SCIENCE CHINA
Information Sciences

• Supplementary File •

Topology Matters: Achieving Fairness in Graph Neural Networks through Heterophily Propagation

Yuge Wang[1], Xibei Yang[1*], Keyu Liu[1], Qiguo Sun[1], Weiping Ding[2] & Yuhua Qian[3]

[1]*School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212003, China*
[2]*School of Information Science and Technology, Nantong University, Nantong 226019, China*
[3]*Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, China*

Appendix A Methods

Topology bias in GNNs, such as the imbalance of homo/heterogeneous neighbors within a node's neighborhood, is a significant cause of unfair embeddings. Determining whether the neighbors of any given node are balanced and mitigating this phenomenon have become two challenging tasks. Our proposed method consists of two core modules, the sensitivity operator module and the adaptive message-passing (MP) mechanism module, designed to address the two aforementioned issues.

Appendix A.1 Sensitivity operator module

We design a neighborhood balance-aware approach that leverages the opposite sensitive attributes. In recent years, many researchers [1,2] have adopted similar ideas, using counterfactual fairness [3] to evaluate the stability of decisions in response to changes in the sensitive attribute values within feature. However, they overlook the fact that the sensitive attribute and other features follow a joint prior distribution $(x, s) \sim prior$ [4]. Merely modifying a single attribute is insufficient to yield a feature that is the opposite counterpart. Therefore, in this module, we aggregate the complete opposite node features to obtain **counter-sensitive feature** $\bar{x}$ for central node v_i:

$$\bar{x}_i = x_i + \text{AGGREGATE}(\{x_j : v_j \in \mathcal{N}^t(v_i), s_j \neq s_i\}), \tag{A1}$$

where s_i and $\mathcal{N}^t(v_i)$ represents the sensitive attribute and set of nodes in the t-hop neighborhood of node v_i, respectively, and AGGREGATE function is the aggregation method (*e.g.*, sum, mean, or max).

To formalize the above process, we first perform MP on the original node feature to obtain the initial embedding h_i:

$$h_i = \text{MP}\left(x_i, \{x_j : v_j \in \mathcal{N}(v_i)\}\right). \tag{A2}$$

Subsequently, we generate counter-sensitive embedding $\bar{h}_i$ by applying the same MP operation on the modified feature $\bar{x}_i$ while preserving the original neighborhood structure:

$$\bar{h}_i = \text{MP}\left(\bar{x}_i, \{x_j : v_j \in \mathcal{N}(v_i)\}\right). \tag{A3}$$

To quantify the discrepancy between the two embeddings, we compute **fairness sensitivity (FS)** through the **logarithmic absolute difference (LAd)** [5] across all dimensions D of the embeddings:

$$\text{FS}_i = \sum_{d=1}^{D} \log\left(1 + |h_{i,d} - \bar{h}_{i,d}|\right), \tag{A4}$$

where $h_{i,d}$ denotes d-th attribute in the embedding h_i.

Due to the superior performance of LAd in amplifying differences, FS is able to capture the embedding disparities at the attribute level. Furthermore, FS quantitatively characterizes node susceptibility to sensitive attributes, thereby providing guidance for the systematic integration of homogeneous and heterogeneous information. To establish methodological foundations, we formally define the normalized FS measure as the **sensitivity operator** $\mathbb{S}$:

$$\mathbb{S} = \text{diag}(\text{FS}_i) \cdot \mathbf{D}^{-1} \in \mathbb{R}^{n \times n}, \tag{A5}$$

where FS_i constitutes the diagonal elements, $\mathbf{D}$ is the degree matrix of $\mathbf{A}$, which is used for normalization.

* Corresponding author (email: jsjxy_yxb@just.edu.cn)

Appendix A.2 Adaptive message-passing mechanism module

To mitigate topology bias induced by homogeneous-heterogeneous neighbor imbalance, we propose an adaptive heterogeneous neighbor integration during message propagation. Formally, FairDHP introduces two supplementary message-passing mechanisms, **local disentangled propagation (LDP)** and **global heterophily-aware propagation (GHP)**:

$$\mathcal{F} = \mathcal{F}_{\mathrm{BAS}} \otimes \mathcal{F}_{\mathrm{LDP}} \otimes \mathcal{F}_{\mathrm{GHP}}, \tag{A6}$$

where $\mathcal{F}_{\mathrm{BAS}}$ is the basic MP, and $\otimes$ denotes the fusion method.

Local disentangled propagation. In undirected graph-based MP systems, information propagates bidirectionally through edges, analogous to flow in channels [6]. This inherent characteristic exacerbates bias amplification under neighborhood imbalance conditions, where homogeneous node signals overwhelmingly dominate their heterogeneous counterparts. To address this limitation, we implement an edge disentanglement strategy that decouples undirected connections into *in-flow* and *out-flow*:

$$\mathcal{E} = \mathcal{E}_{\mathrm{in}} \cup \mathcal{E}_{\mathrm{out}}, \tag{A7}$$

where $\mathcal{E}_{\mathrm{in}}$ governs *in-flow* propagation and $\mathcal{E}_{\mathrm{out}}$ regulates *out-flow* pathways. This architectural innovation strategically attenuates homogeneous signal transmission while preserving heterogeneous information aggregation.

We design a set of message passing strategies to adapt to dual flows. For homogeneous connections, we provide throttling based on sensitivity operator $\mathbb{S}$. For heterogeneous connections, excessive aggregation should also be regularized to prevent the introduction of noise.

For this purpose, the adjacency matrix $\mathbf{A}$ requires transformation. The *in-flow* and the *out-flow* can be achieved by modifying the edge weights. As a result, the refined adjacency weights are computed through:

$$\tilde{\mathbf{A}} = \underbrace{\mathbf{A} \odot [\mathbf{I} \cdot (\mathbf{J} - \mathbb{S})]}_{\text{Homogeneous throttling}} + \underbrace{\mathbf{A} \odot [(\mathbf{J} - \mathbf{I}) \cdot \mathbb{S}]}_{\text{Heterogeneous regularization}}, \tag{A8}$$

where $\mathbf{I}$ is the homogeneity indicator matrix with $\mathbf{I}_{\mathrm{i,j}} = \mathbf{1}\,(\mathrm{s}_i = \mathrm{s}_j)$, $\mathbf{J}$ denotes the diagonal matrix with $\mathbf{J}_{i,j} = 1$, $\odot$ represents Hadamard product.

In practice, the refined adjacency matrix $\tilde{\mathbf{A}}$ is deployed into the GNN framework to generate node embeddings $\tilde{\mathbf{H}}$. For GCN [7] implementation, the layer-wise propagation rule with LDP constraints at the l-th layer is formulated as:

$$\tilde{\mathbf{H}}^{(l+1)} = \mathrm{Softmax}(\tilde{\mathbf{A}}\tilde{\mathbf{H}}^{(l)}\tilde{\mathbf{W}}^{(l)}), \tag{A9}$$

where $\tilde{\mathbf{H}}^{(0)} = \mathbf{X}$.

However, this strategy exhibits limitation: the neighborhood-centric paradigm becomes suboptimal when local heterogeneous neighbors are statistically insufficient or entirely absent, potentially compromising bias mitigation efficacy.

Remark 1. The edge disentanglement approach in LDP enhances the MP process by decoupling information flows, a concept explored in Flow2GNN [6], which uses a two-way flow MP scheme to disentangle topological information in graphs. By separating edge connections into distinct categories [8] (*e.g.*, homophily and heterophily), LDP ensures that information flow is more controlled, leading to more robust and interpretable node representations. In previous researches, the disentanglement strategy primarily focus on disentangling features to achieve fairness. For instance, in the researches of [1,9,10], disentanglement focuses on separating sensitive and non-sensitive features in the latent space, enhancing fairness in representation learning. However, compared to LDP's edge disentanglement strategy, these approaches do not comprehensively consider the graph's structural dependencies. By specifically targeting the edges that connect nodes in a graph, LDP offers a more precise mechanism for ensuring fairness in node representations without relying solely on feature-level disentanglement. This makes it more adaptable to graph-based data, where topology significantly impacts the representation learning process.

