FnRGNN: Distribution-aware Fairness in Graph Neural Network

SoYoung Park
Chungnam National University
Daejeon, Republic of Korea
sypark1452@o.cnu.ac.kr

Sungsu Lim*
Chungnam National University
Daejeon, Republic of Korea
sungsu@cnu.ac.kr

Abstract

Graph Neural Networks (GNNs) excel at learning from structured data, yet fairness in regression tasks remains underexplored. Existing approaches mainly target classification and representation-level debiasing, which cannot fully address the continuous nature of node-level regression. We propose FnRGNN, a fairness-aware in-processing framework for GNN-based node regression that applies interventions at three levels: (i) structure-level edge reweighting, (ii) representation-level alignment via MMD, and (iii) prediction-level normalization through Sinkhorn-based distribution matching. This multi-level strategy ensures robust fairness under complex graph topologies. Experiments on four real-world datasets demonstrate that FnRGNN reduces group disparities without sacrificing performance. Code is available at https://github.com/sybeam27/FnRGNN.

CCS Concepts

• **Computing methodologies** → **Neural networks**; **Supervised learning by regression**.

Keywords

Graph Neural Networks, Node Regression, Trustworthy AI, Fairness

ACM Reference Format:
SoYoung Park and Sungsu Lim. 2025. FnRGNN: Distribution-aware Fairness in Graph Neural Network. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3746252.3760796

1 Introduction

Graph Neural Networks (GNNs) are increasingly deployed in sensitive domains such as social networks [29], recommendation systems [11], and healthcare [35]. While fairness issues have been widely reported in general machine learning systems—such as gender-biased job recommendations [28], ad delivery favoring men [19], and racially skewed online exposure [31]—similar risks exist in GNN-based models. GNNs, in particular, tend to amplify structural biases present in graphs through their message-passing mechanisms, making them especially prone to reinforcing disparities related to sensitive attributes like race and gender [9, 15, 18, 22]. Since such attributes can be implicitly encoded during training, ensuring fairness requires interventions at the algorithmic level—beyond input pre-processing or post-hoc correction [23, 26].

Despite growing interest in in-processing methods for fair GNN training, most existing approaches address bias at only one stage, most often the representation level, by removing sensitive information from node embeddings [1, 5, 38]. However, this overlooks structural bias propagated through message passing. Structure-level debiasing methods have been proposed [15, 25], but they are largely tailored for classification and struggle to generalize to regression. Node-level regression plays a critical role in applications such as disease risk estimation [20], recidivism prediction [37], and credit risk modeling [2], yet fairness in graph-based regression remains underexplored. Unlike classification or link prediction, where fairness is typically defined over discrete labels, regression involves continuous targets, making accuracy alone insufficient. Ensuring equitable outcomes requires evaluating and enforcing distributional parity across sensitive groups [4, 36]. Even small prediction disparities can lead to significant unfairness in high-stakes domains, highlighting the need for a unified, distribution-aware framework that mitigates bias across structural, representational, and predictive stages.

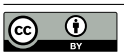
To address this gap, we propose **FnRGNN**, a fairness-aware in-processing framework for node-level regression. Standard GNN regressors often produce disparate output distributions across sensitive groups. FnRGNN mitigates these disparities through a distribution-aware design that enforces fairness throughout training. It integrates three components: (i) *Structure-level* edge reweighting to suppress bias propagation during message passing [22, 38], (ii) *Representation-level* alignment to reduce differences in the embedding space [12], and (iii) *Prediction-level* normalization to match output distributions across groups via full-shape alignment [10, 21]. This unified, multi-level framework extends fair GNN research beyond classification, tackling fairness in continuous prediction where accuracy alone is insufficient. Our contributions are as follows:

(1) **Fairness in Graph Node Regression:** We focus on node-level regression, an important yet underexplored setting in fair GNN, and address its unique fairness challenges from predicting continuous outcomes across diverse groups.
(2) **Multi-level Distribution-aware Framework:** We introduce FnRGNN, a distribution-aware framework operating across structure, representation, and prediction levels.
(3) **Empirical Validation:** Experiments on four real-world datasets demonstrate that FnRGNN effectively reduces group disparities while preserving high predictive accuracy.

