Koh and Liang 2017; Chen et al. 2020) between the estimated and actual $\Delta \Gamma$ to evaluate the effectiveness of node influence estimation. Third, we adopt two traditional fairness metrics, namely $\Delta \Gamma_{SP}$ (the metric for Statistical Parity) (Dwork et al. 2012) and $\Delta \Gamma_{EO}$ (the metric for Equal Opportunity) (Hardt, Price, and Srebro 2016b), to evaluate the effectiveness of debiasing GNNs via harmful nodes deletion. Additionally, the classification accuracy is also employed to evaluate the utility-fairness trade-off.

Efficiency of Node Influence Estimation

To answer RQ1, we evaluate the efficiency of $\Delta \Gamma$ estimation by comparing its running time with that of GNN re-training. The running time of GNN re-training is computed as follows. We first delete the target node from the original input graph $G$ and re-train the GCN to obtain $f_{W}^{′}$. We then obtain $\Delta \Gamma$ based on the values of $\Gamma$ given by $f_{W}$ and $f_{W}^{′}$. The above running time is defined as the time cost of GNN re-training. The running time averaged across all training nodes is compared between GNN re-training and BIND, and we present the running speedup factors of BIND on the four real-world datasets in Fig. 2. We observe that the running speedup factors are over 450× on all four real-world datasets, which corroborates the efficiency superiority of BIND in estimating the value of $\Delta \Gamma$. Additionally, we observe that the estimation on Pokec-z and Pokec-n datasets has higher speedup factors on both $\Delta \Gamma_{SP}$ and $\Delta \Gamma_{EO}$ compared with the other two datasets. A reason could be that nodes in Pokec-z and Pokec-n have lower average degrees (see Appendix B). This facilitates the computation of $\hat{L}_{V}(G, W)$ (the term that characterizes non-i.i.d.) and corresponding derivatives.

Effectiveness of Node Influence Estimation

We now evaluate the effectiveness of $\Delta \Gamma$ estimation. It is worth noting that the numerical values of the estimated influence on model bias are small for most of the nodes (see Appendix B). Here we introduce a strategy to evaluate the

![Figure 2: Evaluation of efficiency: speedup factors of $\Delta \Gamma_{SP}$ and $\Delta \Gamma_{EO}$ estimation over GNN re-training.](image)
We then construct a series of node sets associated with non-overlapping computation graphs are selected in concert with the basic intuition here is that we select node sets and evaluate the helpful and harmful nodes with top-ranked \( \Delta \Gamma \) values. We then construct a series of node sets associated with the largest positive and negative estimated \( \Delta \Gamma \) summations under different set size thresholds. The range of these thresholds is between zero and a maximum possible value (determined by the training set size). It is worth noting that only nodes with non-overlapping computation graphs are selected in constructing each node set. This ensures that these nodes result in an estimated \( \Delta \Gamma \) equivalent to the summation of their estimated \( \Delta \Gamma \) (see Appendix C). We present the Pearson correlation of estimated \( \Delta \Gamma_{SP} \) and \( \Delta \Gamma_{EO} \) with the actual values on four datasets in Fig. 3. It is worth noting that achieving an exact linear correlation (i.e., Pearson correlation equals one) between the estimated and actual \( \Delta \Gamma \) is almost impossible, since we only employ the first-order Taylor expansion in our estimation for \( \Delta \Gamma \). From Fig. 3, we observe that the estimation achieves Pearson correlation values over 0.9 on both \( \Gamma_{SP} \) and \( \Gamma_{EO} \) across all datasets. Such consistencies between estimated and actual values verify the effectiveness of BIND.