Algorithm A1 Local Disentangled Propagation

Require: Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, features $\mathbf{X}$, adjacency matrix $\mathbf{A}$, homogeneous indicator matrix $\mathbf{I}$, sensitivity operator $\mathbb{S}$, training epochs T.
Ensure: Fair embedding matrix $\tilde{\mathbf{H}}$.
1: **Initialize:** Parameter matrices $\tilde{\mathbf{W}}$.
2: Divide undirected connections into *in-flow* and *out-flow*.
3: **for** epoch = 1 to T **do**
4: Calculate the coefficient of throttling $\leftarrow \mathbf{J} - \mathbb{S}$.
5: Calculate the coefficient of regularization $\leftarrow \mathbb{S}$.
6: Modify the edge weights via $\mathbf{A} \odot [\mathbf{I} \cdot (\mathbf{J} - \mathbb{S})]$ and $\mathbf{A} \odot [(\mathbf{J} - \mathbf{I}) \cdot \mathbb{S}]$.
7: Refine the adjacency matrix via Eq. (A8).
8: Generate embeddings via Eq. (A9).
9: **end for**

Global heterophily-aware propagation. To address the locality constraints of LDP, we propose a novel method, global heterophily-aware propagation, that extends the receptive field of heterogeneous nodes from local neighborhoods to the entire graph.

Here, we depart from the conventional graph topology, focusing exclusively on feature-driven approaches to establish latent neighborhood relationships among nodes. We utilize **vector quantization (VQ)** technology [11, 12] to map nodes to several Voronoi regions [13], by optimizing the following formula:

$$\min_{\mathbf{P}\in\{0,1\}^{N\times P},\tilde{\mathbf{X}}\in\mathbb{R}^{P\times k}} \Phi(\mathbf{X}, \mathbf{P}\cdot\tilde{\mathbf{X}}), \tag{A10}$$

where $\mathbf{P}$ is the projection matrix assigning N nodes to P regions, $\tilde{\mathbf{X}}$ is the learnable region feature matrix initialized by random node features, and $\Phi(\cdot,\cdot)$ denotes the distance function that measures the dissimilarity between nodes and regions.

The generated Voronoi regions $\{\mathbf{V}_p\}_{p=1}^{P}$ are formally defined as the partitions of the node space, each corresponding to a centroid vector in $\tilde{\mathbf{X}}$, such that a node belongs to a region if its distance to that centroid (measured by Φ) is minimized over all regions:

$$\mathbf{V}_p = \left\{ v_i \in \mathcal{V} \ : \ \Phi\left(x_i, \tilde{x}_p\right) = \min_{q\in\{1,\ldots,P\}} \Phi\left(x_i, \tilde{x}_q\right) \right\}. \tag{A11}$$

However, Voronoi regions have two characteristics: $\bigcup_{p=1}^{P}\mathbf{V}_p \subseteq \mathbb{R}^n$ and $\bigcap_{p=1}^{P}\mathbf{V}_p = \emptyset$. This results in the node being unable to find neighbors across Voronoi regions. To further expand the receptive field, we employ VQ multiple times with different similarity metrics (*e.g.*, Euclidean and Cosine) to break down region barriers and achieve cross-region node connections. For example, a certain node is mapped to $\mathbf{V}_p^e$ and $\mathbf{V}_q^c$ in multiple VQ, and the information propagates within the region. Therefore, the node can integrate the information of all nodes in $\mathbf{V}_p^e$ and release it into $\mathbf{V}_q^c$, achieving cross-region message passing. We define nodes like this as cross-region nodes.

Let $\mathbf{P}^e \in \{0,1\}^{n\times P^e}$ and $\mathbf{P}^c \in \{0,1\}^{n\times P^c}$ denote the projection matrices derived from Euclidean and cosine similarity metrics respectively. And the cross region node awareness is achieved through concatenation. This leaves us with the following expression:

$$\mathbf{P}^{joint} = [\mathbf{P}^e \parallel \mathbf{P}^c] \in \{0,1\}^{n\times(P^e+P^c)}. \tag{A12}$$

The behavior of searching for neighbors across regions has natural similarities with HGNN. Therefore, we follow its method and construct a message passing path matrix $\mathbf{R}$, by computing the matrix product of the joint projection matrix and its transpose:

$$\mathbf{R} = \mathbf{P}^{joint}\cdot(\mathbf{P}^{joint})^T \in \{0,1\}^{n\times n}, \tag{A13}$$

we establish a generalized adjacency structure that inherently records message passing paths within and across regions.

Aligning with LDP in Eq. (A8), we refine the matrix $\mathbf{R}$ using the sensitivity operator $\mathbb{S}$. However, in order to supplement heterogeneous nodes, we only retain the propagation of heterogeneous information:

$$\hat{\mathbf{A}} = \mathbf{R}\odot[(\mathbf{J}-\mathbf{I})\cdot\mathbb{S}]. \tag{A14}$$

The global heterophily-aware propagation at layer l operates as:

$$\hat{\mathbf{H}}^{(l+1)} = \mathrm{Softmax}(\hat{\mathbf{A}}\hat{\mathbf{H}}^{(l)}\hat{\mathbf{W}}^{(l)}), \tag{A15}$$

where $\hat{\mathbf{H}}^{(0)} = \mathbf{X}$.

Algorithm A2 Global Heterophily-aware Propagation

Require: Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, features $\mathbf{X}$, adjacency matrix $\mathbf{A}$, homogeneous indicator matrix $\mathbf{I}$, sensitivity operator $\mathbb{S}$, training epochs T, number of regions with different distance functions P^e and P^c.

Ensure: Fair embedding matrix $\hat{\mathbf{H}}$.

Initialize: Parameter matrices $\hat{\mathbf{W}}$.
// Establish global latent neighborhood relationships.
Obtain projection matrix $\mathbf{P}^e$ via optimize Eq. (A10) with Euclidean distance and P^e.
Obtain projection matrix $\mathbf{P}^c$ via optimize Eq. (A10) with Cosine distance and P^c.
Concatenate projection matrices to generate cross-region nodes via Eq. (A12).
Calculate inner product to obtain message passing path via Eq. (A13).
// Propagate heterogeneous information.
for epoch = 1 to T **do**
    Calculate the coefficient of regularization $\leftarrow \mathbb{S}$.
    Modify the edge weights via $\mathbf{A}\odot[(\mathbf{J}-\mathbf{I})\cdot\mathbb{S}]$.
    Retain only heterogeneous edges and refine the adjacency matrix via Eq. (A14).
    Generate embeddings via Eq. (A15).
end for

According to Eq. (A6), the final embeddings adaptively combine three mechanisms:

$$\mathbf{H}^{(l)} = \sigma((\mathbf{J} - \mathbb{S}) \cdot \underbrace{\bar{\mathbf{H}}^{(l)}}_{\mathcal{F}_{\text{BAS}}} + \frac{\mathbb{S}}{2} \cdot \underbrace{\hat{\mathbf{H}}^{(l)}}_{\mathcal{F}_{\text{LDP}}} + \frac{\mathbb{S}}{2} \cdot \underbrace{\tilde{\mathbf{H}}^{(l)}}_{\mathcal{F}_{\text{GHP}}}). \tag{A16}$$

Algorithm A3 Training Framework of FairDHP

Require: Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, features $\mathbf{X}$, adjacency matrix $\mathbf{A}$, homogeneous indicator matrix $\mathbf{I}$, sensitivity operator $\mathbb{S}$, training epochs T, the number of regions with different distance functions P^e and P^c.

Ensure: Fair embedding matrix $\mathbf{H}$.

Initialize: Parameter matrices $\overline{\mathbf{W}}, \tilde{\mathbf{W}}, \hat{\mathbf{W}}$.
// Preprocessing Phase.
for $v_i \in \mathcal{V}$ **do**
  Obtain counter-sensitive feature $\bar{x}_i$ via Eq. (A1) with t.
  Generate initial embedding h_i and counter-sensitive embedding $\bar{h}_i$ via Eq. (A2) and (A3).
  Calculate fairness sensitivity via Eq. (A4).
end for
Get sensitivity operator $\mathbb{S}$ via Eq. (A5).
// Refine the adjacency matrices.
Refine Adj $\tilde{\mathbf{A}}$ for LDP via Eq. (A7) - (A8).
Refine Adj $\hat{\mathbf{A}}$ for GHP via Eq. (A10) - (A14).
// Training Phase.
for epoch $= 1$ to T **do**
  Run parallel message passing:
    $\bar{\mathbf{H}} \leftarrow \mathcal{F}_{\text{BAS}}(\mathbf{A}, \mathbf{X}, \overline{\mathbf{W}})$.
    $\tilde{\mathbf{H}} \leftarrow \mathcal{F}_{\text{LDP}}(\tilde{\mathbf{A}}, \mathbf{X}, \tilde{\mathbf{W}})$.
    $\hat{\mathbf{H}} \leftarrow \mathcal{F}_{\text{GHP}}(\hat{\mathbf{A}}, \mathbf{X}, \hat{\mathbf{W}})$.
  $\mathbf{H} \leftarrow$ adaptive fusion $(\bar{\mathbf{H}}, \tilde{\mathbf{H}}, \hat{\mathbf{H}}, \mathbb{S})$.
  Calculate loss.
  Update model parameters $\bar{\mathbf{W}}, \tilde{\mathbf{W}}, \hat{\mathbf{W}}$ with spectral normalization.
end for

Appendix B Theoretical proof

Appendix B.1 Lemma 1

The following assumptions are made for the theoretical findings in this study:

Assumption 1 (*Finite-valued embeddings*). $\|\mathbf{c}_j^{l+1} - \bar{\mathbf{c}}_s^{l+1}\|_\infty \leqslant (\Delta_c^{(s)})^{l+1}$, $\forall v_j \in \mathcal{S}_g$ with $g \in \{0, 1\}$, where $\Delta_c^{l+1} = \max((\Delta_c^{(0)})^{l+1}, (\Delta_c^{(1)})^{l+1})$. $\left\|\mathbf{z}_j^{l+1} - \bar{\mathbf{z}}_s^{l+1}\right\|_\infty \leqslant (\Delta_z^{(s)})^{l+1}$, $\forall v_j \in \mathcal{S}_g$ with $g \in \{0, 1\}$, where $\Delta_z^{l+1} = \max((\Delta_z^{(0)})^{l+1}, (\Delta_z^{(1)})^{l+1})$. Here, $\max(\cdot, \cdot)$ outputs the element-wise maximum of the input vectors.