2 Related Work

GNNs tend to amplify biases in graph-structured data [9, 18], making fairness a critical concern in research. Recent studies focus on in-processing methods that mitigate bias during training, categorized into structure, representation, and prediction levels [4].

*Corresponding author.



CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3760796

(i) *Structure-level* methods mitigate bias by modifying graph topology or aggregation mechanisms. These include adjusting message passing schemes [15], editing the graph structure [7, 8, 25], reweighting or resampling [22, 24], and applying debiasing techniques [3]. Other approaches leverage counterfactual reasoning [26] or multi-level models [14] to account for structural sources of bias.

(ii) *Representation-level* methods focus on learning node embeddings that are invariant to sensitive attributes. Common strategies include adversarial training [5], distribution alignment [38], orthogonalization [27], and normalization techniques [1].

(iii) *Prediction-level* methods directly target output fairness using ranking-based objectives [6], fairness-regularized losses [16], or output distribution alignment via regularization [12].

While prior work has advanced fairness in GNNs, most methods target classification or link prediction and offer limited support for node-level regression, where prediction targets are continuous and biases arise more subtly. Fairness in regression tasks requires more than embedding-level invariance; it must account for structural bias and align output distributions across groups [36], calling for a holistic approach across structure, representation, and prediction.

3 Preliminaries

Graph Neural Networks. We consider a graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges. Each node $i \in \mathcal{V}$ is associated with a feature vector $\mathbf{x}_i \in \mathbb{R}^d$, and the complete feature matrix is denoted by $\mathbf{X} \in \mathbb{R}^{n \times d}$. Each node also has a continuous regression target $y_i \in \mathbb{R}$, forming the label vector $\mathbf{y} \in \mathbb{R}^n$. A GNN learns node representations by aggregating neighborhood information through stacked message-passing layers. At each layer l, the representation of node i is updated as

$$\mathbf{h}_i^{(l)} = \sigma\left(W^{(l)} \cdot \mathrm{AGG}\left(\{\mathbf{h}_u^{(l-1)} : u \in \mathcal{N}(i)\}\right)\right), \tag{1}$$

where $\mathcal{N}(i)$ denotes the set of neighbors of node i, AGG is an aggregation function (e.g., mean, sum, or attention), and σ is a non-linear activation. The final node embedding $\mathbf{h}_i^{(L)}$ is then passed to a regression head to produce the prediction $\hat{y}_i \in \mathbb{R}$.

Fairness in Node Regression. We define fairness as consistency of predictions across groups defined by a binary sensitive attribute $s_i \in \{0, 1\}$. We adopt two criteria: (1) *mean parity* — small difference in group-wise expected predictions, $|\mathbb{E}[\hat{y}_i \mid s_i = 0] - \mathbb{E}[\hat{y}_i \mid s_i = 1]| \approx 0$; (2) *distributional parity* — small divergence between group-wise predictive distributions, $\mathbb{D}(P_{\hat{y}_i|s_i=0}, P_{\hat{y}_i|s_i=1}) \approx 0$.

4 Methodology

We propose FnRGNN, a fairness-aware graph neural network for node-level regression. Its goal is to minimize prediction error while promoting fairness across sensitive groups, as defined in Sec. 3. To achieve this, FnRGNN incorporates fairness interventions at three levels: (i) *structure* (Sec. 4.1), (ii) *representation* (Sec. 4.2), and (iii) *prediction* (Sec. 4.3), as illustrated in Fig. 1.

4.1 Edge Reweighting for Fair Structure

To prevent the amplification of bias through message passing [13, 17], we propose a hybrid edge weighting that jointly considers node feature similarity and group demographic dissimilarity. For each edge $(i, j) \in \mathcal{E}$, we define the adjusted edge weight α_{ij} as:

$$\alpha_{ij} = \mathrm{sim}(\mathbf{x}_i, \mathbf{x}_j) \cdot \exp\left(-\gamma \cdot \mathbb{I}[s_i \neq s_j]\right), \tag{2}$$

where $\mathrm{sim}(\cdot, \cdot)$ captures feature-level closeness between node pairs, while the indicator function $\mathbb{I}[\cdot]$ checks whether their sensitive attributes differ. An exponential penalty controlled by hyperparameter γ reduces the weights of cross-group edges, mitigating bias without hard pruning. This soft reweighting preserves graph connectivity and is directly integrated into standard GNN propagation for fair representation learning [22, 38]. At each layer l, node embeddings are updated as:

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} \cdot \mathbf{h}_j^{(l)} \mathbf{W}^{(l)}\right), \tag{3}$$

where $\mathbf{W}^{(l)}$ is a learnable transformation matrix and $\sigma(\cdot)$ is a non-linear activation function. By integrating fairness-aware edge weights directly into the message passing process, our model effectively modulates structural bias during representation learning.

