TOPOLOGICAL FAIRNESS IN GRAPH REPRESENTATION LEARNING BASED ON EDGE DIFFERENTIATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The fairness is a fundamental problem in graph representation learning (GRL) because it can lead to the failure of downstream tasks for the significant deviations between the results of GRL and the true semantics. However, current research on fairness mainly focuses on the sensitive attributes of node, ignoring the bias caused by the unfairness of the topological structure of the graph. Although some studies usually attribute topological unfairness to homogeneity effects, that is, nodes with the same sensitive attributes are more likely to form connections, the role of topological fairness in graph representation learning still needs to be explored. To address this issue, we first define Topological Fairness Graph (TFG) inspired by the traditional concepts of Statistical Parity and Equal Opportunity. Then, based on TFG, Edge Differential Operator for Topological Fairness (EDO-TF) in Graph representation learning is introduced to measure the variation of topological structure of graph with Probabilistic Distribution Disparit. Finally, we combine the EDO-TF into the loss function in GRL to eliminate the topological unfairness of graph. Our empirical results are tested on eight datasets, indicating that the introduction of TFG can correct many erroneous information orientations, which is a great extension of research on graph fairness solely based on feature information.

1 INTRODUCTION

The goal of Graph Representation Learning (GRL) is to map a graph into an embedding space by integrating the structure and feature information, which can be used in various downstream tasks, such as node classification, graph classification, link prediction, and outlier detection (Wu et al., 2021). GRL has been widely applied in many fields such as social computing, biomedicine, and traffic management. The main methods in GRL include from classical matrix factorization (Chen et al., 2015) and random walk methods (Feng et al., 2020) to modern Graph Neural Networks (GNNs) (Ju et al., 2025). Among these techniques, the GNNs, including Graph Convolutional Networks (GCN) (Zhang et al., 2022), Graph Sampling and Aggregation (GraphSAGE) (Hamilton et al., 2017), Graph Attention Networks (GAT) (Veličković et al., 2018), and among others, have shown breakthrough performance and become the mainstream in current GRL.

However, owing to the data-driven nature of GNNs, bias and unfairness inevitably arise during the process of model training and application(Dai & Wang, 2021; Agarwal et al., 2021; Dong et al., 2022a; 2023c). These biases primarily stem from unequal treatment in historical data, societal biases, sampling biases, which have long-term effects on the model's training, decision-making, and prediction processes, potentially leading to unfair decisions for specific demographic groups. To tackle the challenge of unfairness, researchers have proposed a series of classic fairness concepts, mainly including group fairness(Zemel et al., 2013; Zafar et al., 2017) and individual fairness(Dwork et al., 2012; Kearns et al., 2019; Kusner et al., 2017), which only focus on unfairness caused by attribute sensitivity. Meanwhile, significant progresses have been made in methods for addressing fairness problems. These techniques are mainly divided into three categories: adversarial debiasing, fairness constraints, and other innovative debiasing strategies, such as Bayesian-based methods and data preprocessing-based methods(Dong et al., 2023a). However, because the result of information extraction is strongly dependent on graph structure, how to reasonably quantify and eliminate the impact of complex graph structures on information extraction remains an urgent challenge. The core issue lies in the lack of definitions and methods for structural fairness for graph in previous studies.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To our knowledge, the earliest definition of topological fairness was topological bias which is commonly attributed to the homophily effect of sensitive attributes, where nodes with same sensitive attribute are more likely to establish connections (Spinelli et al., 2022a). Although homogeneity effect is an important source of bias, it cannot cover all types of topological biases because topological structure can be influenced by various factors and is not always highly correlated with sensitive attributes. For instance, in a community of technology enthusiasts, individuals connect based on shared interests. Although such connections are not closely associated with sensitive attributes, they can still inadvertently introduce bias.

So, in this paper, we introduce a definition inspired by the traditional concepts of Statistical Parity (SP) and Equal Opportunity (EO), Topological Fairness Graph (TFG), for the research on topological fairness of graph by using the term topology to describe the structure of graph. It is important to note that, unlike Structural Fairness (Wang et al., 2024), which primarily addresses long-tail effects, topological fairness focuses on exploring the unbiased nature of topological structures. Then, Edge Differential Operator for Topological Fairness (EDO-TF) in GNNs is employed to measure the variation of topological structure of graph with Probabilistic Distribution Disparit (PDD). Using EDO-TF, neighborhood information can be divided into two classes, appropriate information and inappropriate information. The former is beneficial for fair prediction and the latter is harmful for fairness which can be considered as a confounding factor that amplifies bias. The overall motivation of our reaserch is displayed in Figure 1. Our contributions can be summarized as follows,

- After rethinking topological bias from a fine-grained perspective, we introduce the concept of Topological Fairness Graph and present a theoretical analysis.
- Edge Differential Operator for Topological Fairness was proposed to eliminate the influence of confounding factors and can be combined with GNNs to determine the optimal fairness graph.
- We conduct experiments on eight datasets, and the results demonstrate that our method achieves a competitive fairness-utility trade-off, validating the effectiveness of the proposed approach.

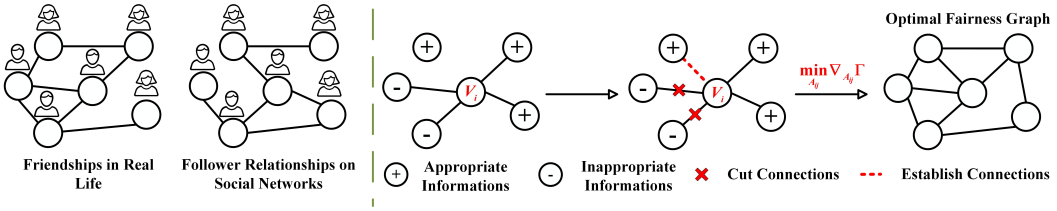


Figure 1: In this figure, we illustrate the motivation in this research. There exist multiple types of relationships within a single group of nodes, including but not limited to friendships, familial connections, and social media interactions. For each node v_i, during the neighborhood information aggregation, our goal is to exclude all inappropriate information while incorporating all appropriate information. By minimizing the value of EDO-TF , we can achieve this objective.

2 Related Work

2.1 Graph Neural Networks

Graph Neural Networks encode attributes and structure into vector representations, effectively capturing useful graph structural information. Some well-known classical models, such as GCN (Zhang et al., 2022), GraphSAGE (Hamilton et al., 2017), GAT (Veličković et al., 2018), and Graph Isomorphism Network (GIN) (Wang & Zhang, 2022), have achieved significant success. The GCN aggregates features of neighboring nodes, extending the concept of traditional convolution to graph data through spectral methods. The GAT incorporates attention mechanisms to dynamically weigh the importance of adjacent nodes for feature aggregation. The GraphSAGE samples a fixed number of neighboring nodes and aggregates their features, suitable for large-scale graph data. The GIN learns graph isomorphism through a powerful aggregation function, aiming to elevate GNN capabilities to

the level of the Weisfeiler-Lehman test. These methods are widely used for processing graph data, such as social networks, chemical structures, and transportation networks. Although the process of aggregating information varies among different methods, structure based information aggregation is the fundamental approach in graph representation learning. Because these models explicitly propagate information along graph edges, their predictions may be affected by topology-induced group disparities. The magnitude and direction of this effect, however, depend on the graph, the task, and the learned model rather than on topology alone.