Assumption 2 (*Connection Homophily*). η and η' are the sample mean of heterogeneous neighbors normalized by degree, $\eta = \text{mean}\left(\frac{|\mathcal{N}(i) \cap \mathcal{S}_{\bar{g}}|}{|\mathcal{N}(i)|} \middle| v_i \in \mathcal{V}\right)$, $\eta' = \text{mean}\left(\frac{|\hat{\mathcal{N}}(i) \cap \mathcal{S}_{\bar{g}}|}{|\mathcal{N}(i)|} \middle| v_i \in \mathcal{V}\right)$, where $\bar{g} = 1 - s_i$. Owing to the inherent homophily in network connectivity patterns [4, 14, 15], where heterogeneous node counts within neighborhoods are strictly less than homogeneous ones, η and η' are theoretically bounded within $(0, 0.5)$.

Based on these, Lemma 1 demonstrates the factors that contribute to the statistical parity (represented by δ_h) between the embeddings of different sensitive groups obtained at the l-th $\mathcal{F}_{\text{LDP}}$ and $\mathcal{F}_{\text{GHP}}$. Specifically, Theorem 1 upper bounds the term $\delta_h^{l+1} := \left\|\text{mean}(\mathbf{h}_j^{l+1} \mid s_j = 0) - \text{mean}(\mathbf{h}_j^{l+1} \mid s_j = 1)\right\|_2$.

Lemma 1. The statistical parity δ_h^{l+1} between the embeddings of different sensitive groups that are output by the l-th $\mathcal{F}_{\text{LDP}}$ and $\mathcal{F}_{\text{GHP}}$, can be upper bounded by:

$$\delta_h^{l+1} \leqslant L\Big(\sigma_{max}(\mathbf{W}^l) \cdot |(1 - \alpha^+)(1 - \eta) + (1 - \min(R_0^\chi, R_1^\chi))\eta\alpha^+ - (R_0^\chi + R_1^\chi)\eta'\alpha^+|\delta_h^l + 2\sqrt{N}\Delta_c^{l+1} + 2\sqrt{N}\Delta_z^{l+1}\Big),$$

where L is the Lipschitz constant of nonlinear activation σ, $\sigma_{max}(\cdot)$ denotes the largest singular value of the parameter matrix, $R_1^\chi := \frac{|S_1^\chi|}{|\mathcal{S}_1|}, R_0^\chi := \frac{|S_0^\chi|}{|\mathcal{S}_0|}$, and $\alpha^+ = \text{mean}(\text{FS} \mid v_i \in \mathcal{V})$.

Proof. Here, without loss of generality, we will consider the l-th $\mathcal{F}_{\text{LDP}}$ and $\mathcal{F}_{\text{GHP}}$, where the input embeddings are denoted by $\mathbf{H}^l$ and output embeddings are $\mathbf{H}^{l+1}$. The disparity between the output embeddings follows as:

$$\begin{aligned}\delta_h^{l+1} &:= \left\|\text{mean}(\mathbf{h}_j^{l+1} \mid s_j = 0) - \text{mean}(\mathbf{h}_j^{l+1} \mid s_j = 1)\right\|_2 \\ &= \left\|\frac{1}{|\mathcal{S}_0|}\sum_{v_j\in\mathcal{S}_0}\mathbf{h}_j^{l+1} - \frac{1}{|\mathcal{S}_1|}\sum_{v_j\in\mathcal{S}_1}\mathbf{h}_j^{l+1}\right\|_2.\end{aligned}$$

According to Lemma A.1. in [16], δ_h^{l+1} can be upper bounded by the following term:

$$\left\|\frac{1}{|\mathcal{S}_0|}\sum_{v_j\in\mathcal{S}_0}\mathbf{h}_j^{l+1} - \frac{1}{|\mathcal{S}_1|}\sum_{v_j\in\mathcal{S}_1}\mathbf{h}_j^{l+1}\right\|_2 \leqslant L\left(\left\|\frac{1}{|\mathcal{S}_0|}\sum_{v_j\in\mathcal{S}_0}\mathbf{z}_j^{l+1} - \frac{1}{|\mathcal{S}_1|}\sum_{v_j\in\mathcal{S}_1}\mathbf{z}_j^{l+1}\right\|_2 + 2\sqrt{N}\Delta_z^{l+1}\right). \tag{B1}$$

Based on this upper bound, the term $\left\|\frac{1}{|\mathcal{S}_0|}\sum_{v_j\in\mathcal{S}_0}\mathbf{z}_j^{l+1} - \frac{1}{|\mathcal{S}_1|}\sum_{v_j\in\mathcal{S}_1}\mathbf{z}_j^{l+1}\right\|_2$ will be analyzed. We first consider the terms $\frac{1}{|\mathcal{S}_1|}\sum_{v_j\in\mathcal{S}_1}\mathbf{z}_j^{l+1}$ and $\frac{1}{|\mathcal{S}_0|}\sum_{v_j\in\mathcal{S}_0}\mathbf{z}_j^{l+1}$ individually.

Let's re-define aggregated embeddings for node i at l-th $\mathcal{F}_{\text{LDP}}$ and $\mathcal{F}_{\text{GHP}}$ as:

$$\mathbf{z}_i^{l+1} := \underbrace{\sum_{j\in\mathcal{N}(i)}\frac{\alpha_{ij}}{D_i}\mathbf{h}_j^l\tilde{\mathbf{W}}^l}_{\mathcal{F}_{\text{LDP}}} + \underbrace{\sum_{j\in\hat{\mathcal{N}}(i)}\frac{\alpha_{ij}}{D_i}\mathbf{h}_j^l\hat{\mathbf{W}}^l}_{\mathcal{F}_{\text{GHP}}}.$$

We make $\mathbf{c}_i^{l+1} := \mathbf{h}_i^l\mathbf{W}^l$ to obtain a simplified form. The sample means of $\mathbf{c}^{l+1}$ and $\mathbf{z}^{l+1}$ vectors are represented by $\bar{\mathbf{c}}_g^{l+1} := \text{mean}(\mathbf{c}_j^{l+1} \mid v_j \in \mathcal{S}_g)$ and $\bar{\mathbf{z}}_g^{l+1} := \text{mean}(\mathbf{z}_j^{l+1} \mid v_j \in \mathcal{S}_g)$ for the nodes in sensitive group $\mathcal{S}_g$ for $g = 0, 1$. While considering assumption 1, the aggregation of node representations $\mathbf{z}_j^{l+1}$ over set $\mathcal{S}_1$ can be written:

$$\begin{aligned}&\frac{1}{|\mathcal{S}_1|}\sum_{v_j\in\mathcal{S}_1}\mathbf{z}_j^{l+1} \\ &\in \underbrace{\frac{1}{|\mathcal{S}_1|}\left(\sum_{v_k\in S_1^{\mathcal{X}}}\Big(\sum_{a\in\mathcal{N}(k)\cap\mathcal{S}_0}\frac{\alpha_{ka}}{|\mathcal{N}(k)|}\bar{\mathbf{c}}_0^{l+1} + \sum_{b\in\mathcal{N}(k)\cap\mathcal{S}_1}\frac{\alpha_{kb}}{|\mathcal{N}(k)|}\bar{\mathbf{c}}_1^{l+1}\Big) + \sum_{v_k\in S_1^{\omega}}\sum_{a\in\mathcal{N}(k)\cap\mathcal{S}_1}\frac{\alpha_{ka}}{|\mathcal{N}(k)|}\bar{\mathbf{c}}_1^{l+1}\right)}_{\mathcal{F}_{\text{LDP}}} \\ &+ \underbrace{\frac{1}{|\mathcal{S}_1|}\left(\sum_{v_k\in S_1^{\mathcal{X}}}\sum_{a\in\hat{\mathcal{N}}(k)\cap\mathcal{S}_0}\frac{\alpha_{ka}}{|\mathcal{N}(k)|}\bar{\mathbf{c}}_0^{l+1} + \sum_{v_k\in S_1^{\omega}}\sum_{a\in\mathcal{N}(k)\cap\mathcal{S}_0}\frac{\alpha_{ka}}{|\mathcal{N}(k)|}\bar{\mathbf{c}}_1^{l+1}\right)}_{\mathcal{F}_{\text{GHP}}} \pm\Delta_c^{l+1}\mathbf{1}.\end{aligned}$$