4.2 Group Alignment for Fair Representation

To mitigate representational bias, we introduce a kernel-based regularization using Maximum Mean Discrepancy (MMD) [12]. Given a binary sensitive attribute s_i, we define group $\mathcal{G}_a = \{i \in \mathcal{V} \mid s_i = a\}$ for $a \in \{0, 1\}$. MMD encourages the embedding distributions of these groups to be statistically aligned—not only in terms of the mean but also higher-order moments—thus promoting fairer representations. The MMD loss is computed as:

$$\begin{aligned}\mathcal{L}_{\mathrm{MMD}} = \mathbb{E}_{\mathbf{x},\mathbf{x}' \in \mathcal{G}_0}[k(\mathbf{x}, \mathbf{x}')] + \mathbb{E}_{\mathbf{y},\mathbf{y}' \in \mathcal{G}_1}[k(\mathbf{y}, \mathbf{y}')] \\ - 2\mathbb{E}_{\mathbf{x} \in \mathcal{G}_0, \mathbf{y} \in \mathcal{G}_1}[k(\mathbf{x}, \mathbf{y})],\end{aligned} \tag{4}$$

where $k(\cdot, \cdot)$ is a positive-definite kernel function, typically instantiated as the Gaussian RBF kernel [30], defined by $k(\mathbf{X}, \mathbf{y}) = \exp\left(-\|\mathbf{X} - \mathbf{y}\|^2/(2\sigma^2)\right)$, to measure similarity. Minimizing $\mathcal{L}_{\mathrm{MMD}}$ encourages the model to learn group-invariant representations by reducing distributional discrepancies in embedding space. This regularization serves as a soft fairness constraint and integrates seamlessly with the main regression objective.

4.3 Distribution Matching for Fair Prediction

To ensure output-level fairness, we apply a dual regularization strategy that aligns global distributions and preserves local statistical consistency across sensitive groups. Unlike standard moment matching, our approach accounts for higher-order differences in the output space. To address this, we combine moment-based matching with Sinkhorn divergence [10], a smoothed version of optimal transport (OT) [34] that measures the geometric cost of transforming one distribution into another under entropic regularization. Specifically, the output-level fairness loss is defined as $\mathcal{L}_{\mathrm{dist}} = \mathcal{L}_{\mathrm{sinkhorn}} + \mathcal{L}_{\mathrm{moment}}$, where the Sinkhorn term is given by:

$$\begin{aligned}\mathcal{L}_{\mathrm{sinkhorn}} = OT_\varepsilon(\hat{\mathbf{y}}_{\mathcal{G}_0}, \hat{\mathbf{y}}_{\mathcal{G}_1}) \\ - \tfrac{1}{2}\left[OT_\varepsilon(\hat{\mathbf{y}}_{\mathcal{G}_0}, \hat{\mathbf{y}}_{\mathcal{G}_0}) + OT_\varepsilon(\hat{\mathbf{y}}_{\mathcal{G}_1}, \hat{\mathbf{y}}_{\mathcal{G}_1})\right],\end{aligned} \tag{5}$$

with OT_ε denoting the entropic OT distance computed with regularization parameter ε. This formulation stabilizes training by penalizing over-concentration and sampling noise. In parallel, we