Additionally, to understand how the non-i.i.d. characterization benefits the estimation, we also estimate \( \Delta \Gamma \) with BIND after the non-i.i.d. characterization being disabled (i.e., setting the \( L_{\text{PDD}}(G, W) \) term in Eq. 4 as 0). We present the estimated \( \Delta \Gamma \) v.s. actual \( \Delta \Gamma \) on Income dataset with non-i.i.d. characterization being enabled and disabled in Fig. 4a and 4b, respectively. We observe the correlation decreases between the estimated and actual \( \Delta \Gamma \) after the non-i.i.d. characterization is disabled. Such a decrease is also observed on other datasets in terms of both statistical parity and equal opportunity. Such an observation verifies the contribution of non-i.i.d. characterization to the estimation of \( \Delta \Gamma \).

Finally, we evaluate how well the values of the proposed PDD matches the values of traditional fairness metrics. We collect the value pairs of \( (\Delta_{SP}, \Gamma_{SP}) \) and \( (\Delta_{EO}, \Gamma_{EO}) \) during the GNN re-training process. The values of \( \Delta_{SP} \) v.s. actual \( \Gamma_{SP} \) are presented in Fig. 4c, and the values of \( \Delta_{EO} \) v.s. actual \( \Gamma_{EO} \) are shown in Fig. 4d. We observe a satisfying match between \( \Gamma \) and traditional metrics, which corroborates that PDD is a valid indicator of the fairness level depicted by traditional fairness metrics.

Debiasing via Harmful Nodes Deletion
In this subsection, we demonstrate how BIND could be employed for GNN debiasing. The basic intuition here is to identify and delete those harmful nodes according to the estimated node influence on model bias, and evaluate whether GNNs can be debiased when they are trained on this new graph. Specifically, we set \( \Gamma = \lambda \Gamma_{SP} + (1 - \lambda) \Gamma_{EO} \) and estimate the node influence on \( \Gamma \) to consider both statistical parity and equal opportunity. We then set a budget \( k \), and follow the strategy adopted in Section to select and delete a set of training nodes with the largest positive influence summation on \( \Gamma \) under this budget. We set \( \lambda = 0.5 \) to assign statistical parity and equal opportunity the same weight, and perform experiments with \( k \) being 1% (denoted as BIND 1%) and 10% (denoted as BIND 10%) of the total number of training nodes. We present the results on the four adopted datasets in Table 1. The following observations are made: (1) compared with other baselines, BIND achieves competitive performance (i.e., lower values) on both \( \Delta_{SP} \) and \( \Delta_{EO} \). Hence, training GNNs on a new graph after deleting harmful nodes (to fairness) is an effective approach for GNN debiasing; (2) there is no obvious performance decrease on the model utility of BIND compared with other baselines. We thus argue that deleting harmful nodes can also lead to a satisfying fairness-utility trade-off.

Related Work

Graph Neural Networks. GNNs can be divided into spectral-based and spatial-based ones (Wu et al. 2020; Zhou et al. 2020). Spectral GNNs inherit the insights from Convolutional Neural Networks (CNNs) (Bruna et al. 2013), and followed by many works (Defferrard, Bresson, and Vandergheynst 2016; Levie et al. 2018; Kipf and Welling 2017). Their goal is to design graph filters to extract task-related information from the input graphs (Chung and Graham 1997). Differently, spatial GNNs design message-passing mechanisms in the spatial domain to extract information from each node’s neighbors (Wu et al. 2020; Zhou et al. 2020). Various aggregation strategies contribute to different tasks (Velickovic et al. 2017; Xu et al. 2019b; Suresh et al. 2021; Park and Neville 2020).