The aggregation coefficients are defined through the fairness sensitivity $\text{FS}(\cdot)$:

$$\alpha_{ka} = \begin{cases}\alpha_k^+ = \text{FS}(k) & if\ s_k \neq s_a, \\ \alpha_k^- = 1 - \text{FS}(k) & if\ s_k = s_a.\end{cases}$$

Consequently, $\frac{1}{|\mathcal{S}_1|}\sum_{v_j\in\mathcal{S}_1}\mathbf{z}_j^{l+1}$ satisfies:

$$\begin{aligned}&\frac{1}{|\mathcal{S}_1|}\sum_{v_j\in\mathcal{S}_1}\mathbf{z}_j^{l+1} \\ &\in \frac{1}{|\mathcal{S}_1|}\left(\sum_{v_k\in S_1^{\mathcal{X}}}\Big(\sum_{a\in\mathcal{N}(k)\cap\mathcal{S}_0}\frac{\alpha_k^+}{|\mathcal{N}(k)|}\bar{\mathbf{c}}_0^{l+1} + \sum_{b\in\mathcal{N}(k)\cap\mathcal{S}_1}\frac{\alpha_k^-}{|\mathcal{N}(k)|}\bar{\mathbf{c}}_1^{l+1}\Big) + \sum_{v_k\in S_1^{\omega}}\sum_{a\in\mathcal{N}(k)\cap\mathcal{S}_1}\frac{\alpha_k^-}{|\mathcal{N}(k)|}\bar{\mathbf{c}}_1^{l+1}\right) \\ &+ \frac{1}{|\mathcal{S}_1|}\left(\sum_{v_k\in S_1^{\mathcal{X}}}\sum_{a\in\hat{\mathcal{N}}(k)\cap\mathcal{S}_0}\frac{\alpha_k^+}{|\mathcal{N}(k)|}\bar{\mathbf{c}}_0^{l+1} + \sum_{v_k\in S_1^{\omega}}\sum_{a\in\hat{\mathcal{N}}(k)\cap\mathcal{S}_0}\frac{\alpha_k^+}{|\mathcal{N}(k)|}\bar{\mathbf{c}}_1^{l+1}\right) \pm \Delta_c^{l+1}\mathbf{1}.\end{aligned}$$

By further decomposing the neighborhood terms, we obtain the expanded formulation:

$$
\begin{aligned}
&\frac{1}{|\mathcal{S}_1|}\sum_{v_j\in\mathcal{S}_1}\mathbf{z}_j^{l+1}\\
&\in\frac{1}{|\mathcal{S}_1|}\left(\sum_{v_k\in S_1^{\chi}}(\frac{|\mathcal{N}(k)\cap\mathcal{S}_0|}{|\mathcal{N}(k)|}\alpha_k^{+}\bar{\mathbf{c}}_0^{l+1}+\frac{|\mathcal{N}(k)\cap\mathcal{S}_1|}{|\mathcal{N}(k)|}\alpha_k^{-}\bar{\mathbf{c}}_1^{l+1})+\sum_{v_k\in S_1^{\omega}}\frac{|\mathcal{N}(k)\cap\mathcal{S}_1|}{|\mathcal{N}(k)|}\alpha_k^{-}\bar{\mathbf{c}}_1^{l+1}\right)\\
&+\frac{1}{|\mathcal{S}_1|}\left(\sum_{v_k\in S_1^{\chi}}\frac{|\hat{\mathcal{N}}(k)\cap\mathcal{S}_0|}{|\mathcal{N}(k)|}\alpha_k^{+}\bar{\mathbf{c}}_0^{l+1}+\sum_{v_k\in S_1^{\omega}}\frac{|\hat{\mathcal{N}}(k)\cap\mathcal{S}_0|}{|\mathcal{N}(k)|}\alpha_k^{+}\bar{\mathbf{c}}_1^{l+1}\right)\pm\Delta_c^{l+1}\mathbf{1}.
\end{aligned}
$$

Let's define $R_1^{\chi}:=\frac{|S_1^{\chi}|}{|\mathcal{S}_1|}$ and $R_0^{\chi}:=\frac{|S_0^{\chi}|}{|\mathcal{S}_0|}$, where $R_1^{\omega}=1-R_1^{\chi}$, $R_0^{\omega}=1-R_0^{\chi}$; α^{+} is the mean of α_k^{+}, where $\alpha^{+}=1-\alpha^{-}$. Similarly, the proportion of heterogeneous node neighborhoods on can be measured using the sample mean in the following way:

$$
\begin{aligned}
\eta&=\text{mean}\left(\frac{|\mathcal{N}(i)\cap\mathcal{S}_{\bar{g}}|}{|\mathcal{N}(i)|}\Big|v_i\in\mathcal{V}\right),\\
\eta'&=\text{mean}\left(\frac{|\hat{\mathcal{N}}(i)\cap\mathcal{S}_{\bar{g}}|}{|\mathcal{N}(i)|}\Big|v_i\in\mathcal{V}\right),
\end{aligned}
$$

where $\bar{g}=1-s_i$. The sample mean of homogeneous node is $1-\eta$.

So, the equation can be written as:

$$
\frac{1}{|\mathcal{S}_1|}\sum_{v_j\in\mathcal{S}_1}\mathbf{z}_j^{l+1}\in R_1^{\chi}\eta\alpha^{+}\bar{\mathbf{c}}_0^{l+1}+R_1^{\chi}(1-\eta)\alpha^{-}\bar{\mathbf{c}}_0^{l+1}+R_1^{\omega}(1-\eta)\alpha^{-}\bar{\mathbf{c}}_0^{l+1}+R_1^{\chi}\eta'\alpha^{+}\bar{\mathbf{c}}_0^{l+1}+R_1^{\omega}\eta'\alpha^{+}\bar{\mathbf{c}}_0^{l+1}\pm\Delta_c^{l+1}\mathbf{1}.
$$

Replace R_1^{ω}, α^{-} with R_1^{χ}, α^{+}:

$$
\begin{aligned}
\frac{1}{|\mathcal{S}_1|}\sum_{v_j\in\mathcal{S}_1}\mathbf{z}_j^{l+1}&\in R_1^{\chi}\eta\alpha^{+}\bar{\mathbf{c}}_0^{l+1}+R_1^{\chi}(1-\eta)(1-\alpha^{+})\bar{\mathbf{c}}_0^{l+1}+(1-R_1^{\chi})(1-\eta)(1-\alpha^{+})\bar{\mathbf{c}}_0^{l+1}+R_1^{\chi}\eta'\alpha^{+}\bar{\mathbf{c}}_0^{l+1}+(1-R_1^{\chi})\eta'\alpha^{+}\bar{\mathbf{c}}_0^{l+1}\pm\Delta_c^{l+1}\mathbf{1},\\
&\in R_1^{\chi}\eta\alpha^{+}\bar{\mathbf{c}}_0^{l+1}+\bar{\mathbf{c}}_1^{l+1}-\alpha^{+}\bar{\mathbf{c}}_1^{l+1}-\eta\bar{\mathbf{c}}_1^{l+1}+2\eta\alpha^{+}\bar{\mathbf{c}}_1^{l+1}+R_1^{\chi}\eta'\alpha^{+}\bar{\mathbf{c}}_0^{l+1}-R_1^{\chi}\eta'\alpha^{+}\bar{\mathbf{c}}_1^{l+1}\pm\Delta_c^{l+1}\mathbf{1}.
\end{aligned}
$$