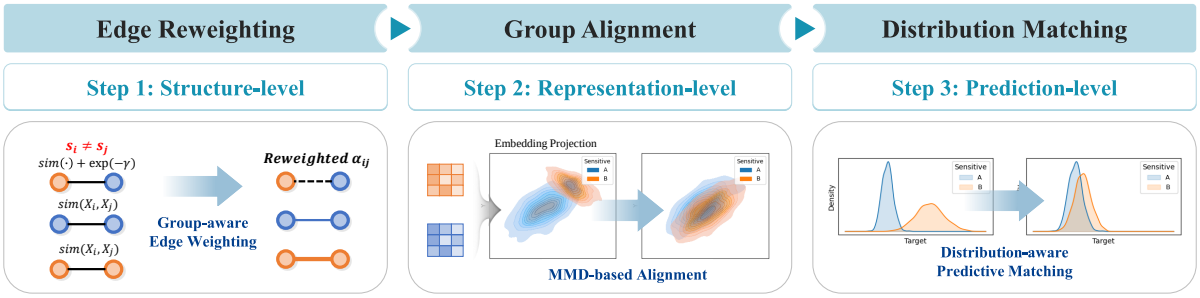


Figure 1: Overview of FnRGNN, which integrates fairness interventions at three levels: (i) structure, (ii) representation, and (iii) prediction. Each component addresses group-level disparities in node-level regression, as introduced in Sec. 4.

Algorithm 1: Training of FnRGNN

Input: Graph $\mathcal{G}$, features $\mathbf{X}$, target $\mathbf{y}$, sensitive attribute s
Output: Trained model parameters θ

Initialize model parameters θ ;
for *each epoch* **do**
  1. Structure-level:
  $\alpha_{ij} \leftarrow \text{sim}(\mathbf{x}_i, \mathbf{x}_j) e^{-\gamma \mathbb{I}[s_i \neq s_j]}$; $h^{(2)} \leftarrow \text{GCNLayer}(X, A_\alpha)$
  2. Representation-level: $\mathcal{L}_{\text{MMD}} = \text{MMD}(\mathbf{H}^{(2)}_{\mathcal{G}_0}, \mathbf{H}^{(2)}_{\mathcal{G}_1})$
  3. Prediction-level: $\mathcal{L}_{\text{dist}} = \mathcal{L}_{\text{sinkhorn}} + \mathcal{L}_{\text{moment}}$
  4. Optimize: $\mathcal{L}_{\text{total}} = \text{MSE} + \lambda_{\text{MMD}} \mathcal{L}_{\text{MMD}} + \lambda_{\text{dist}} \mathcal{L}_{\text{dist}}$

apply a moment-based regularizer [21] to enforce consistency in the first and second moments:

$$\mathcal{L}_{\text{moment}} = \left|\mathbb{E}[\hat{\mathbf{y}}_{\mathcal{G}_0}] - \mathbb{E}[\hat{\mathbf{y}}_{\mathcal{G}_1}]\right| + \left|\text{Var}[\hat{\mathbf{y}}_{\mathcal{G}_0}] - \text{Var}[\hat{\mathbf{y}}_{\mathcal{G}_1}]\right|. \quad (6)$$

This hybrid approach ensures robust fairness by aligning predictions and preserving group-level statistics.

4.4 Training Procedure

In Alg. 1, FnRGNN is trained in four stages. **Step 1** constructs a fairness-aware graph by reweighting edges with a similarity-based coefficient α_{ij} that penalizes cross-group connections. The resulting adjacency matrix is used in a GCN to generate hidden representations, followed by an MLP that outputs predictions $\hat{\mathbf{y}}$. **Step 2** promotes representation-level fairness by minimizing the MMD loss $\mathcal{L}_{\text{MMD}}$, aligning embeddings across sensitive groups. **Step 3** ensures prediction-level fairness via distributional alignment using $\mathcal{L}_{\text{dist}}$, which combines Sinkhorn distance and moment matching. **Step 4** defines the total loss $\mathcal{L}_{\text{total}}$ as the sum of prediction error and fairness terms. Model parameters θ are optimized via Adam, with fairness losses computed per mini-batch for stability.

5 Experiments

5.1 Experiment Settings

Datasets. We evaluate on four real-world graph datasets with diverse structures and sensitive attributes (Tab.1): **Pokec-z** and **Pokec-n**[32] are Slovak social network subsets for profile completion; Pokec-z shows mean-shift bias, while Pokec-n has near-identical gender distribution. **NBA**[5] contains basketball data for predicting minutes per game, with nationality as the sensitive attribute; despite group size imbalance, distributions are similar. **German**[1] is credit data with gender as the sensitive attribute; group distributions differ in shape and mean, indicating moderate bias.