Algorithmic Fairness. Algorithmic fairness can be defined from different perspectives (Pessach and Shmueli 2020; M. et al. 2021; Du et al. 2020; Caton and Haas 2020; Corbett-Davies and Goel 2019; Mitchell et al. 2021), where Group Fairness and Individual Fairness are two popular notions (Dwork et al. 2012). Generally, group fairness enforces similar statistics (e.g., positive prediction rate in binary classification tasks) across different demographic subgroups (Dwork et al. 2012). Typically, these demographic subgroups are described by certain sensitive attributes, such as gender, race, and religion. Individual fairness argues for similar outputs for similar individuals (Dwork et al. 2012). Algorithmic fairness can be considered in different stages of learning pipelines, including pre-processing (Dong et al. 2022a), in-processing (Dong et al. 2021; Lahoti, Gummadhi, and Weikum 2019; Dai and Wang 2021b), and post-processing (Kang et al. 2020). Particularly, re-weighting train-
we also analyze the non-IID characteristic between nodes to understand how each training node influences model bias. To the best of our knowledge, this is a first-of-its-kind study. Moreover, most of existing methods based on re-weighting training samples are developed under the IID assumption. However, in this paper, we develop a node deletion strategy to achieve GNN debiasing based on influence estimation. We first propose a strategy named Probabilistic Distribution Disparity (PDD), which can be instantiated with any existing GNN. We then propose a novel framework for statistical parity and equal opportunity under different node deletion budgets.

**Interpretation of Deep Learning Models.** Deep learning models have huge parameter size and high complexity (Buhlmester, Münch, and Arens 2021; Samek, Wiegand, and Müller 2017; Fong and Vedaldi 2017; Xu et al. 2019a). To make these models more trustworthy and controllable, many studies have been devoted to improving their transparency (Fong and Vedaldi 2017). Generally, these works are divided into transparency design and post-hoc explanation (Xu et al. 2019a). The basic goal of transparency design is to understand the model in terms of model structure (Liu et al. 2021; Zhang et al. 2019) and training algorithms (Plumb et al. 2019), while post-hoc explanation aims to explain specific prediction results via visualization (Ding et al. 2017) and explanatory examples (Chen et al. 2018). In the realm of learning on graphs, some existing works aim to interpret GNNs (Ying et al. 2019; Luo et al. 2020; Yuan et al. 2020a), and they mainly focus on understanding the utility (e.g., node classification accuracy) of GNNs on the test set. Our work is different from them in two aspects: (1) we focus on interpreting the model bias instead of the utility for GNNs; (2) we aim to understand the model bias via attributing to the training set instead of only focusing on the test set.

**Conclusion**

In this paper, we study a novel problem of characterizing how each training node influences the bias exhibited in a trained GNN. We first propose a strategy named Probabilistic Distribution Disparity (PDD), which can be instantiated with different existing fairness notions, to quantify the node influence on the model bias. We then propose a novel framework named BIND to achieve an efficient influence estimation for each training node. We also develop a node deletion strategy to achieve GNN debiasing based on influence estimation. Extensive experiments verify (1) the consistency between the proposed PDD and traditional fairness metrics; (2) the efficiency and effectiveness of the influence estimation algorithm; and (3) the performance of the proposed strategy on GNN debiasing. We leave interpreting how the unfairness arises in other graph learning tasks as future works.
Acknowledgments
This work is supported by the National Science Foundation under grants IIS-2006844, IIS-2144209, IIS-2223768, IIS-2223769, CNS-2154962, and BCS-2228534, the JP Morgan Chase Faculty Research Award, and the Cisco Faculty Research Award.