2.2 Fairness in Graph Neural Networks

Fairness is a multilevel connotation concept that encompasses various aspects, notably Group Fairness (Zemel et al., 2013; Zafar et al., 2017) and Individual Fairness (Dwork et al., 2012; Kearns et al., 2019; Kusner et al., 2017). Frontier research in the field, such as FairGNN (Dai & Wang, 2021), NIFTY (Agarwal et al., 2021), BIND (Dong et al., 2023c), FairSIN (Yang et al., 2024) has made significant progress, achieving debiasing while ensuring model efficiency and effectiveness. FairGNN employs an adversarial component to filter out sensitive attribute information from node embeddings, ensuring that the GNN classifier makes predictions independent of sensitive attribute, thereby achieving model debiasing. NIFTY simultaneously considers fairness and stability by introducing a novel objective function to optimize consistency between predictions made with perturbed sensitive attribute and their unperturbed counterparts. BIND introduces a novel strategy for quantifying the influence of each training node on model bias and debiasing GNNs by removing nodes that are harmful to fairness. FairSIN proposes an innovative neutralization-based paradigm, where additional fairness-facilitating features are incorporated into node features or representations before message passing.

Some recent studies have empirically demonstrated that message passing process can exacerbate fairness bias (Yang et al., 2024). Consequently, addressing bias in graph structure has emerged as a novel and pressing challenge. However, this issue still lacks a formal definition, and thus we propose the concept of topological fairness graph. It is worth noting that, unlike the structural fairness (Wang et al., 2024), which primarily focuses on the issue of long-tail effects, TFG aims to explore the unbiased nature of topological structures. Recent methods, such as EDITS (Dong et al., 2022a), FairDrop (Spinelli et al., 2022b), REFEREE (Dong et al., 2022b), Graphair (Fenyu et al., 2021), have made some progress. FairDrop proposes an edge masking algorithm to counter-act homophily and improve fairness. REFEREE provides structural explanations of topology bias and proposes a novel post-hoc explanation framework. EDITS focuses on input graph data, aiming to preprocess the data to enable fairer GNN outcomes by feeding the model with less biased information. Graphair considers both feature masking and topology modification, automatically discovers fairness-aware augmentations from input graphs. Although several studies have explored the fairness implications of graph topology structure, these methods have not considered the connectivity of edges, which means that removing one edge may result in changes to other edges. Therefore, we introduce the method of EDO-TF to address this issue at the graph-level.

3 Methodology

3.1 Preliminaries

A graph is given by $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ represents the set of nodes with $|\mathcal{V}| = n$ indicating the total number of nodes, $\mathcal{E}$ represents the set of edges with $|\mathcal{E}| = m$ indicating the total number of edges. $A \in \mathcal{R}^{n \times n}$ denotes the adjacency matrix of the input graph $\mathcal{G}$, where $A_{ij} = 1$ if there is an edge between nodes v_i and v_j, and $A_{ij} = 0$ otherwise. $X \in \mathcal{R}^{n \times d}$ denotes the node feature matrix, where d is the node attribute dimension and $x_i \in \mathcal{R}^d$ is the feature vector for node $v_i \in \mathcal{V}$. Furthermore, $\mathcal{S} \in \{0, 1\}^n$ represents the binary sensitive attribute of the nodes, Y is the true labels and $\hat{Y}$ is the predicted labels of nodes respectively.

When assessing fairness, two pivotal concepts commonly introduced are Statistical Parity (SP, also known as Demographic Parity) (Dwork et al., 2012) and Equal Opportunity (EO) (Pearl, 2009). Statistical Parity primarily focuses on whether distribution of predictions is consistent across different sensitive subgroups. Specifically, it aims to ensure that, for each value of a sensitive attribute, the predictions maintain roughly equal proportions between positive and negative classes. Mathemati-

cally, Statistical Parity can be expressed as follows,

$$P(\hat{y} = 1|\mathcal{S} = 0) = P(\hat{y} = 1|\mathcal{S} = 1). \tag{1}$$

Equal Opportunity is another important fairness metric that is usually applied to binary classification problems. It concerns the model's ability to distinguish positive instances in different subgroups defined by sensitive attribute. Specifically, EO requires the model to maintain similar True Positive Rate (also known as sensitivity or recall) across different subgroups defined by sensitive attribute. This means that the model should correctly identify positive instances with similar probabilities regardless of the subgroup. In mathematical terms, EO can be formulated as follows,

$$P(\hat{y} = 1|S = 0, Y = 1) = P(\hat{y} = 1|S = 1, Y = 1). \tag{2}$$

Probabilistic Distribution Disparit (PDD) is a novel bias quantification strategy proposed by (Dong et al., 2023c), which can be instantiated with various fairness notions to depict the model bias from different perspectives. We adopt the PDD as the fairness metric in our study. Specifically, PDD is defined as the Wasserstein-1 distance between the probability distributions of a variable of interest in different sensitive subgroups. In this paper, we denote the PDD instantiated with SP as Γ_{SP} and with EO as Γ_{EO} respectively. The details of this subsection can be obtained in Appendix A.

The Structural Causal Model (SCM), conceived by Judea Pearl (Pearl, 2009), is proposed for modeling causal relationships among variables. Central to SCM is the utilization of a directed graph, typically a Directed Acyclic Graph (DAG).

3.2 Theoretical Analysis of Topological Fairness of Graph

Topological bias is commonly attributed to the homophily effect, where nodes with the same sensitive attribute are more likely to establish connections (Spinelli et al., 2022b; Dong et al., 2023b). This phenomenon can be formalized as follows,

$$P(A_{ij} = 1|S_i = S_j) \gg P(A_{ij} = 1|S_i \neq S_j). \tag{3}$$

Given that the information aggregation process can exacerbate fairness bias (Yang et al., 2024), the fairness issues inherent in complex graph structures have become increasingly urgent to address. A key question in this context is how topological structures influence fairness predictions.

To answer this question, we introduce the concept of topological fairness Graph (TFG). The goal of topological fairness is to explore the unbiased nature of topological structures and to discover the potential invariant rationales (Wu et al., 2022; Li et al., 2022) within graphs. In this context, we categorize neighborhood information into appropriate and inappropriate information. Inappropriate information is regarded as a confounding factor that causes bias during the information aggregation process. GNNs can eliminate topological bias and achieve topological fairness if and only if all appropriate information is considered and inappropriate information is excluded.

Definition 1 (Topological Fairness Graph) Let $\mathcal{G}$ be a simple graph, and v be a node of $\mathcal{G}$. Supposed N_v is the neighbor set of v in $\mathcal{G}$, N'_v is any subset of V which can be regarded as a potential neighbor set of v, for a given GNN, $\mathcal{G}$ is a Topological Fairness Graph if and only if that any v satisfies the inequality as follows,

$$P(\hat{y} = y|\Psi(v, N_v)) \geq P(\hat{y} = y|\Psi(v, N'_v)), \tag{4}$$

where $\Psi(\cdot, \cdot)$ is the information aggregation method of GNN.

For a given information aggregation method of GNN, this definition shows that a graph is not topologically fair if there is a node whose neighbor set cannot lead to an optimal prediction, which indicates there exists a new neighbor set for the node to make a better prediction. Obviously, it is difficult to directly use this definition for identification topological fairness graph and cannot be combined with GRL. So we will discuss the way of identification on topological fairness graph based on SCM and provide a method, Edge differential operator for Topological Fairness, to combine the topological fairness graph with GNN.

We provide a theoretical analysis of our definition from a causal perspective using SCM to model the causal relationships among six variables. The details of the discussion is listed in Appendix B with Figure5.

To effectively identify whether a graph is a TFG and integrate with GNN, we propose the method of Edge Differential Operator for Topological Fairness (EDO-TF).

Definition 2 (Edge Differential Operator for Topological Fairness) Let Γ represent the PDD values and A denote the adjacency matrix of graph $\mathcal{G}$. Assuming Γ is differentiable with respect to A, we define Edge Differential Operator for Topological Fairness as $\nabla_{A_{ij}}\Gamma$.