In the same manner, $\frac{1}{|\mathcal{S}_0|}\sum_{v_j\in\mathcal{S}_0}\mathbf{z}_j^{l+1}$ can be represented by:

$$
\frac{1}{|\mathcal{S}_0|}\sum_{v_j\in\mathcal{S}_0}\mathbf{z}_j^{l+1}\in R_1^{\chi}\eta\alpha^{+}\bar{\mathbf{c}}_0^{l+1}+\bar{\mathbf{c}}_1^{l+1}-\alpha^{+}\bar{\mathbf{c}}_1^{l+1}-\eta\bar{\mathbf{c}}_1^{l+1}+2\eta\alpha^{+}\bar{\mathbf{c}}_1^{l+1}+R_1^{\chi}\eta'\alpha^{+}\bar{\mathbf{c}}_0^{l+1}-R_1^{\chi}\eta'\alpha^{+}\bar{\mathbf{c}}_1^{l+1}\pm\Delta_c^{l+1}\mathbf{1}.
$$

Define $\varepsilon^{l+1}:=\frac{1}{|\mathcal{S}_1|}\sum_{v_j\in\mathcal{S}_1}\mathbf{z}_j^{l+1}-\frac{1}{|\mathcal{S}_0|}\sum_{v_j\in\mathcal{S}_0}\mathbf{z}_j^{l+1}$, the following can be written:

$$
\begin{aligned}
\varepsilon^{l+1}\in&\,\bar{\mathbf{c}}_1^{l+1}(1-\alpha^{+}-\eta+2\eta\alpha^{+}-R_0^{\chi}\eta\alpha^{+}-R_0^{\chi}\eta'\alpha^{+}-R_1^{\chi}\eta'\alpha^{+})\\
&-\bar{\mathbf{c}}_0^{l+1}(1-\alpha^{+}-\eta+2\eta\alpha^{+}-R_1^{\chi}\eta\alpha^{+}-R_0^{\chi}\eta'\alpha^{+}-R_1^{\chi}\eta'\alpha^{+})\\
&\pm2\Delta_c^{l+1}\mathbf{1}.
\end{aligned}
$$

Merge similar items to obtain:

$$
\begin{aligned}
\varepsilon^{l+1}\in&\underbrace{(\bar{\mathbf{c}}_1^{l+1}-\bar{\mathbf{c}}_0^{l+1})\left[(1-\alpha^{+})(1-\eta)+\eta\alpha^{+}-(R_0^{\chi}+R_1^{\chi})\eta'\alpha^{+}\right]}_{\text{term }i}\\
&+\underbrace{\left[\bar{\mathbf{c}}_1^{l+1}(-R_0^{\chi}\eta\alpha^{+})-\bar{\mathbf{c}}_0^{l+1}(-R_1^{\chi}\eta\alpha^{+})\right]}_{\text{term }ii}\pm2\Delta_c^{l+1}\mathbf{1}.
\end{aligned}
$$

We will enlarge term ii into a form that matches term i. Let's assume $R_1^{\chi}<R_0^{\chi}$, we have $\bar{\mathbf{c}}_1^{l+1}(-R_0^{\chi}\eta\alpha^{+})-\bar{\mathbf{c}}_0^{l+1}(-R_1^{\chi}\eta\alpha^{+})=(\bar{\mathbf{c}}_1^{l+1}-\bar{\mathbf{c}}_0^{l+1})(-R_1^{\chi}\eta\alpha^{+})$. Due to symmetry, term ii can be exchange into:

$$
\bar{\mathbf{c}}_1^{l+1}(-R_0^{\chi}\eta\alpha^{+})-\bar{\mathbf{c}}_0^{l+1}(-R_1^{\chi}\eta\alpha^{+})<(\bar{\mathbf{c}}_1^{l+1}-\bar{\mathbf{c}}_0^{l+1})(-\min(R_0^{\chi},R_1^{\chi})\eta\alpha^{+}).
$$

Therefore, $\|\varepsilon^{l+1}\|_2$ can be upper bounded by:

$$
\|\varepsilon^{l+1}\|_2\leqslant\|\bar{\mathbf{c}}_0^{l+1}-\bar{\mathbf{c}}_1^{l+1}\|_2\cdot|(1-\alpha^{+})(1-\eta)+(1-\min(R_0^{\chi},R_1^{\chi}))\eta\alpha^{+}-(R_0^{\chi}+R_1^{\chi})\eta'\alpha^{+}|+2\sqrt{N}\Delta_c^{l+1}.\tag{B2}
$$

Furthermore, consider the term $\|\bar{\mathbf{c}}_0^{l+1} - \bar{\mathbf{c}}_1^{l+1}\|_2$, where $\bar{\mathbf{c}}_0^{l+1} = \text{mean}(\mathbf{c}_j^{l+1} \mid v_j \in \mathcal{S}_s)$ and $\mathbf{c}_j^{l+1} = \mathbf{h}_j^l \mathbf{W}^l$.

$$\begin{aligned} \|\frac{1}{|\mathcal{S}_0|}\sum_{v_j\in\mathcal{S}_0}\mathbf{c}_j^{l+1} - \frac{1}{|\mathcal{S}_1|}\sum_{v_j\in\mathcal{S}_1}\mathbf{c}_j^{l+1}\|_2 &= \|\frac{1}{|\mathcal{S}_0|}\sum_{v_j\in\mathcal{S}_0}\mathbf{h}_j^l\mathbf{W}^l - \frac{1}{|\mathcal{S}_1|}\sum_{v_j\in\mathcal{S}_1}\mathbf{h}_j^l\mathbf{W}^l\|_2 \\ &= \|\mathbf{W}^l(\frac{1}{|\mathcal{S}_0|}\sum_{v_j\in\mathcal{S}_0}\mathbf{h}_j^l - \frac{1}{|\mathcal{S}_1|}\sum_{v_j\in\mathcal{S}_1}\mathbf{h}_j^l)\|_2 \\ &\leqslant \sigma_{max}(\mathbf{W}^l)\|\frac{1}{|\mathcal{S}_0|}\sum_{v_j\in\mathcal{S}_0}\mathbf{h}_j^l - \frac{1}{|\mathcal{S}_1|}\sum_{v_j\in\mathcal{S}_1}\mathbf{h}_j^l\|_2, \end{aligned} \tag{B3}$$

where $\sigma_{max}(.)$ outputs the largest singular value of the input matrix. Based on Eq. (B2) and Eq. (B3), it follows that:

$$\begin{aligned} \|\varepsilon^{l+1}\|_2 \leqslant\ & \sigma_{max}(\mathbf{W}^l) \\ & \cdot |(1-\alpha^+)(1-\eta) + (1-\min(R_0^\chi, R_1^\chi))\eta\alpha^+ - (R_0^\chi + R_1^\chi)\eta'\alpha^+| \\ & \cdot \|\frac{1}{|\mathcal{S}_0|}\sum_{v_j\in\mathcal{S}_0}\mathbf{h}_j^l - \frac{1}{|\mathcal{S}_1|}\sum_{v_j\in\mathcal{S}_1}\mathbf{h}_j^l\|_2 + 2\sqrt{N}\Delta_c^{l+1}. \end{aligned} \tag{B4}$$

Finally, combining the results in Eq. (B1) and Eq. (B4), deviation between the output representations from different sensitive groups can be upper bounded by:

$$\begin{aligned} &\left\|\frac{1}{|\mathcal{S}_0|}\sum_{v_j\in\mathcal{S}_0}\mathbf{h}_j^{l+1} - \frac{1}{|\mathcal{S}_1|}\sum_{v_j\in\mathcal{S}_1}\mathbf{h}_j^{l+1}\right\|_2 \\ &\leqslant L\Bigg(\sigma_{\max}(\mathbf{W}^l)\left|(1-\alpha^+)(1-\eta) + (1-\min(R_0^\chi, R_1^\chi))\eta\alpha^+ - (R_0^\chi + R_1^\chi)\eta'\alpha^+\right|\left\|\frac{1}{|\mathcal{S}_0|}\sum_{v_j\in\mathcal{S}_0}\mathbf{h}_j^l - \frac{1}{|\mathcal{S}_1|}\sum_{v_j\in\mathcal{S}_1}\mathbf{h}_j^l\right\|_2 \\ &\qquad + 2\sqrt{N}\Delta_c^{l+1} + 2\sqrt{N}\Delta_z^{l+1}\Bigg). \end{aligned} \tag{B5}$$

Appendix B.2 Theorem 1

Theorem 1. Given parameters $\alpha^+, R_0^\chi, R_1^\chi \in (0,1)$ and $\eta, \eta' \in (0, 0.5)$, the absolute value accumulation term is bounded by:

$$|\underbrace{(1-\alpha^+)(1-\eta)}_{\text{term } i} + \underbrace{[1-\min(R_0^\chi, R_1^\chi)]\eta\alpha^+}_{\text{term } ii} - \underbrace{(R_0^\chi + R_1^\chi)\eta'\alpha^+}_{\text{term } iii}| \in [0,1]. \tag{B6}$$

Proof. To rigorously characterize its bounding mechanism, we initiate our analysis by formally defining the formula of the following expression:

$$\mathcal{C} = (1-\alpha^+)(1-\eta) + [1-\min(R_0^\chi, R_1^\chi)]\eta\alpha^+ - (R_0^\chi + R_1^\chi)\eta'\alpha^+. \tag{B7}$$

It is evident that the formula for $\mathcal{C}$ is continuous, which facilitates the determination of its extremum.