Table 1: Dataset statistics and label distributions by sensitive group. The rightmost column shows target value differences across sensitive attributes.

Dataset	Group $\mathcal{G}$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	dim(X)	Dist.
Pokec-z	Region	67,796	1,303,712	276	
Pokec-n	Gender	66,569	1,100,663	265	
NBA	Country	403	19,357	95	
German	Gender	1,000	44,484	29	

Baselines and Implementation. We compare FnRGNN with representative fair GNN baselines, categorized into three fairness strategies levels: (i) structure — FMP [15], EDITS [7]; (ii) representation — FairGNN [5], GMMD [38]; and (iii) prediction — REDRESS [6], and GCN_{mean}, a GCN-based regressor that performs group-wise mean matching. All models were implemented in PyTorch Geometric, adapting regression by replacing classification loss with MSE. Training used Adam optimizer for 500 epochs with early stopping, and results were averaged over five runs. Unless noted, we used a 2-layer GCN [17] encoder (64 hidden units, learning rate 10^{-3}, weight decay 10^{-5}). FnRGNN extends this base with (i) MMD [12] alignment, (ii) Sinkhorn regularization [10], and (iii) edge reweighting. Hyperparameters ($\lambda_{\text{MMD}}, \lambda_{\text{dist}}, \gamma$) were tuned via NSGA-II in Optuna, with losses computed from 500 sampled nodes per group.

Evaluation Metrics. To evaluate fairness in regression, we adopt group-wise metrics suitable for continuous outputs, since classification-based notions (e.g., demographic parity) are not directly applicable [4]. Given a binary sensitive attribute $s_v \in \{0, 1\}$, we define groups $\mathcal{G}_0$ and $\mathcal{G}_1$ and compute: (i) Mean Gap (**MG**) = $\left|\bar{\mathbf{y}}_{\mathcal{G}_0} - \bar{\mathbf{y}}_{\mathcal{G}_1}\right|$, (ii) Variance Gap (**VG**) = $\left|\text{Var}(\hat{\mathbf{y}}_{\mathcal{G}_0}) - \text{Var}(\hat{\mathbf{y}}_{\mathcal{G}_1})\right|$, and (iii) Wasserstein Distance (**WD**) [33], which quantifies distributional divergence as $\inf_{\gamma \in \Pi} \mathbb{E}_{(i,j)\sim\gamma}[|\hat{y}_i - \hat{y}_j|]$, where γ denotes a

Table 2: Results on four datasets (average of five runs). MSE/MAE assess accuracy; MG, VG, and WD assess fairness. (i) Structure-level: FMP [15], EDITS [7]; (ii) Representation-level: FairGNN [5], GMMD [38]; (iii) Prediction-level: REDRESS [6], GCN$_{mean}$. EDITS and REDRESS are omitted for Pokec datasets due to GPU memory limits.

Metrics (↓)	Dataset	FMP	FairGNN	GMMD	GCN	FnRGNN	Dataset	FMP	EDITS	FairGNN	GMMD	REDRESS	GCN	FnRGNN
MSE	Pokec-z	0.3178	0.6921	**0.0342**	0.4333	0.0622	NBA	1.0001	746.28	1.1417	0.3496	0.9977	1.6414	**0.1495**
MAE		0.4986	0.7052	**0.1247**	0.5040	0.1829		0.8570	21.97	0.8177	0.3803	0.8555	0.8016	**0.3024**
MG		0.0269	0.0416	0.0747	**0.0084**	0.0171		0.0019	13.80	0.2204	0.1156	**0.0001**	0.5501	0.0415
VG		**0.0026**	0.0392	0.0150	0.1074	0.0046		0.0003	216.40	0.3509	0.2891	**0.0000**	4.3050	0.1876
WD		0.0063	0.0055	**0.0028**	0.0604	0.0055		**0.0038**	13.64	0.0185	0.1085	0.0106	0.5279	0.0070
MSE	Pokec-n	0.4857	0.5877	**0.0301**	0.4439	0.0511	German	1.0934	304.53	1.0141	1.7481	1.1700	1.0490	**0.6994**
MAE		0.6217	0.6427	**0.1140**	0.5040	0.1557		0.9511	13.05	0.8803	1.0308	0.8993	0.8550	**0.6995**
MG		0.0232	0.0064	0.0437	**0.0008**	0.0133		0.0007	3.2408	0.1643	0.1812	**0.0001**	0.0010	0.1421
VG		0.0012	0.0156	0.0242	0.0046	**0.0003**		0.0001	28.98	0.0012	0.3665	**0.0000**	0.0250	0.0730
WD		0.0143	0.0312	0.0107	0.0240	**0.0018**		0.0465	0.7483	0.1090	0.0881	0.0662	0.0574	**0.0217**