The rationale behind this definition is that modifications to the topological structure of an input graph lead to changes in PDD values. If a topological change results in a higher PDD value, it indicates that the change leads to more unfair outcomes. This suggests that the modified topological structure contains more inappropriate neighborhood information. Therefore, by achieving the minimal value of $\nabla_{A_{ij}}\Gamma$, we can effectively eliminate the influence of inappropriate neighborhood information. Specifically, to calculate the value of $\nabla_{A_{ij}}\Gamma$, we apply the chain rule and decompose the influence of the topology A on the Γ into two parts: the influence of A on the optimal parameters $\hat{W}$, and the effect of these optimal parameters $\hat{W}$ on the Γ . It is important to note that Γ is a function of $\hat{W}$ for a trained GNN, and the optimal parameters $\hat{W}$ minimize the objective function in the node classification task (Dong et al., 2023c). Mathematically, the optimal parameters $\hat{W}$ take the form as follows,

$$\hat{W} \triangleq \underset{W}{\operatorname{argmin}} L_{\mathcal{V}}(\mathcal{G}, W) = \underset{W}{\operatorname{argmin}} L_{\mathcal{V}}(X, A, W). \tag{5}$$

Subsequently, to determine how W changes with alterations in A, we employ the method of implicit differentiation. The steps of the derivative of $\hat{W}$ with respect to A and the algorithm for calculating $\nabla_{A_{ij}}\Gamma$ in Algorithm 1 are listed in the Appendix C. In the final, for a given $\epsilon > 0$, we can obtain the formula as follows,

$$\nabla_{A_{ij}}\Gamma = -(\frac{\partial \Gamma}{\partial W})^{\intercal} \cdot (\frac{\partial^2 L_{\mathcal{V}}(X, A, \hat{W})}{\partial W^2})^{-1} \cdot \frac{\frac{\partial L_{\mathcal{V}}}{\partial W}(X, A_{ij} + \epsilon \cdot E_{ij}, \hat{W}) - \frac{\partial L_{\mathcal{V}}}{\partial W}(X, A_{ij} - \epsilon \cdot E_{ij}, \hat{W})}{2\epsilon}. \tag{6}$$

According to formula (6), the computational complexity of $\nabla_{A_{ij}}\Gamma$ mainly lies in the inverse of the Hessian matrix. We refer to the work of (Dong et al., 2023c) and adopt the efficient estimation method proposed by (Martens, 2010) which has a time complexity linearly dependent on the number of parameters (Koh & Liang, 2017; Chen et al., 2020). Assuming the number of parameters is p, the time complexity for computing the inverse of the Hessian matrix is $\mathcal{O}(p)$. The computational complexity of the first-order derivative is $\mathcal{O}(np)$ and the computational complexity of the second-order derivative is $\mathcal{O}(n^2p)$. So the total time complexity of $\nabla_{A_{ij}}\Gamma$ is about $\mathcal{O}(np + n^2p)$.

3.3 Implementations

We employ the cross-entropy loss to pre-train the model, as it facilitates faster convergence during training and effectively handles the softmax output layer commonly used in classification tasks. Mathematically, the cross-entropy loss is expressed as follows,

$$L_{CE} = -\sum_{i=1}^{n} [y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)], \tag{7}$$

where y_i represents the ground truth label and $\hat{y}_i$ represents the predicted probability for node v_i.

To improve the fairness performance of the model, we encourage GNNs to make fairer predictions by optimizing the topological structure. Specifically, for each input graph, we mitigate bias by minimizing $\nabla_{A_{ij}}\Gamma$ within its domain. The optimization problem can be formulated as follows,

$$\underset{A_{ij}}{\operatorname{argmin}} (\nabla_{A_{ij}}\Gamma + \lambda \parallel A_{ij} \parallel_F^2), \tag{8}$$

where $\| \cdot \|_F$ represents the Frobenius norm, and λ denotes the regularization parameter. Based on this optimization objective, the fairness loss function can be expressed as,

$$L_{EDO-TF} = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \nabla_{A_{ij}}\Gamma. \tag{9}$$

To preserve the intrinsic geometric structure, ensuring that connected nodes remain close in the feature space, we introduce a graph regularization term (Pei et al., 2015) as follows,

$$L_{REG} = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} \parallel X_i - X_j \parallel_2^2 = \text{tr}(XLX^{\mathsf{T}}), \tag{10}$$

where $\| \cdot \|_2$ denotes the L2 norm, tr$(\cdot)$ denotes the trace of a matrix, $L = D - A$ represents the graph Laplacian matrix with D to be a diagonal matrix with diagonal elements equal to the row sums of A.

By combining all of the above loss terms, the total loss function can be expressed as,

$$L = L_{CE} + \alpha L_{EDO-TF} + \beta L_{REG}, \tag{11}$$

where α and β are hyper-parameters.

4 EXPERIMENTS

Table 1: Comparison among baseline methods with EDO-TF. (↑) denotes the larger, the better; (↓) denotes the opposite. Best ones are in bold.

Encoder	Method	Recidivism			Income			Pokec-n			Pokec-z		
		(↑) ACC	(↓) Δ_{SP}	(↓) Δ_{EO}	(↑) ACC	(↓) Δ_{SP}	(↓) Δ_{EO}	(↑) ACC	(↓) Δ_{SP}	(↓) Δ_{EO}	(↑) ACC	(↓) Δ_{SP}	(↓) Δ_{EO}
GCN	Vanilla	**88.9±0.5**	7.55±0.2	5.26±0.7	74.7±1.4	25.9±1.9	32.3±0.8	64.7±0.4	3.77±0.9	2.96±1.1	62.8±1.0	3.95±1.0	2.76±0.9
	FairGNN	85.5±1.7	7.31±0.1	5.17±0.1	69.1±0.6	23.9±2.7	25.6±2.8	63.3±0.5	3.29±2.9	2.46±2.6	**63.3±1.5**	2.07±1.9	1.67±1.4
	NIFTY	84.3±2.9	5.78±1.3	4.72±1.0	70.8±0.9	24.4±1.6	26.9±3.7	63.2±0.2	2.34±1.0	2.79±1.3	62.7±0.3	6.50±2.1	7.58±1.7
	EDITS	87.3±0.1	6.64±0.3	4.51±0.9	68.3±0.8	24.0±1.9	25.9±1.7	64.6±0.9	2.52±0.8	2.61±1.6	62.8±0.9	2.75±1.8	2.24±1.5
	BIND	88.7±0.0	7.40±0.0	5.09±0.1	**75.2±0.0**	19.2±0.6	26.4±0.4	63.5±0.4	6.75±2.3	5.41±3.4	60.6±0.8	5.85±2.0	1.15±0.7
	EDO-TF	88.0±0.9	**5.31±0.3**	**3.82±0.9**	74.2±1.2	**15.7±2.0**	**19.3±2.1**	**64.8±0.5**	**1.96±1.5**	**1.92±1.7**	62.8±1.2	**1.46±1.5**	**1.07±1.2**
GIN	Vanilla	83.5±0.1	7.05±0.6	6.44±0.7	72.1±0.4	27.7±1.2	30.6±0.7	**64.1±0.8**	3.94±1.1	2.58±1.2	62.8±1.3	3.67±1.1	3.12±0.5
	FairGNN	77.9±2.2	6.63±1.4	6.23±1.4	69.8±1.6	23.6±4.5	26.2±4.8	62.1±1.2	3.82±2.4	3.62±2.7	**63.4±1.5**	3.61±2.9	3.29±3.5
	NIFTY	74.5±0.9	5.27±1.1	3.46±1.5	70.3±0.9	26.5±2.3	27.9±2.7	61.7±0.5	3.84±1.0	3.24±1.6	62.5±1.3	2.57±1.2	3.01±1.9
	EDITS	83.6±0.1	5.02±0.3	2.89±0.1	68.8±0.8	25.8±1.8	27.2±1.7	59.8±0.9	2.75±0.8	2.24±1.5	61.9±0.3	3.33±0.8	3.65±1.6
	BIND	**83.7±0.1**	6.94±0.3	5.66±0.2	71.5±0.3	21.2±0.9	25.7±1.0	63.6±0.1	6.68±1.9	5.32±2.2	62.7±0.1	3.76±0.9	1.99±1.5
	EDO-TF	83.1±0.1	**4.79±0.5**	**2.06±1.0**	**72.1±0.1**	**18.8±2.0**	**20.2±1.7**	63.2±0.7	**1.45±1.1**	**2.12±1.6**	62.5±1.0	**1.64±1.8**	**1.16±2.3**