[**Upper bound**] The supremum occurs when:

$$\begin{aligned} &\text{term } i \to 1 \ (\alpha^+ \to 0,\ \eta \to 0) \\ &\text{term } ii \to 0 \ (\alpha^+ \to 0) \\ &\text{term } iii \to 0 \ (\alpha^+ \to 0) \\ &\Rightarrow \sup\mathcal{C} = 1. \end{aligned}$$

[**Lower bound**] The infimum occurs when:

$$\begin{aligned} &\text{term } i \to 0 \ (\alpha^+ \to 1) \\ &\text{term } ii \to 0 \ (\min(R_0^\chi, R_1^\chi) \to 1) \\ &\text{term } iii \to -1 \ (R_{0,1}^\chi \to 1,\ \eta' \to 0.5,\ \alpha^+ \to 1) \\ &\Rightarrow \inf\mathcal{C} = -1. \end{aligned}$$

The complete domain can be derived as $\mathcal{C} \in [-1, 1]$. Therefore, confirmed by extremum theorems, the strict boundary of the absolute value term can be derived as $|\mathcal{C}| \in [0, 1]$.

Furthermore, we consider the rationale behind the values of each parameter. Notably, $R_0^{\mathcal{X}}$, $R_1^{\mathcal{X}}$ and η are related to the neighborhood balance parameter α^+, which follows the relationships: $\alpha^+ \to 1 \Rightarrow R_0^{\mathcal{X}}, R_1^{\mathcal{X}} \to 1,\ \eta \to 0.5$, and conversely, $\alpha^+ \to 0 \Rightarrow R_0^{\mathcal{X}}, R_1^{\mathcal{X}} \to 0,\ \eta \to 0$. On the other hand, η' represents the proportion of heterogeneous nodes reconstructed by the GHP module, which requires to be considered separately.

In the derivation of the **upper bound**, the core assumption is that $\alpha^+ \to 0$, indicating that the neighborhood balance is broken and the nodes in the neighborhood tend to be homogeneous ($R_0^{\mathcal{X}}, R_1^{\mathcal{X}} \to 0$, $\eta \to 0$). Therefore, all parameter values in the upper bound inference are reasonable.

In the derivation of the **lower bound**, the assumption is the opposite: $\alpha^+ \to 1$ implies $R_0^{\mathcal{X}}, R_1^{\mathcal{X}} \to 1$ and $\eta \to 0.5$. In this case, term i and term $ii \to 0$, and we need to analyze the value of η', which influences term iii. If the topology reconstructed by the GHP module exhibits high heterophily, then term $iii \to -1$; otherwise, term $iii \to 0$. Based on this analysis, we conclude that these boundary conditions are practically feasible.

By inspecting equation (B6), we interpret the roles of $\mathcal{F}_{\text{LDP}}$ and $\mathcal{F}_{\text{GHP}}$. $\mathcal{F}_{\text{LDP}}$ incorporates the throttling of homogeneous nodes and the regularization of heterogeneous nodes, which correspond to the first two additive terms in the expression. In term i, $(1-\alpha^+)$ and $(1-\eta)$ represent the aggregation coefficient of homogeneous nodes and the average number of homogeneous neighbors, respectively, explicitly capturing the influence of homogeneous nodes on the results. Term ii reflects the role of heterogeneous nodes through $\eta\alpha^+$, which is further regularized by $[1-\min(R_0^{\mathcal{X}}, R_1^{\mathcal{X}})]$. Term iii integrates subtractive components associated with heterogeneous nodes, counterbalancing the first two terms to effectively reduce bias. This aligns precisely with the objective of $\mathcal{F}_{\text{GHP}}$, which aims to mitigate neighborhood imbalance by introducing additional heterogeneous nodes.

In conclusion, through the analysis of these three terms, we establish that both $\mathcal{F}_{\text{LDP}}$ and $\mathcal{F}_{\text{GHP}}$ can effectively reduce statistical parity δ_h between the embeddings of different sensitive groups.

At this stage, we quantitatively study the debaising capability of our proposed MP layer ($\mathcal{F}_{\text{LDP}} + \mathcal{F}_{\text{GHP}}$). Specifically, we train a standard GNN on four public datasets and report the impact of using different numbers of MP layers on fairness. As shown in Figure C1 (a), compared to using standard message aggregation function (0 MP layer), statistical parity is significantly reduced with only 1 to 2 MP layers. In addition, when a large number of MP layers are stacked, fairness can also be guaranteed without the occurrence of over-smoothing phenomenon like classification accuracy. This provides strong experimental evidence for our theoretical proof.

Appendix C Experiments

We conduct experiments on four real-world datasets to evaluate the performance of FairDHP, particularly focusing on accuracy and fairness. Codes can be found at GitHub [1)]. We aim to address the following research questions through empirical investigations:

RQ1: How effectively can FairDHP achieve fair node classification while optimizing the fairness-utility trade-off?
RQ2: Are the components of FairDHP essential for enhancing model performance?
RQ3: Can FairDHP maintain stable performance under varying parameter configurations?

Appendix C.1 Datasets

• **Bail** dataset originates from U.S. criminal justice records, containing demographic attributes (race, gender), criminal history, and bail decision outcomes for $100,000^+$ defendants [17]. Race and bail decision were selected as sensitive attributes and classification labels, respectively.

• **Credit** dataset comes from an important bank credit data in Hong Kong, China, including 30000 credit card holders, with 13 attributes covering payment history, credit utilization, and debt and income indicators [18]. It is mainly used for predicting future default and supports the classification research of fairness perception by combining sensitive attributes such as age.

• **German** dataset has 1000 nodes representing customers of a German bank, who are connected based on the similarity of their credit accounts [19]. Its task is to classify customers into good credit risk and bad credit risk based on their gender sensitive attribute.

• **Pokec_n** dataset is from the Pokec social network in Slovakia, including user profiles, friendship relationships, and interaction timestamps [20]. It is a normalized version of its religion with anonymous sensitive attributes, used for predicting working fields.

Appendix C.2 Baselines

• **FairGNN** [21] (2021, WSDM): The FairGNN develops an adversarial debiasing framework, which leverages limited partial sensitive and topology-aware data augmentation to eliminate biased correlations.

• **EDITS** [22] (2022, WWW): The EDITS mitigates data bias by jointly optimizing attribute distribution alignment and structural fairness regularization through trainable adjustments to both adjacency matrix and node features.

• **NIFTY** [23] (2021, UAI): The NIFTY establishes a unified paradigm through adversarial debiasing modules and graph augmentation techniques that enhance both counterfactual fairness and robustness against structural perturbations.

• **FairVGNN** [24] (2022, KDD): The FairVGNN mitigates discrimination by automatically masking sensitive-correlated features through adaptive view generation and encoder weight clamping, achieving enhanced fairness-utility trade-offs.

1) https://github.com/Wangshiyi1116/FairDHP

Table C1 Statistics of four datasets.

Dataset	Bail	Credit	German	Pokec_n
# Nodes	18,876	30,000	1,000	66,569
# Edges	321,308	1,436,858	22,242	729,129
# Features	18	13	27	59
Sensitive Attribute	Race	Age	Gender	Region
Label	Bail decision	Future default	Credit status	Working field
Avg. Degree	34.04	95.79	44.48	16.53
Avg. Hete Degree	15.79	3.83	8.49	0.73
# w/o Hete Neighbors	32	16,738	30	46,134

• **FairSIN** [4] (2024, AAAI): The FairSIN proposes a neutralization-based paradigm that can debias sensitive attributes and provide additional non-sensitive information for learning fair GNNs.
• **FairGT** [25] (2024, IJCAI): The FairGT incorporates a meticulous structural feature selection strategy and a multi-hop node feature integration method, ensuring independence of sensitive features and bolstering fairness considerations.
• **FDGNN** [1] (2025, NN): The FDGNN designs a counterfactual augmentation strategy for constructing instances with varying sensitive values while preserving the same adjacency matrices, thereby balancing the distribution of sensitive values across different groups.

Appendix C.3 GNN backbones

In our experiments, three efficient and stable GNN backbones are employed as basic encoders: **GCN** [7], **GIN** [26] and **GraphSAGE** [27]. The superior performance of these encoders has been proven in previous works [28–31].