References
Agarwal, C.; Lakkaraju, H.; and Zitnik, M. 2021. Towards a Unified Framework for Fair and Stable Graph Representation Learning. In UAI.
Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein GAN. In ICML.
Bruna, J.; Zaremba, W.; Szlam, A.; and LeCun, Y. 2013. Spectral networks and locally connected networks on graphs. arXiv preprint arXiv:1312.6203.
Buhrmester, V.; Münch, D.; and Arens, M. 2021. Analysis of explainers of black box deep neural networks for computer vision: A survey. MLKE.
Caton, S.; and Haas, C. 2020. Fairness in machine learning: A survey. CSUR.
Chen, H.; Si, S.; Li, Y.; Chelba, C.; Kumar, S.; Boning, D.; and Hsieh, C.-J. 2020. Multi-stage influence function. NeurIPS.
Chen, J.; Song, L.; Wainwright, M.; and Jordan, M. 2018. Learning to explain: An information-theoretic perspective on model interpretation. In ICML.
Cheng, D.; Wang, X.; Zhang, Y.; and Zhang, L. 2020. Graph neural network for fraud detection via spatial-temporal attention. TKDE.
Chung, F. R.; and Graham, F. C. 1997. Spectral graph theory. 92. American Mathematical Soc.
Corbett-Davies, S.; and Goel, S. 2019. The measure and mismeasure of fairness: A critical review of fair machine learning. In NeurIPS.
Dai, E.; and Wang, S. 2021a. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In WSDM.
Dai, E.; and Wang, S. 2021b. Towards Self-Explainable Graph Neural Network. In CIKM.
Dai, E.; and Wang, S. 2022. Learning fair graph neural networks with limited and private sensitive attribute information. TKDE.
Defferrard, M.; Bresson, X.; and Vandergheynst, P. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. NeurIPS.
Ding, Y.; Liu, Y.; Luan, H.; and Sun, M. 2017. Visualizing and understanding neural machine translation. In ACL.
Do, K.; Tran, T.; and Venkatesh, S. 2019. Graph transformation policy network for chemical reaction prediction. In SIGKDD.
Dong, Y.; Kang, J.; Tong, H.; and Li, J. 2021. Individual Fairness for Graph Neural Networks: A Ranking based Approach. In SIGKDD.
Dong, Y.; Liu, N.; Jalaian, B.; and Li, J. 2022a. EDITS: Modeling and Mitigating Data Bias for Graph Neural Networks. In The Web Conf.
Li, P.; Wang, Y.; Zhao, H.; Hong, P.; and Liu, H. 2021. On dyadic fairness: Exploring and mitigating bias in graph connections. In *ICLR*.

Liu, G.; Sun, X.; Schulte, O.; and Poupart, P. 2021. Learning Tree Interpretation from Object Representation for Deep Reinforcement Learning. *NeurIPS*.

Liu, N.; Feng, Q.; and Hu, X. 2022. Interpretablity in Graph Neural Networks. *Graph Neural Networks: Foundations, Frontiers, and Applications*.

Loveland, D.; Pan, J.; Bhathena, A. F.; and Lu, Y. 2022. FairEdit: Preserving Fairness in Graph Neural Networks through Greedy Graph Editing. *arXiv preprint arXiv:2201.03681*.

Luo, D.; Cheng, W.; Xu, D.; Yu, W.; Zong, B.; Chen, H.; and Zhang, X. 2020. Parameterized explainer for graph neural network. In *NeurIPS*.

M., N.; M., F.; S., N.; L., K.; and G., A. 2021. A survey on bias and fairness in machine learning. *CSUR*.

Ma, J.; Deng, J.; and Mei, Q. 2021. Subgroup generalization and fairness of graph neural networks. In *NeurIPS*.

Mitchell, S.; Potash, E.; Barocas, S.; D’Amour, A.; and Lum, K. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annu. Rev. Stat. Appl.*

Park, H.; and Neville, J. 2020. Role Equivalence Attention for Label Propagation in Graph Neural Networks. In *PAKDD*.

Pessach, D.; and Shmueli, E. 2020. Algorithmic fairness. *arXiv preprint arXiv:2001.09784*.

Petrović, A.; Nikolić, M.; Radovanović, S.; Delibašić, B.; and Jovanović, M. 2022. FAIR: Fair adversarial instance re-weighting. *Neurocomputing*.

Plumb, G.; Al-Shedivat, M.; Cabrera, A. A.; Perer, A.; Xing, E.; and Talwalkar, A. 2019. Regularizing black-box models for improved interpretability. *arXiv preprint arXiv:1902.06787*.

Pourhabibi, T.; Ong, K.-L.; Kam, B. H.; and Boo, Y. L. 2020. Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*.

Samek, W.; Wiegand, T.; and Müller, K.-R. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Shi, C.; Xu, M.; Guo, H.; Zhang, M.; and Tang, J. 2020. A graph to graphs framework for retrosynthesis prediction. In *ICML*.