Encoder	Method	Credit			Occupation			NBA			UCSD34		
		(↑) ACC	(↓) Δ_{SP}	(↓) Δ_{EO}	(↑) ACC	(↓) Δ_{SP}	(↓) Δ_{EO}	(↑) ACC	(↓) Δ_{SP}	(↓) Δ_{EO}	(↑) ACC	(↓) Δ_{SP}	(↓) Δ_{EO}
GCN	Vanilla	73.0±0.3	12.9±0.1	10.7±0.5	76.1±0.4	11.8±0.6	12.9±0.8	72.3±0.1	8.17±1.7	19.9±1.9	63.3±0.3	1.27±0.1	1.40±0.1
	FairGNN	72.9±0.2	11.7±0.3	10.2±0.4	76.0±0.6	8.99±0.5	11.7±1.0	71.2±0.5	7.90±1.9	17.7±1.6	61.7±0.5	1.01±0.1	1.29±0.1
	NIFTY	70.6±0.4	9.73±1.0	8.04±1.3	**77.3±0.6**	8.74±1.3	8.98±2.1	70.9±0.2	6.50±0.3	15.39±2.0	61.6±0.4	0.96±0.1	1.27±0.1
	EDITS	70.1±0.7	9.13±0.2	7.88±0.6	74.3±0.1	8.75±1.0	9.53±1.5	70.2±0.4	6.99±1.8	16.74±2.6	60.4±0.3	0.87±0.1	0.89±0.1
	BIND	**73.3±0.0**	7.36±0.1	5.44±0.1	76.9±0.0	6.03±0.4	7.35±0.4	72.0±0.1	5.35±2.0	12.17±1.9	**63.6±0.2**	0.99±0.1	0.64±0.1
	EDO-TF	72.7±0.2	**7.29±0.5**	**5.06±0.4**	76.4±1.0	**4.07±0.3**	**4.46±1.1**	**72.7±0.5**	**4.04±1.2**	**10.98±1.7**	63.3±0.1	**0.36±0.1**	**0.43±0.1**
GIN	Vanilla	**74.1±0.2**	13.0±0.2	10.9±0.6	76.8±0.5	12.0±0.4	13.2±0.7	**73.1±0.3**	8.50±1.5	20.5±1.8	**64.0±0.2**	1.35±0.1	1.45±0.1
	FairGNN	73.8±0.3	12.1±0.4	10.3±0.5	76.2±0.6	9.10±0.5	12.0±1.0	72.5±0.4	8.00±1.7	18.0±1.5	62.9±0.4	1.05±0.1	1.33±0.1
	NIFTY	71.4±0.5	10.0±1.1	8.20±1.4	76.7±0.7	9.00±1.1	9.20±2.2	71.5±0.3	6.60±0.4	16.1±1.9	61.8±0.3	0.98±0.1	1.30±0.1
	EDITS	70.3±0.6	9.20±0.3	8.00±0.7	74.6±0.2	8.80±1.1	9.80±1.4	70.5±0.3	7.20±1.9	17.0±2.5	60.8±0.2	0.89±0.1	0.91±0.1
	BIND	73.5±0.1	**7.40±0.2**	5.50±0.2	77.2±0.1	6.10±0.5	7.40±0.5	72.3±0.1	5.50±2.1	12.5±1.8	63.4±0.1	1.00±0.1	0.65±0.1
	EDO-TF	73.0±0.2	7.40±0.4	**5.20±0.5**	**76.9±0.9**	**3.89±0.2**	**4.60±1.2**	72.5±0.4	**4.10±1.1**	**11.2±1.5**	63.2±0.1	**0.38±0.1**	**0.45±0.1**

4.1 EXPERIMENTAL SETTINGS

We employed eight graph datasets in our experiments, Recidivism, Income, Pokec-n, Pokec-z, Credit, Occupation, NBA, UCSD34. These datasets have been extensively utilized in previous studies on graph fairness learning, covering a diverse range of domains such as finance, criminal justice, and social network. The details of these datasets is listed in in Appendix D including the Table2 .

In our experiments, we utilized two prominent GNNs as the encoding backbone, GCN (Zhang et al., 2022) and GIN (Wang & Zhang, 2022). These two models are widely recognized and employed within the research community for their robust performance on various graph-related tasks. We compared our approach with four methods for addressing fairness issues: FairGNN (Dai & Wang, 2021), NIFTY (Agarwal et al., 2021), EDITS (Dong et al., 2022a), and BIND (Dong et al., 2023c).

The implementation is based on PyTorch. Specifically, we replicate each experiment using three different random seeds, 1, 10, and 100. All GNNs are optimized with Adam optimizer. The training

is conducted for 1,000 epochs with learning rate chosen from $\{0.1, 0.01, 0.001\}$. The GNNs feature a hidden dimension of 16 and employ dropout rates selected from $\{0.2, 0.5, 0.8\}$. Additionally, we experiment with iteration numbers for Hessian matrix inverse estimation ranging from $\{10, 50, 100, 500, 1000, 5000\}$ to optimize performance. We set $\epsilon = 0.001$ and $\lambda = 1$ in all experiments. The hyper-parameters, α and β, are analysed in Subsection 4.5. All experiments are conducted on a device equipped with two GeForce RTX 4080 Super GPUs.