Appendix C.4 Fairness metrics

Following the existing works [1, 16, 32], we utilize $\mathbf{y}_i \in \{0, 1\}$ as ground-truth label. At the same time, in order to simplify the expression, sensitive attribute is represented as $s_i \in \{0, 1\}$ for $v_i \in \mathcal{V}$.

Definition 1. Statistical Parity (SP)

Statistical parity [33] (also called demographic parity) requires that the predicted outcomes of a model are statistically independent of sensitive attributes. Formally, for a binary classifier with prediction $\hat{\mathbf{y}}_i$ and sensitive attribute s_i, statistical parity denotes as:

$$\Delta\mathsf{SP} = \Big|\mathbb{P}(\hat{\mathbf{y}}_i = 1 \mid s_i = 1) - \mathbb{P}(\hat{\mathbf{y}}_i = 1 \mid s_i = 0)\Big|, \tag{C1}$$

where $s_i = 1$ denote distinct groups.

Definition 2. Equal Opportunity (EO)

Equal opportunity [34] is a fairness criterion that requires true positive rates to be equal across protected groups. Formally, for ground-truth label $\mathbf{y}_i$, equal opportunity denotes as:

$$\Delta\mathsf{EO} = \Big|\mathbb{P}(\hat{\mathbf{y}}_i = 1 \mid \mathbf{y}_i = 1, s_i = 1) - \mathbb{P}(\hat{\mathbf{y}}_i = 1 \mid \mathbf{y}_i = 1, s_i = 0)\Big|, \tag{C2}$$

Appendix C.5 Implementation details

We employ a five-fold cross-validation methodology for each experiment to ensure the validity of the experimental results. Our FairDHP architecture was developed within PyTorch, leveraging 100 GB VRAM NVIDIA A100 for all computational experiments.

Appendix C.6 RQ1: Results on node classification

To validate the effectiveness of FairDHP, we compare our method with SOTA methods on node classification task. As demonstrated in Table C2, classification accuracy, SP, and EO metrics are reported, along with the fairness-utility trade-off values (the pink column) calculated as follow:

$$\Delta TO = ACC - \lambda(\Delta\mathsf{SP} + \Delta\mathsf{EO}), \tag{C3}$$

where $\lambda = 1$.

Cross-architecture evaluations reveal that FairDHP achieves a great improvement of fairness metrics compared to baseline methods, while retaining a high classification accuracy across GNN backbones and four datasets. Specifically, FairDHP obtains optimal fairness metrics reductions in SP and EO on both the Bail and Pokec_n datasets, surpassing contemporary fairness-aware graph learning methods. Concurrently, it maintains competitive classification accuracy with a improvement over vanilla GNN backbones.

However, we observe that FairDHP exhibits suboptimal performance on the Credit dataset, particularly when compared with FairSIN. This performance discrepancy manifests in two critical aspects: (1) a 6.17 average reduction in trade-off; (2) significantly higher variance ($\pm$1.51%) across repeated trials compared to FairSIN's stable performance ($\pm$0.70%). According

Table C2 Comparison of model performance (mean ± standard deviation). (bold: best)

Encoder	Method	German				Credit				Bail				Pokec_n			
		ACC (↑)	ΔSP(↓)	ΔEO(↓)	Δ*TO*(↑)	ACC (↑)	ΔSP(↓)	ΔEO(↓)	Δ*TO*(↑)	ACC (↑)	ΔSP(↓)	ΔEO(↓)	Δ*TO*(↑)	ACC (↑)	ΔSP(↓)	ΔEO(↓)	Δ*TO*(↑)
GCN	vanilla	72.28±1.52	32.43±10.29	24.69±7.74	15.16	74.13±0.04	12.44±0.06	10.24±0.09	51.45	87.55±0.54	6.85±0.47	5.26±0.78	75.44	68.55±0.51	3.75±0.94	2.93±1.15	61.87
	FairGNN	69.68±0.30	3.49±2.15	3.40±2.15	62.79	73.41±1.24	12.64±2.11	10.41±2.03	50.36	82.94±1.67	6.90±0.17	4.65±0.14	71.39	67.36±2.06	3.29±2.95	2.46±2.64	61.61
	EDITS	71.60±0.89	4.05±4.48	3.89±4.23	63.66	73.51±0.30	10.90±1.22	8.75±1.21	53.86	84.49±2.27	6.64±0.39	7.51±1.20	70.34	OOM	OOM	OOM	-
	NIFTY	69.92±1.14	5.73±5.25	5.08±4.29	59.11	73.45±0.06	11.68±0.07	9.39±0.07	52.38	82.36±3.91	5.78±1.29	4.72±1.08	71.86	67.24±0.49	1.22±0.94	2.79±1.24	63.23
	FairVGNN	70.16±0.86	1.71±1.68	0.88±0.58	67.57	**78.04±0.33**	5.02±5.22	3.60±4.31	69.42	84.73±0.46	6.53±0.67	4.95±1.22	73.25	66.10±1.45	1.69±0.79	1.78±0.70	62.63
	FairSIN	70.08±0.16	0.22±0.43	0.02±0.04	69.84	77.87±0.01	**0.50±0.70**	**0.25±0.34**	**77.12**	87.67±0.26	4.56±0.75	2.79±0.89	80.32	69.34±0.32	**0.57±0.19**	0.43±0.41	68.34
	FairGT	69.99±0.11	0.13±0.24	0.09±0.19	69.77	74.64±2.86	6.29±3.21	4.43±2.31	63.92	91.52±0.19	1.7±0.53	1.67±0.29	88.15	66.37±3.75	1.0±0.31	1.07±1.15	64.30
	FDGNN	69.6±1.01	0.39±0.51	0.29±0.37	68.92	75.14±2.9	5.6±4.11	4.51±3.77	65.03	92.73±0.08	**0.74±0.39**	1.77±0.84	**90.22**	OOM	OOM	OOM	-
	OURS	**70.08±0.16**	**0.13±0.26**	**0.0±0.0**	**69.95**	77.95±0.72	4.12±1.51	2.88±0.97	70.95	**92.74±0.1**	2.9±0.08	**1.1±0.18**	88.74	**69.35±0.77**	0.58±0.3	**0.38±0.31**	**68.39**
GIN	vanilla	**72.96±1.14**	13.94±6.81	9.08±6.04	49.94	77.39±1.00	5.66±1.82	3.47±1.72	68.26	83.52±0.87	7.55±0.51	6.17±0.69	69.80	69.25±1.75	3.71±1.20	2.55±1.52	62.99
	FairGNN	72.24±1.44	6.88±4.42	2.06±1.46	63.30	70.33±5.50	4.67±3.06	3.94±1.49	61.72	77.90±2.21	6.33±1.49	4.74±1.64	66.83	67.10±3.25	3.82±2.44	3.62±2.78	59.66
	EDITS	72.08±0.66	0.86±0.76	1.72±1.14	69.50	74.07±0.98	14.11±14.45	15.40±15.76	44.56	73.74±5.12	6.71±2.35	5.98±3.66	61.05	OOM	OOM	OOM	-
	NIFTY	69.92±3.64	5.26±3.24	5.34±5.67	59.32	75.59±0.66	7.09±4.62	6.22±3.26	62.28	74.46±9.98	5.57±1.11	3.41±1.43	65.48	66.37±1.51	3.84±1.05	3.24±1.60	59.29
	FairVGNN	70.16±0.32	0.43±0.54	0.34±0.41	69.39	78.18±0.20	2.85±2.01	1.72±1.80	73.61	83.86±1.57	5.67±0.76	5.77±0.76	72.42	68.37±0.97	1.88±0.99	1.24±1.06	65.25
	FairSIN	70.40±0.80	0.30±0.29	0.19±0.33	69.91	77.88±0.12	0.36±0.72	**0.23±0.45**	77.29	86.52±0.48	4.35±0.71	4.17±0.96	78.00	69.58±0.57	1.11±0.31	0.97±0.59	67.50
	FairGT	70.8±0.43	0.38±0.39	0.63±0.21	69.79	77.9±0.81	0.73±0.52	1.25±0.75	75.92	77.2±2.24	4.17±2.12	4.18±2.71	68.85	69.55±0.45	1.0±0.98	**0.77±0.37**	67.78
	FDGNN	69.84±0.19	0.27±0.31	0.34±0.41	69.23	78.58±0.7	1.07±0.41	0.95±0.95	76.56	85.74±1.27	3.59±0.82	3.35±2.49	78.80	OOM	OOM	OOM	-
	OURS	72.08±0.93	**0.18±0.32**	**0.17±0.23**	**71.73**	**79.0±0.43**	**0.3±0.24**	0.61±0.36	**78.09**	**86.7±0.28**	**3.44±1.68**	**2.88±1.32**	**80.38**	**69.6±0.35**	**0.72±0.31**	0.85±0.97	**68.03**
SAGE	vanilla	72.12±1.76	20.33±11.82	14.86±10.96	36.93	76.77±0.68	14.31±6.55	11.78±5.71	50.68	88.13±1.12	1.13±0.48	2.61±1.16	84.39	69.03±0.77	3.09±1.29	2.21±1.60	63.73
	FairGNN	70.64±0.74	7.65±8.07	4.18±4.86	58.81	75.29±1.62	6.17±5.57	5.06±4.46	64.06	87.68±0.73	1.94±0.82	1.72±0.70	84.02	67.03±2.61	2.97±1.28	2.06±3.02	62.00
	EDITS	71.68±1.25	8.42±7.35	5.69±2.16	57.57	74.13±0.59	11.34±6.36	9.38±5.39	53.41	84.42±2.87	3.74±3.54	4.46±3.50	76.22	OOM	OOM	OOM	-
	NIFTY	69.60±1.50	7.74±7.80	5.17±2.38	56.69	74.39±1.35	10.65±1.65	8.10±1.91	55.64	84.11±5.49	5.74±0.38	4.07±1.28	74.30	68.48±1.11	3.84±1.05	3.90±2.18	60.74
	FairVGNN	70.00±0.25	1.36±1.90	1.22±1.49	67.42	79.94±0.30	4.94±1.10	2.39±0.71	72.61	88.41±1.29	1.14±0.67	1.69±1.13	85.58	68.50±0.71	1.12±0.98	1.13±1.02	66.25
	FairSIN	70.40±0.62	0.32±0.25	**0.08±0.33**	70.00	78.91±0.61	1.38±1.71	**0.79±0.94**	76.74	88.74±0.42	0.58±0.60	1.49±0.34	86.67	69.12±1.16	1.04±0.83	1.04±0.42	67.04
	FairGT	68±0.11	0.25±0.22	0.31±0.37	67.44	78.52±0.53	2.39±1.94	1.34±1.11	74.79	87.92±0.47	0.88±0.79	1.44±0.88	85.60	66.89±1.28	0.74±0.65	1.36±0.79	64.79
	FDGNN	70.00±0.44	0.34±0.37	0.46±0.25	69.20	78.54±1.06	0.74±0.63	0.92±0.64	76.88	86.43±0.24	0.59±0.71	1.54±1.07	84.30	OOM	OOM	OOM	-
	OURS	**72.56±0.54**	**0.17±0.34**	0.19±0.38	**72.20**	**79.98±0.52**	**1.46±1.16**	0.72±0.85	**77.80**	**90.76±1.01**	**0.4±0.3**	**1.4±0.96**	**88.96**	**69.41±0.81**	**0.61±0.81**	**0.87±0.56**	**67.93**