Table 3: Ablation results on four datasets: Case 1 — without edge reweighting; Case 2 — without group embedding alignment; Case 3 — with mean-only distribution matching.

Metric (↓)	Dataset	Vanilla	Case 1	Case 2	Case 3	Full
MSE	Pokec-z	0.4676	0.4646	0.0846	0.4333	**0.0622**
WD		0.0206	**0.0053**	0.0112	0.0604	0.0055
MSE	Pokec-n	0.4837	0.4417	0.0789	0.4439	**0.0511**
WD		0.0181	0.0123	0.0036	0.0240	**0.0018**
MSE	NBA	5.3053	0.5787	0.1695	0.1641	**0.1495**
WD		1.0585	0.0321	0.0561	0.5279	**0.0070**
MSE	German	1.4856	0.8654	0.8072	1.0490	**0.6994**
WD		0.0372	0.0284	0.0524	0.0574	**0.0217**

transport plan between group-wise prediction distributions. Together, MG and VG assess moment disparities, while WD captures full-shape differences in prediction distributions. For predictive accuracy, we report Mean Squared Error (**MSE**) = $\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$ and Mean Absolute Error (**MAE**) = $\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$.

5.2 Experiment Results

We compare FnRGNN with representative fair GNN baselines on four datasets with varying graph structures and demographic imbalances. As shown in Tab. 2, FnRGNN consistently achieves a favorable trade-off between prediction accuracy and fairness across all three evaluation levels. At the (i) *structure-level*, FnRGNN clearly outperforms FMP and EDITS, especially on real-world datasets such as NBA and German. While FMP and EDITS suffer from high MSE and fairness gaps, FnRGNN effectively leverages edge reweighting to enhance both accuracy and group parity. At the (ii) *representation-level*, GMMD and FairGNN partially reduce fairness gaps, but often at the cost of prediction accuracy. For example, GMMD achieves low MSE on Pokec but fails to minimize group disparities (e.g., MG, WD). In contrast, FnRGNN achieves both low error and consistently small group-wise gaps, suggesting stronger alignment in the latent space. At the (iii) **prediction-level**, FnRGNN significantly outperforms REDRESS and GCN$_{\text{mean}}$, especially on datasets like German and NBA. While REDRESS struggles with fairness, GCN$_{\text{mean}}$ underperforms in accuracy. FnRGNN achieves the lowest MSE and MAE while maintaining strong fairness scores across MG, VG, and WD. These results demonstrate that FnRGNN's multi-level design enables robust performance under distributional imbalance, jointly addressing structural, representational, and predictive biases.

5.3 Ablation Study

We conduct ablation studies on four regression datasets to evaluate the contribution of each component in FnRGNN. As shown in Tab. 3, the full model consistently achieves the best trade-off between MSE and WD. Case 1 occasionally lowers WD but results in higher MSE, suggesting that this module enhances training stability and overall predictive accuracy. Case 2 leads to a notable increase in MSE, especially on the Pokec datasets, emphasizing the importance of alignment in the embedding space for both fairness and accuracy. Case 3 increases WD in most datasets, underscoring the necessity of aligning full output distributions to mitigate bias effectively. Overall, these results confirm that the three components of FnRGNN are complementary and jointly contribute to fairness in regression.

6 Conclusion

We present FnRGNN, a fairness-aware GNN for node regression that applies multi-level interventions to mitigate bias. Our framework integrates edge reweighting for fair message passing, MMD-based representation alignment, and Sinkhorn-augmented distribution matching for prediction fairness. Experiments on real-world datasets show that each module improves fairness without compromising predictive accuracy. These results demonstrate the effectiveness of structured fairness regularization without the need for adversarial training or complex architectures. Future work will extend FnRGNN to multi-class and dynamic graph scenarios and investigate theoretical fairness guarantees under distribution shifts.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00214065) and by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-00155857, Artificial Intelligence Convergence Innovation Human Resources Development (Chungnam National University)).