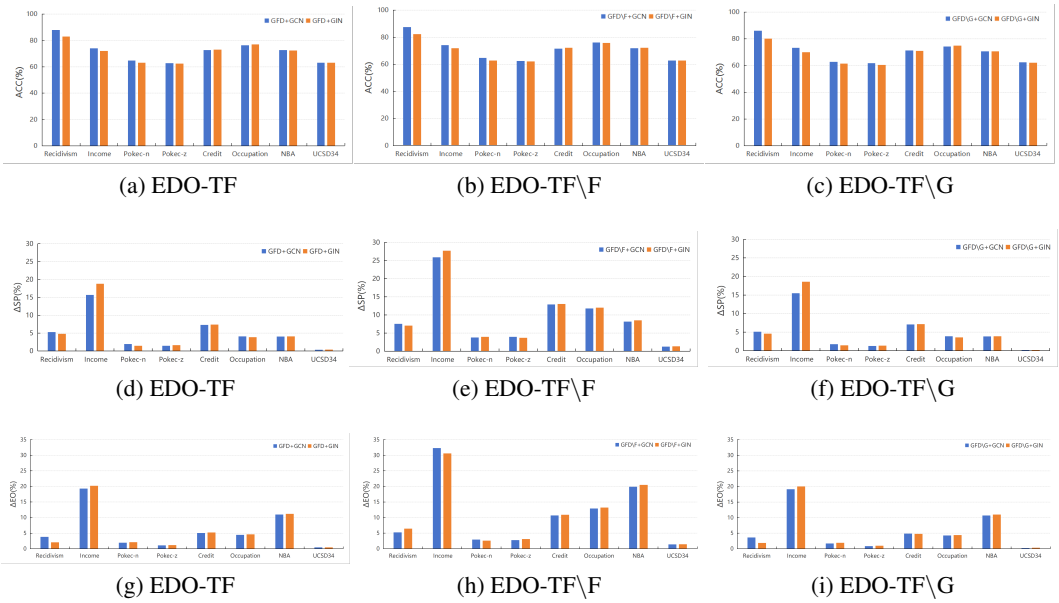
4.2 Effectiveness Analysis

We evaluated the performance of EDO-TF on the accuracy and fairness in node classification task with results listed in Table 1. In the table, Vanilla refers to using GNNs directly for representation learning without considering fairness issues. From these results, we observed the following facts. (1) In the term of the ACC metric, EDO-TF achieved the optimal results 4 times in the 16 comparative experiments which ranked 2nd among all algorithms and 1st among fairness-aware algorithms respectively. (2) In the term of the standard deviation of ACC, which can be regarded as the robustness of the algorithm, the means of EDO-TF ranks 3rd and the difference between EDO-TF and the minimum is only 0.41%. (3) In the two metrics, Δ_{SP} and Δ_{EO}, EDO-TF achieved the first place 15 times and 16 times respectively across the 16 experiments. It clear that our algorithm has gained an overwhelming advantage in addressing the two core fairness issues of SP and EO. Compared with the other four fairness-aware algorithms, the improvement in Δ_{SP} reached a maximum of 8.7% while the improvement in Δ_{EO} reached 7.7%. Even on the Credit dataset where our algorithm did not rank first in Δ_{SP}, its variance was merely 0.2% higher than the top-performing algorithm. (4) When using different architectures as backbones, our method delivered highly consistent performance and did not exhibit the backbone-dependent behavior observed in other algorithms. The maximum ACC difference across different architectures was 4.9%, while the minimum was merely 0.1%. This demonstrates that our algorithm boasts excellent generalization capability and is compatible with multiple GNN architectures, enabling it to adapt to a wide range of tasks. To summarize, EDO-TF has achieved satisfactory performance in terms of correctness metrics and overwhelming advantages in fairness metrics. It demonstrates that our algorithm has conducted more in-depth exploration into fairness issues with introducing topological fairness. Consequently, it provides fairer results than other algorithms and alleviates fairness bias in graph neural networks at a more fundamental level. These results also verify the effectiveness of the fine-grained differentiable intervention mechanism for topological connections based on differential signs in EDO-TF. This mechanism can accurately identify and suppress inappropriate information in network structures that adversely affects fairness, thereby showcasing the great potential of EDO-TF as a strong generalization ability, interpretable, robust, and high-performance paradigm for achieving topological fairness.

To further visualize the debiasing effect of EDO-TF, we employed t-SNE to project the node embeddings from the Recidivism dataset into a 2-D space, as shown in Figure. 6 with discussion in Appendix D.

4.3 Ablation Studies

In order to gain a deeper understanding of how the individual loss terms influence the model's predictions, we performed an ablation study by separately removing the fairness loss and the graph regularization loss from the overall loss function. These two variants are denoted as EDO-TF\F and EDO-TF\G, respectively. It is important to note that the accuracy loss term acts on the pretrained model, meaning that its removal would likely induce uncontrollable fluctuations in the model's performance. Therefore, further discussion on it is omitted. The ablation results of EDO-TF are shown in Fig. 2, with GCN and GIN as the backbone. We can observe the following performance. (1) The performance of the EDO-TF method in the three metrics using the two backbones is approximate for the eight datasets. This once again demonstrates the universal applicability of EDO-TF for GNNs. (2) In the term of the ACC metric, the difference between EDO-TF and the other two methods is relatively small (approximately 1% on average). It means that removing two loss terms has a relatively small impact on accuracy. (3) A comparison between EDO-TF and EDO-TF\F reveals that the introduction of topological fairness leads to a significant changes in Δ_{SP} and Δ_{EO}. For example, on the Rec dataset with GCN as the backbone, the Δ_{SP} value increased from 5.31% to 7.55% (a 42% rise), which is the smallest increase across the eight datasets. (4) A comparison between EDO-TF and EDO-TF\G indicates that there are certain changes in Δ_{SP} and Δ_{EO}, but the magnitude of



(a) EDO-TF (b) EDO-TF\F (c) EDO-TF\G

(d) EDO-TF (e) EDO-TF\F (f) EDO-TF\G

(g) EDO-TF (h) EDO-TF\F (i) EDO-TF\G

Figure 2: Ablation results of loss terms in EDO-TF. EDO-TF\F denotes the removal of fairness loss, and EDO-TF\G denotes the removal of graph regularization loss.

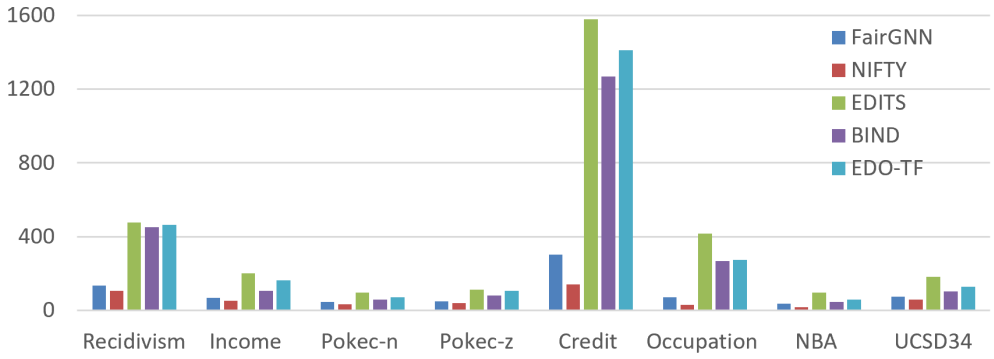
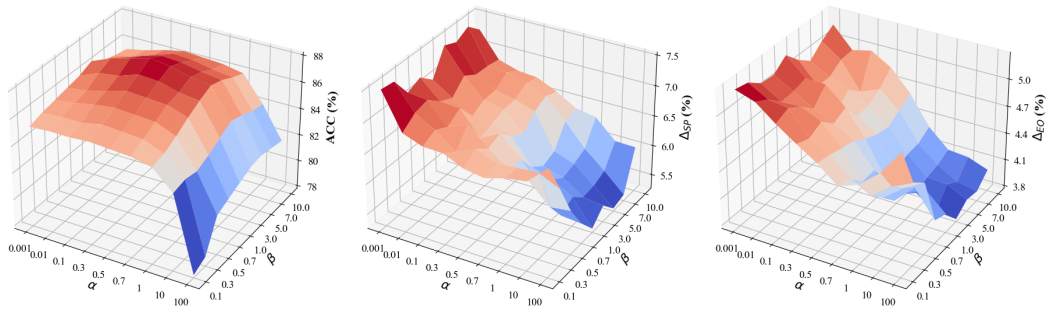


Figure 3: Comparison of training time costs for EDO-TF and baselines(in seconds).

change is relatively small. Comprehensive analysis of the experimental results leads to the following conclusions:(1) Removing the fairness loss expands biases, which confirms the effectiveness of the fairness loss in debiasing.(2) Removing the graph regularization loss affects accuracy but slightly improves fairness, indicating that the introduction of graph regularization has certain value for enhancing accuracy performance. In summary, considering these loss terms simultaneously enables a better balance between fairness and utility.