Shumovskaia, V.; Fedyanin, K.; Sukharev, I.; Berestnev, D.; and Panov, M. 2020. Linking bank clients using graph neural networks powered by rich transactional data. In *DSSA*.

Song, W.; Dong, Y.; Liu, N.; and Li, J. 2022. GUIDE: Group Equality Informed Individual Fairness in Graph Neural Networks. In *KDD*.

Song, W.; Xiao, Z.; Wang, Y.; Charlin, L.; Zhang, M.; and Tang, J. 2019. Session-based social recommendation via dynamic graph attention networks. In *WSDM*.

Sun, Y.; Wang, S.; Tang, X.; Hsieh, T.-Y.; and Honavar, V. 2020. Adversarial attacks on graph neural networks via node injections: A hierarchical reinforcement learning approach. In *The Web Conf*.

Suresh, S.; Budde, V.; Neville, J.; Li, P.; and Ma, J. 2021. Breaking the limit of graph neural networks by improving the assortativity of graphs with local mixing patterns. *arXiv preprint arXiv:2106.06586*.

Takac, L.; and Zabovský, M. 2012. Data analysis in public social networks. In *International. scient. workshop*.

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

Wang, D.; Lin, J.; Cui, P.; Jia, Q.; Wang, Z.; Fang, Y.; Yu, Q.; Zhou, J.; Yang, S.; and Qi, Y. 2019. A semi-supervised graph attentive network for financial fraud detection. In *ICDM*.

Wang, H.; Wu, Z.; and He, J. 2022. Training Fair Deep Neural Networks by Balancing Influence. *arXiv preprint arXiv:2201.05759*.

Wang, Y.; Zhao, Y.; Dong, Y.; Chen, H.; Li, J.; and Derr, T. 2022. Improving Fairness in Graph Neural Networks via Mitigating Sensitive Attribute Leakage. In *KDD*.

Wu, F.; Souza, A.; Zhang, T.; Fifty, C.; Yu, T.; and Weinberger, K. 2019. Simplifying graph convolutional networks. In *ICML*.

Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; and Philip, S. Y. 2020. A comprehensive survey on graph neural networks. *IEEE Trans Neural Netw Learn Syst*.

Xu, F.; Uszkoreit, H.; Du, Y.; Fan, W.; Zhao, D.; and Zhu, J. 2019a. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *NLPPC*.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2019b. How Powerful are Graph Neural Networks? In *ICLR*.

Xu, K.; Li, C.; Tian, Y.; Sonobe, T.; Kawarabayashi, K.-i.; and Jegelka, S. 2018. Representation learning on graphs with jumping knowledge networks. In *ICML*.

Yan, B.; Seto, S.; and Apostoloff, N. 2022. FORML: Learning to Reweight Data for Fairness. *arXiv preprint arXiv:2202.01719*.

Ying, R.; Bourgeois, D.; You, J.; Zitnik, M.; and Leskovec, J. 2019. GnneXplainer: Generating explanations for graph neural networks. In *NeurIPS*.

Yuan, H.; Tang, J.; Hu, X.; and Ji, S. 2020a. Xggn: Towards model-level explanations of graph neural networks. In *SIGKDD*.

Yuan, H.; Yu, H.; Gui, S.; and Ji, S. 2020b. Explainability in graph neural networks: A taxonomic survey. *arXiv preprint arXiv:2012.15445*.

Zhang, Q.; Yang, Y.; Ma, H.; and Wu, Y. N. 2019. Interpreting cnns via decision trees. In *CVPR*.

Zhang, S.; Zhu, F.; Yan, J.; Zhao, R.; and Yang, X. 2022a. DOTIN: Dropping Task-Irrelevant Nodes for GNNs. *arXiv preprint arXiv:2204.13429*.

Zhang, W.; Weiss, J. C.; Zhou, S.; and Walsh, T. 2022b. Fairness Amidst Non-IID Graph Data: A Literature Review. *arXiv preprint arXiv:2202.07170*.

Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; and Sun, M. 2020. Graph neural networks: A review of methods and applications. *AI Open*.