Generative AI Usage Disclosure

This paper made limited and transparent use of generative AI tools—specifically ChatGPT, Grok, and Grammarly—for the following purposes: (i) minor language editing, including grammar, spelling, and clarity improvements; (ii) preliminary exploration of research directions; (iii) verification and clarification of technical concepts to support the author's understanding; and (iv) assistance with LaTeX formatting. No AI-generated text was used verbatim, and all core ideas, arguments, and interpretations are solely the author's own. AI tools were used strictly as assistive technologies and did not influence the conceptual development or analytical reasoning of the work.

References

[1] Chirag Agarwal, Himabindu Lakkaraju, and Marinka Zitnik. 2021. Towards a unified framework for fair and stable graph representation learning. In *UAI*, Vol. 161. AUAI Press, 2114–2124.

[2] Vicente Balmaseda, María Coronado, and Gonzalo de Cadenas-Santiago. 2023. Predicting systemic risk in financial systems using Deep Graph Learning. *ISWA* 19 (2023), 200240.

[3] Maarten Buyl and Tijl De Bie. 2020. Debayes: a bayesian method for debiasing network embeddings. In *ICML*, Vol. 119. PMLR, 1220–1229.

[4] April Chen, Ryan A Rossi, Namyong Park, Puja Trivedi, Yu Wang, Tong Yu, Sungchul Kim, Franck Dernoncourt, and Nesreen K Ahmed. 2024. Fairness-aware graph neural networks: A survey. *TKDD* 18, 6 (2024), 1–23.

[5] Enyan Dai and Suhang Wang. 2021. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *WSDM*. ACM, 680–688.

[6] Yushun Dong, Jian Kang, Hanghang Tong, and Jundong Li. 2021. Individual fairness for graph neural networks: A ranking based approach. In *KDD*. ACM, 300–310.

[7] Yushun Dong, Ninghao Liu, Brian Jalaian, and Jundong Li. 2022. Edits: Modeling and mitigating data bias for graph neural networks. In *WWW*. ACM, 1259–1269.

[8] Yushun Dong, Song Wang, Jing Ma, Ninghao Liu, and Jundong Li. 2023. Interpreting unfairness in graph neural networks via training node attribution. In *AAAI*. AAAI Press, 7441–7449.

[9] Yushun Dong, Song Wang, Yu Wang, Tyler Derr, and Jundong Li. 2022. On structural explanation of bias in graph neural networks. In *KDD*. ACM, 316–326.

[10] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. 2019. Interpolating between optimal transport and mmd using sinkhorn divergences. In *AISTATS*, Vol. 89. PMLR, 2681–2690.

[11] Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, et al. 2020. Fairness-aware explainable recommendation over knowledge graphs. In *SIGIR*. ACM, 69–78.

[12] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. 2006. A kernel method for the two-sample-problem. *NeurIPS* 19 (2006), 513–520.

[13] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *NeurIPS* 30 (2017), 1024–1034.

[14] Yuntian He, Saket Gurukar, and Srinivasan Parthasarathy. 2023. FairMILE: Towards an efficient framework for fair graph representation learning. In *EAAMO*. ACM, 1–10.

[15] Zhimeng Jiang, Xiaotian Han, Chao Fan, Zirui Liu, Na Zou, Ali Mostafavi, and Xia Hu. 2024. Chasing fairness in graphs: A gnn architecture perspective. In *AAAI*. AAAI Press, 21214–21222.

[16] Jian Kang, Yan Zhu, Yinglong Xia, Jiebo Luo, and Hanghang Tong. 2022. Rawlsgcn: Towards rawlsian difference principle on graph convolutional network. In *WWW*. ACM, 1214–1225.

[17] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*. OpenReview.net.

[18] Charlotte Laclau, Christine Largeron, and Manvi Choudhary. 2022. A survey on fairness for machine learning on graphs. *arXiv preprint arXiv:2205.05396* (2022).

[19] Anja Lambrecht and Catherine Tucker. 2019. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management science* 65, 7 (2019), 2966–2981.