4.4 Computational Efficiency Analysis

In order to compare the computational efficiency of the algorithms, we conducted comparative experiments with results shown in Figure 3. For convenience, only the results using GCN as the backbone are listed because GIN has similar performance. Two key observations can be drawn from the results.(1) EDO-TF did not achieve the shortest running time and ranked fourth among the five algorithms. This indicates that there is still room for further improvement in the computational complexity of the algorithm. (2) The average running time of EDO-TF is 4, although its computational efficiency is not the highest, the increase in running time is only several times of the fastest algorithm. Overall, EDO-TF is not the most effective among all methods, but its time complexity is still



(a) ACC (b) Δ_{SP} (c) Δ_{EO}

Figure 4: Hyper-parameter analysis on Recidivism.

within an acceptable range. The main reason for the high computational complexity of EDO-TF is to calculate the inverse of the Hessian matrix for each A_{ij}. Therefore, improving the efficiency of matrix inversion and avoiding solving for each A_{ij} separately is an important direction for further improving computational efficiency.

4.5 Hyper-parameter Analysis

We conducted hyperparameter analysis on two parameters, α and β. Due to similar observations across different datasets, We specifically present the parameter sensitivity results for the Recidivism and Income datasets. Using GCN as the backbone, we vary α and β within the ranges of $\{0.001, 0.01, 0.1, 0.3, 0.5, 0.7, 1, 10, 100\}$ and $\{0.1, 0.3, 0.5, 0.7, 1, 3, 5, 7, 10\}$, respectively. As shown in Figure 4, the following observations can be made: (1) the overall performance of EDO-TF remains consistent despite the wide range of variations in α and β. Specifically, EDO-TF maintains stable utility and fairness when α varies from 0.001 to 10 and β from 0.5 to 10. (2) The best ACC performance is observed at $\alpha = 0.3$ and $\beta = 3$, while optimal Δ_{SP} and Δ_{EO} are achieved at $\alpha = 100$ with $\beta = 3$ and $\beta = 5$, respectively. (3) When α is greater than or equal to 10, although EDO-TF continues to improve fairness, the utility performance decreases significantly. When β is less than or equal to 0.5, both fairness and utility performance sharply decline. This emphasizes the importance of selecting appropriate values for α and β to optimize the fairness-utility trade-off, and highlights the cruciality of graph regularization in preventing unpredictable model behavior. Overall, EDO-TF demonstrates stability with respect to the broad variation in α and β.

5 Conclusion

In this paper, we revisited topological bias in GNNs from a fine-grained perspective, with a particular focus on how different types of per-edge influence fairness. Specifically, we first introduced the concept of Topological Fairness, which categorized neighborhood information into appropriate or inappropriate information, treating the latter as a confounding factor that introduced bias. Subsequently, to effectively differentiate between these two types of information and intervene in the inappropriate ones, we proposed a novel method called EDO-TF. By evaluating the sign of the derivative, we can effectively determine whether the information is helpful or harmful. Minimizing this value globally allowed us to mitigate topological bias and identify the optimal fairness graph. Our experiments on eight widely-used datasets demonstrated an exceptional fairness-utility trade-off, showcasing the effectiveness and competitiveness of our method. In summary, the definition of Topological Fairness not only laid the foundation for future research but also had the potential to open up the scope of research in this area.

6 AI USE STATEMENT

In this work, we used generative AI tools for improving the clarity, grammar, and readability of the manuscript. We have not used generative AI tools for generating the main scientific contributions, designing the core methodology, conducting experiments, analyzing experimental results, or drawing scientific conclusions, and the generation of research data, experimental results, and claims without human verification are not applicable to this work.

7 ETHICS STATEMENT

This work studies fairness in graph neural networks using eight previously released graph datasets: Recidivism, Income, Pokec-n, Pokec-z, Credit, Occupation, NBA, and UCSD34. We do not collect new data, conduct experiments involving human participants, or make decisions about identifiable individuals. The experiments are conducted solely for methodological evaluation.

8 REPRODUCIBILITY STATEMENT

The paper provides the mathematical formulation of Topological Fairness and Graph Fairness Differentiation, the corresponding objective functions, the complexity analysis, dataset statistics in Table 1, and the experimental settings in Section 4.

REFERENCES

C. Agarwal, H. Lakkaraju, and M. Zitnik. Towards a unified framework for fair and stable graph representation learning. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, Proceedings of Machine Learning Research, pp. 2114–2124, 2021.

H. Chen, S. Si, Y. Li, C. Chelba, S. Kumar, D. Boning, and C. Hsieh. Multi-stage influence function. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 1068, Vancouver, BC, Canada, 2020.

M. Chen, IW. Tsang, M. Tan, and TJ. Cham. A unified feature selection framework for graph embedding on high dimensional data. *IEEE Transactions on Knowledge and Data Engineering*, 27(6):1465–1477, 2015.

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

Y. Dong, N. Liu, B. Jalaian, and J. Li. Edits: Modeling and mitigating data bias for graph neural networks. In *Proceedings of the ACM Web Conference 2022*, pp. 1259–1269, Virtual Event, Lyon, France, 2022a.

Y. Dong, S. Wang, Y. Wang, T. Derr, and J. Li. On structural explanation of bias in graph neural networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 316–326, Washington, DC, USA, 2022b.

Y. Dong, J. Ma, S. Wang, C. Chen, and J. Li. Fairness in graph mining: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10583–10602, 2023a.

Y. Dong, J. Ma, S. Wang, C. Chen, and J. Li. Fairness in graph mining: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10583–10602, 2023b.

Y. Dong, S. Wang, J. Ma, N. Liu, and J. Li. Interpreting unfairness in graph neural networks via training node attribution. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, volume 37, pp. 7441–7449, Washington, DC, USA, 2023c. doi: 10.1609/aaai.v37i6.25905.

D. Dua and C. Graff. Uci machine learning repository. UCI Machine Learning Repository, 2017.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, Cambridge, MA, USA, 2012.

X. Feng, J. Liu, H. Nie, Y. Fu, L. Wan, and X. Kong. Random walks: A review of algorithms and applications. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 4(2):95–107, 2020.

H. Fenyu, Y. Zhu, S. Wu, W. Huang, L. Wang, and T. Tan. Graphair: Graph representation learning with neighborhood aggregation and interaction. *Pattern Recognition*, 112:107745, 2021.

W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 1025–1035, Long Beach, CA, USA, 2017.

K. Jordan and T. Freiburger. The effect of race/ethnicity on sentencing: Examining sentence type, jail length, and prison length. *Journal of Ethnicity in Criminal Justice*, 13:1–18, 2014.

W. Ju, S. Yi, Y. Wang, Z. Xiao, Z. Mao, and H. Li. A survey of graph neural networks in real world: Imbalance, noise, privacy and ood challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3036–3055, 2025.

M. Kearns, A. Roth, and S. Sharifi-Malvajerdi. Average individual fairness: Algorithms, generalization and experiments. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, 2019.

P. Koh and P. Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1885–1894, Sydney, NSW, Australia, 2017.

M. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4069–4079, Long Beach, CA, USA, 2017.

S. Li, X. Wang, A. Zhang, Y. Wu, X. He, and T. Chua. Let invariant rationale discovery inspire graph contrastive learning. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 13052–13065, Baltimore, MD, USA, 2022.

J. Martens. Deep learning via hessian-free optimization. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 735–742, Haifa, Israel, 2010.

J. Pearl. *Causality*. Cambridge University Press, London, U.K., 2009.

Y. Pei, N. Chakraborty, and K. Sycara. Nonnegative matrix tri-factorization with graph regularization for community detection in social networks. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pp. 2083–2089, Buenos Aires, Argentina, 2015.

X. Qian, Z. Guo, J. Li, H. Mao, B. Li, S. Wang, and Y. Ma. Addressing shortcomings in fair graph learning datasets: Towards a new benchmark. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5602–5612, Barcelona, Spain, 2024.

I. Spinelli, S. Scardapane, A. Hussain, and A. Uncini. Fairdrop: Biased edge dropout for enhancing fairness in graph representation learning. *IEEE Transactions on Artificial Intelligence*, 3(3):344–354, 2022a.

I. Spinelli, S. Scardapane, A. Hussain, and A. Uncini. Fairdrop: Biased edge dropout for enhancing fairness in graph representation learning. *IEEE Transactions on Artificial Intelligence*, 3(3):344–354, 2022b.

L. Takac and M. Zabovsky. Data analysis in public social networks. In *International Scientific Conference and International Workshop*, Lomza, Poland, 2012.

A. L. Traud, P. J. Mucha, and M. A. Porter. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180, 2012.

P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In *The Sixth International Conference on Learning Representations*, Vancouver, BC, Canada, 2018.

R. Wang, X. Wang, C. Shi, and L. Song. Uncovering the structural fairness in graph contrastive learning. New Orleans, LA, USA, 2024.

X. Wang and M. Zhang. How powerful are spectral graph neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 23341–23362, Baltimore, MD, USA, 2022.

Y. Wu, X. Wang, A. Zhang, X. He, and T. Chua. Discovering invariant rationales for graph neural networks. In *The Tenth International Conference on Learning Representations*, Vienna, Austria, 2022.

Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2021.

C. Yang, J. Liu, Y. Yan, and C. Shi. Fairsin: Achieving fairness in graph neural networks through sensitive information neutralization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9241–9249, Philadelphia, PA, USA, 2024.

I. Yeh and C. Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.

M. B. Zafar, I. Valera, R. M. Gomez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1171–1180, Perth, Australia, 2017.

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 325–333, Atlanta, GA, USA, 2013.

H. Zhang, G. Lu, M. Zhan, and B. Zhang. Semi-supervised classification of graph convolutional networks with laplacian rank constraints. *Neural Processing Letters*, 54(4):2645–2656, 2022.

A Appendix

A.1 Appendix A

Because PDD can be defined as the Wasserstein-1 distance between the probability distribution, we can give the calculation formula for SP and EO as follows,

$$\Gamma_{SP} = Wasserstein_1(P_{\hat{y}}^{(\mathcal{S}=0)}, P_{\hat{y}}^{(\mathcal{S}=1)}), \tag{12}$$

and

$$\Gamma_{EO} = Wasserstein_1(P_{\hat{y}}^{(\mathcal{S}=0,Y=1)}, P_{\hat{y}}^{(\mathcal{S}=1,Y=1)}). \tag{13}$$

where $P_{\hat{y}}^{(\mathcal{S}=0)} = P(\hat{y}=1|S=0)$, $P_{\hat{y}}^{(\mathcal{S}=1)} = P(\hat{y}=1|S=1)$, $P_{\hat{y}}^{(\mathcal{S}=0,Y=1)} = P(\hat{y}=1|S=0,Y=1)$, and $P_{\hat{y}}^{(\mathcal{S}=1,Y=1)} = P(\hat{y}=1|S=1,Y=1)$.

A.2 Appendix B

We provide a theoretical analysis of our definition from a causal perspective using SCM to model the causal relationships among six variables. With these assumptions listed in subsection 3.2, the SCM can be depicted in Figure. 5. Specifically, we focus on the causal relationships among six variables: node features X, appropriate information $\overline{N}$, inappropriate information $\tilde{N}$, node embedding E, biased prediction $\tilde{Y}$ and exogenous variable U. We employ SCM to model the variables, where nodes represent abstract variables and directed edges denote the causal relationship between

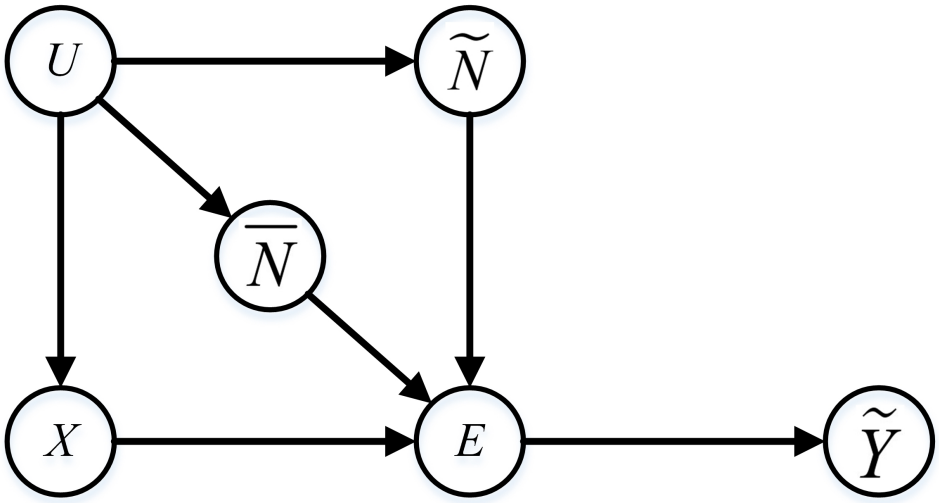


Figure 5: SCM for the information aggregation process.

these variables. Our hypothetical SCM is based on the following assumptions: (1) U represents all unobserved latent factors, and no variables outside the model influence the causal relationships. (2) The statistical independencies and dependencies among the variables in the causal graph are fully determined by the structure of the graph. (3) The causal relationships in the model are identifiable, meaning all causal paths and effects can be inferred from the observed data. Our analysis centers on three causal paths:

- $\tilde{N} \rightarrow E \rightarrow \tilde{Y}$. Direct path: Inappropriate neighborhood information $\tilde{N}$ directly influences node embedding E, which in turn leads to the biased prediction $\tilde{Y}$.
- $\tilde{N} \leftarrow U \rightarrow X \rightarrow E \rightarrow \tilde{Y}$ and $\tilde{N} \leftarrow U \rightarrow \overline{N} \rightarrow E \rightarrow \tilde{Y}$. Indirect path: Unobserved factors U can indirectly cause biased prediction results.

A.3 Appendix C

To derive the relationship between $\hat{W}$ and A, we start from the first-order optimality condition for $\hat{W}$ as follows,

$$\frac{\partial L_V(X, A, \hat{W})}{\partial W} = 0. \tag{14}$$

Since $\hat{W}$ implicitly depends on A, we use implicit differentiation to calculate $\frac{\partial \hat{W}}{\partial A}$. Differentiate both side of (8) with respect to A, we obtain the equation as follows,

$$\frac{\partial}{\partial A}\left(\frac{\partial L_V(X, A, \hat{W})}{\partial W}\right) = \frac{\partial^2 L_V(X, A, \hat{W})}{\partial W \partial A} + \frac{\partial^2 L_V(X, A, \hat{W})}{\partial W^2} \cdot \frac{\partial \hat{W}}{\partial A} = 0. \tag{15}$$

Solving for $\frac{\partial \hat{W}}{\partial A}$, we can obtain the first-order partial derivative as follows,

$$\frac{\partial \hat{W}}{\partial A} = -\left(\frac{\partial^2 L_V(X, A, \hat{W})}{\partial W^2}\right)^{-1} \cdot \frac{\partial^2 L_V(X, A, \hat{W})}{\partial W \partial A}. \tag{16}$$