[20] Yang Li, Buyue Qian, Xianli Zhang, and Hui Liu. 2020. Graph neural network-based diagnosis prediction. *Big data* 8, 5 (2020), 379–390.

[21] Yujia Li, Kevin Swersky, and Rich Zemel. 2015. Generative moment matching networks. In *ICML*, Vol. 37. PMLR, 1718–1727.

[22] Xiao Lin, Jian Kang, Weilin Cong, and Hanghang Tong. 2024. Bemap: Balanced message passing for fair graph neural network. In *LoG*, Vol. 231. PMLR, 37:1–37:18.

[23] Hongyi Ling, Zhimeng Jiang, Youzhi Luo, Shuiwang Ji, and Na Zou. 2023. Learning fair graph representations via automated data augmentations. In *ICLR*.

[24] Zemin Liu, Trung-Kien Nguyen, and Yuan Fang. 2023. On generalized degree fairness in graph neural networks. In *AAAI*, Vol. 37. AAAI Press, 4525–4533.

[25] Donald Loveland, Jiayi Pan, Aaresh Farrokh Bhathena, and Yiyang Lu. 2022. Fairedit: Preserving fairness in graph neural networks through greedy graph editing. *arXiv preprint arXiv:2201.03681* (2022).

[26] Jing Ma, Ruocheng Guo, Mengting Wan, Longqi Yang, Aidong Zhang, and Jundong Li. 2022. Learning fair node representations with graph counterfactual fairness. In *WSDM*. ACM, 695–703.

[27] John Palowitch and Bryan Perozzi. 2020. Debiasing graph representations via metadata-orthogonal training. In *ASONAM*. IEEE, 435–442.

[28] Clara Rus, Jeffrey Luppes, Harrie Oosterhuis, and Gido H. Schoenmacker. 2022. Closing the Gender Wage Gap: Adversarial Fairness in Job Recommendation. In *HR@RecSys (CEUR Workshop Proceedings, Vol. 3218)*. CEUR-WS.org.

[29] Akrati Saxena, George Fletcher, and Mykola Pechenizkiy. 2024. Fairsna: Algorithmic fairness in social network analysis. *Comput. Surveys* 56, 8 (2024), 1–45.

[30] Bernhard Scholkopf, Kah-Kay Sung, Christopher JC Burges, Federico Girosi, Partha Niyogi, Tomaso Poggio, and Vladimir Vapnik. 1997. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *TSP* 45, 11 (1997), 2758–2765.

[31] Latanya Sweeney. 2013. Discrimination in online ad delivery. *CACM* 56, 5 (2013), 44–54.

[32] Lubos Takac and Michal Zabovsky. 2012. Data analysis in public social networks. In *International scientific conference and international workshop present day trends of innovations*.

[33] Cédric Villani. 2021. *Topics in optimal transportation*. Vol. 58. American Mathematical Soc.

[34] Cédric Vincent-Cuaz, Rémi Flamary, Marco Corneli, Titouan Vayer, and Nicolas Courty. 2022. Template based graph neural network with optimal transport distances. *NeurIPS* 35 (2022), 11800–11814.

[35] Yan Wang, Ruochi Zhang, Qian Yang, Qiong Zhou, Shengde Zhang, Yusi Fan, Lan Huang, Kewei Li, and Fengfeng Zhou. 2024. FairCare: Adversarial training of a heterogeneous graph neural network with attention mechanism to learn fair representations of electronic health records. *IPM* 61, 3 (2024), 103682.

[36] Jiaxing Zhang, Zhuomin Chen, Hao Mei, Longchao Da, Dongsheng Luo, and Hua Wei. 2024. RegExplainer: Generating Explanations for Graph Neural Networks in Regression Tasks. In *NeurIPS*.

[37] Binbin Zhou, Hang Zhou, Weikun Wang, Liming Chen, Jianhua Ma, and Zengwei Zheng. 2024. HDM-GNN: A Heterogeneous Dynamic Multi-view Graph Neural Network for Crime Prediction. *TOSN* (2024).

[38] Huaisheng Zhu, Guoji Fu, Zhimeng Guo, Zhiwei Zhang, Teng Xiao, and Suhang Wang. 2023. Fairness-aware Message Passing for Graph Neural Networks. *arXiv preprint arXiv:2306.11132* (2023).