For each element A_{ij} in matrix A, we estimate the partial derivatives using the central difference formula as follows,

$$\frac{\partial^2 L_V(X, A, \hat{W})}{\partial W \partial A_{ij}} \approx \frac{\frac{\partial L_V}{\partial W}\left(X, A_{ij} + \epsilon \cdot E_{ij}, \hat{W}\right) - \frac{\partial L_V}{\partial W}\left(X, A_{ij} - \epsilon \cdot E_{ij}, \hat{W}\right)}{2\epsilon}. \tag{17}$$

Algorithm 1 Computation of $\nabla_{A_{ij}}\Gamma$

Input: the processed graph data $\mathcal{G}$, the trained GNN model $f_{\hat{W}}$, the set of training nodes $\mathcal{V}$
Output: $\nabla_{A_{ij}}\Gamma$

1: Initialize $\nabla_{A_{ij}}\Gamma = \varnothing$
2: Compute $\frac{\partial\Gamma}{\partial W}$ based on $f_{\hat{W}}$
3: Compute $\frac{\partial L_{\mathcal{V}}(X,A,\hat{W})}{\partial W}$ based on $f_{\hat{W}}$
4: Estimate the inverse of the Hessian matrix,
5: $\left(\frac{\partial^2 L_{\mathcal{V}}(X,A,\hat{W})}{\partial W^2}\right)^{-1}$
6: **for** each element A_{ij} in matrix A **do**
7: Compute $\frac{\partial^2 L_{\mathcal{V}}(X,A,\hat{W})}{\partial W \partial A_{ij}}$
8: **end for**
9: Compute $\nabla_{A_{ij}}\Gamma$
10: Append the computed value to $\nabla_{A_{ij}}\Gamma$
11: **return** $\nabla_{A_{ij}}\Gamma$

Finally, we can obtain the first-order partial derivative as follows,:

$$\frac{\partial \hat{W}}{\partial A_{ij}} = -(\frac{\partial^2 L_V(X,A,\hat{W})}{\partial W^2})^{-1} \cdot \frac{\frac{\partial L_V}{\partial W}(X, A_{ij}+\epsilon\cdot E_{ij}, \hat{W}) - \frac{\partial L_V}{\partial W}(X, A_{ij}-\epsilon\cdot E_{ij}, \hat{W})}{2\epsilon}, \tag{18}$$

where E_{ij} is a matrix of the same type as W, with all elements set to 0 except that the element at position (i, j) is set to 1, ϵ is a sufficiently small number which means only minor increases and decreases of ϵ are applied to the elements in row i and column j of W. Consequently, $\nabla_{A_{ij}}\Gamma$ can be calculated as follows,

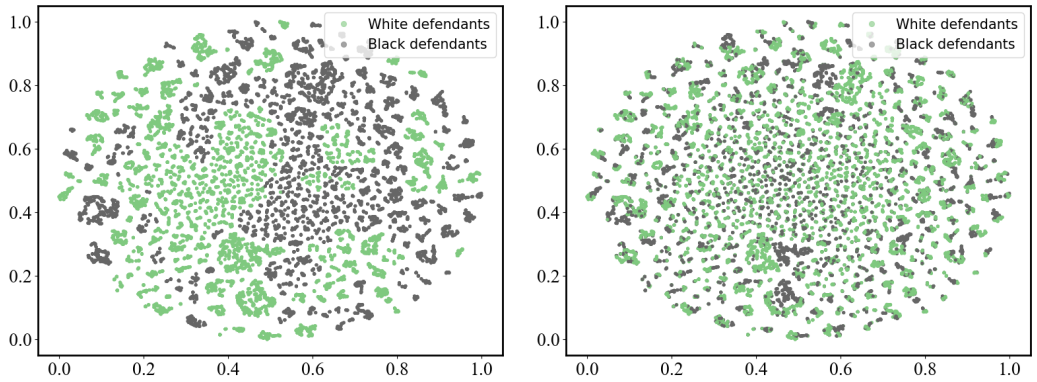
$$\begin{aligned}\nabla_{A_{ij}}\Gamma = (\frac{\partial\Gamma}{\partial W})^{\intercal} \cdot \frac{\partial \hat{W}}{\partial A_{ij}} = -(\frac{\partial\Gamma}{\partial W})^{\intercal} \cdot (\frac{\partial^2 L_{\mathcal{V}}(X,A,\hat{W})}{\partial W^2})^{-1} \\ \cdot \frac{\frac{\partial L_{\mathcal{V}}}{\partial W}(X, A_{ij}+\epsilon\cdot E_{ij}, \hat{W}) - \frac{\partial L_{\mathcal{V}}}{\partial W}(X, A_{ij}-\epsilon\cdot E_{ij}, \hat{W})}{2\epsilon}.\end{aligned} \tag{19}$$

A.4 Appendix D

We employed eight graph datasets in our experiments, Recidivism, Income, Pokec-n, Pokec-z, Credit, Occupation, NBA, UCSD34. These datasets have been extensively utilized in previous studies on graph fairness learning, covering a diverse range of domains such as finance, criminal justice, and social network. The Recidivism dataset (Jordan & Freiburger, 2014) designates 'race' as the sensitive attribute, with the task of classifying whether a defendant is granted bail. The income dataset (Dua & Graff, 2017) also identifies 'race' as the sensitive attribute with the task of predicting whether an individual's annual income exceeds $50,000. The datasets of Pokec-n and Pokec-z, collected from the popular Slovakian social network Pokec (Takac & Zabovsky, 2012), list 'locating region' as the sensitive attribute, which refers to the geographical area where users are located, and the associated task is to classify the users' fields of employment. The dataset of Credit dataset (Yeh & Lien, 2009) identifies 'age' as the sensitive attribute, with the task of predicting whether a user will default on their credit card payment. The dataset of Occupation (Qian et al., 2024) also designates 'gender' as the sensitive attribute, with the task of predicting a user's field of work without gender-based bias. The NBA (Dai & Wang, 2021) identifies 'nationality' (binarized as U.S. vs. overseas players) as the sensitive attribute, with the task of predicting whether a player's salary exceeds the median. The UCSD34 dataset (Traud et al., 2012) identifies 'gender' as the sensitive attribute, with the task of predicting whether a user belongs to a specific major. The details are listed in Table2 .

Table 2: Description of Dataset Details

Dataset	Recidivism	Income	Pokee-n	Pokee-z	Credit	Occupation	NBA	UCSD34
Nodes	18,876	14,821	6,185	7,659	30,000	6,951	403	4,132
Edges	321,308	100,483	21,844	29,476	1,436,858	44,166	10,621	108,383
Attributes	18	14	59	768	18	59	96	7
Avg. degree	34.0	13.6	7.06	7.70	34.0	12.7	52.7	52.5
Sens.	Race	Race	Region	Region	Age	Gender	Nationality	Gender
Label	Bail decision	Income level	Working field	Working field	Future default	Working field	Salary	Major



(a) GCN without EDO-TF (b) GCN with EDO-TF

Figure 6: Visualization of embeddings learned on Recidivism. Node color represents the sensitive attributes of nodes.

To visualize the debiasing effect of EDO-TF, we employed t-SNE to project the node embeddings from the Recidivism dataset into a 2-D space, as shown in Figure. 6. The left subplot displays embeddings learned by GCN without any debiasing, while the right noe demonstrates the debiased embeddings obtained by applying the EDO-TF method. Node colors indicate race categories. Without debiasing, embeddings are distinctly clustered by sensitive attribute. In contrast, with EDO-TF, embeddings with different sensitive attribute are blended together, demonstrating that our method effectively achieves model debiasing.