OOM: Out of memory.

to Table C1, we note that the Credit dataset has an astonishing proportion of homo/heterogeneous neighbors (reach to 24:1). This makes it difficult for both LDP and GHP to find suitable heterogeneous neighbors for a node with imbalance neighborhood. In contrast, FairSIN adopts a generative approach, not only allowing nodes rich in heterogeneous neighbors to transfer their knowledge to other nodes but also increasing the stability of the model.

Appendix C.7 RQ2: Ablation study

Table C3 Comparison of variants performance (mean ± standard deviation). (bold: best).

Encoder	Method	SOT	LDP	GHP	German			Credit			Bail			Pokec_n		
					ACC (↑)	ΔSP(↓)	ΔEO(↓)	ACC (↑)	ΔSP(↓)	ΔEO(↓)	ACC (↑)	ΔSP(↓)	ΔEO(↓)	ACC (↑)	ΔSP(↓)	ΔEO(↓)
GCN	w/o-S		✓	✓	70.02±0.64	0.81±1.61	0.17±0.34	72.49±4.12	5.29±3.73	3.59±3.29	87.47±0.28	5.25±1.94	5.52±2.5	65.5±2.06	0.94±0.52	1.25±0.89
	w/o-D	✓		✓	70.24±0.48	1.06±2.13	0.71±1.43	71.93±3.45	7.32±3.68	5.26±3.44	85.07±0.68	4.25±1.81	4.91±1.35	66.85±0.8	2.74±0.32	1.99±0.6
	w/o-H	✓	✓		70.08±0.53	1.40±0.80	0.38±0.42	76.34±2.46	5.79±2.31	4.81±2.26	90.15±0.19	3.44±1.3	4.19±1.87	66.09±0.75	0.93±0.72	1.76±1.16
	FairDHP	✓	✓	✓	**70.08±0.16**	**0.13±0.26**	**0.0±0.0**	**77.95±0.72**	**4.12±1.51**	**2.88±0.97**	**92.74±0.1**	**2.90±0.08**	**1.1±0.18**	**69.35±0.77**	**0.58±0.30**	**0.38±0.31**
GIN	w/o-S		✓	✓	70.88±0.12	0.43±0.85	0.36±0.71	78.33±1.44	2.31±2.05	1.71±1.28	86.37±1.09	**3.23±0.82**	3.35±2.49	69.46±0.38	1.0±0.88	0.79±0.32
	w/o-D	✓		✓	70.00±0.9	0.85±0.7	0.71±0.43	77.11±1.37	2.24±1.17	1.84±1.02	82.02±0.66	4.40±2.05	4.86±2.56	68.97±0.91	1.90±1.13	1.03±0.8
	w/o-H	✓	✓		69.92±0.16	0.64±1.27	0.57±1.13	72.93±1.7	1.22±0.61	0.93±0.46	82.11±0.28	3.84±1.34	2.9±1.29	67.55±0.24	0.92±0.76	**0.78±0.62**
	FairDHP	✓	✓	✓	**72.08±0.93**	**0.18±0.32**	**0.17±0.23**	**79.0±0.43**	**0.30±0.24**	**0.61±0.36**	**86.70±0.28**	3.44±1.68	**2.88±1.32**	**69.6±0.35**	**0.72±0.31**	0.85±0.97
SAGE	w/o-S		✓	✓	70.16±0.2	0.51±0.64	0.36±0.71	75.83±3.24	4.92±3.35	3.3±2.94	90.10±0.41	1.05±0.67	1.41±0.91	69.2±0.52	2.19±0.57	0.97±0.54
	w/o-D	✓		✓	69.76±0.48	0.97±1.95	0.78±1.55	78.52±0.53	2.39±1.94	1.34±1.11	88.15±0.34	1.08±1.04	2.72±0.63	67.51±0.36	0.78±0.2	2.14±0.69
	w/o-H	✓	✓		69.92±0.16	0.64±0.85	0.71±0.87	77.92±0.31	3.94±2.4	2.66±2.05	88.61±0.74	0.80±0.48	3.47±2.23	69.41±0.93	1.65±0.57	**0.67±0.38**
	FairDHP	✓	✓	✓	**72.56±0.54**	**0.17±0.34**	**0.19±0.38**	**79.98±0.52**	**1.46±1.16**	**0.72±0.85**	**90.76±1.01**	**0.40±0.30**	**1.40±0.96**	**69.41±0.81**	**0.61±0.81**	0.87±0.56

In order to better demonstrate the role of each component in FairDHP, we conduct ablation experiments. Specifically, we design three ablated models by removing individual modules:

- w/o-S: FairDHP without sensitive operator;
- w/o-D: FairDHP without local disentangled propagation;
- w/o-H: FairDHP without global heterophily-aware propagation.

To ensure experimental feasibility, the w/o-S utilizes the identity matrix $\mathbb{E}$ in place of SOT. Ablation results are shown in Table C3. We have the following observations.

(1) The accuracy and fairness indicators of w/o-S show significant declines across three GNN backbones and four datasets. Taking GCN encoder as an example, a 6.0% degradation in classification accuracy can be observed on the Credit dataset, while the Bail dataset exhibits 1.3 and 4.4 improvements in SP and EO metrics respectively. This is due to the sensitive operator is extremely important for our method, as it not only assists in the generation of LDP and GHP adjacency matrices, but also guides the fusion of subsequent fair embeddings.

(2) Similar to the above, the w/o-D and the w/o-H cannot fully complete the task for debiasing. The LDP and the GHP constitute dual pillars of FairDHP's fairness-aware MP framework, collaboratively addressing bias mitigation. Losing either of these two MPs, variants naturally reduce the ability of mitigating the harm caused by sensitive attributes.

In addition, we design some comparison methods for replaceable technologies within component.

Effectiveness analysis of LAd. In Eq. (A4), LAd is utilized to calculate the discrepancy between embeddings h and $\bar{h}$. Here, we replace it with cosine similarity and Euclidean distance. As shown in Figure C1 (b), these two methods

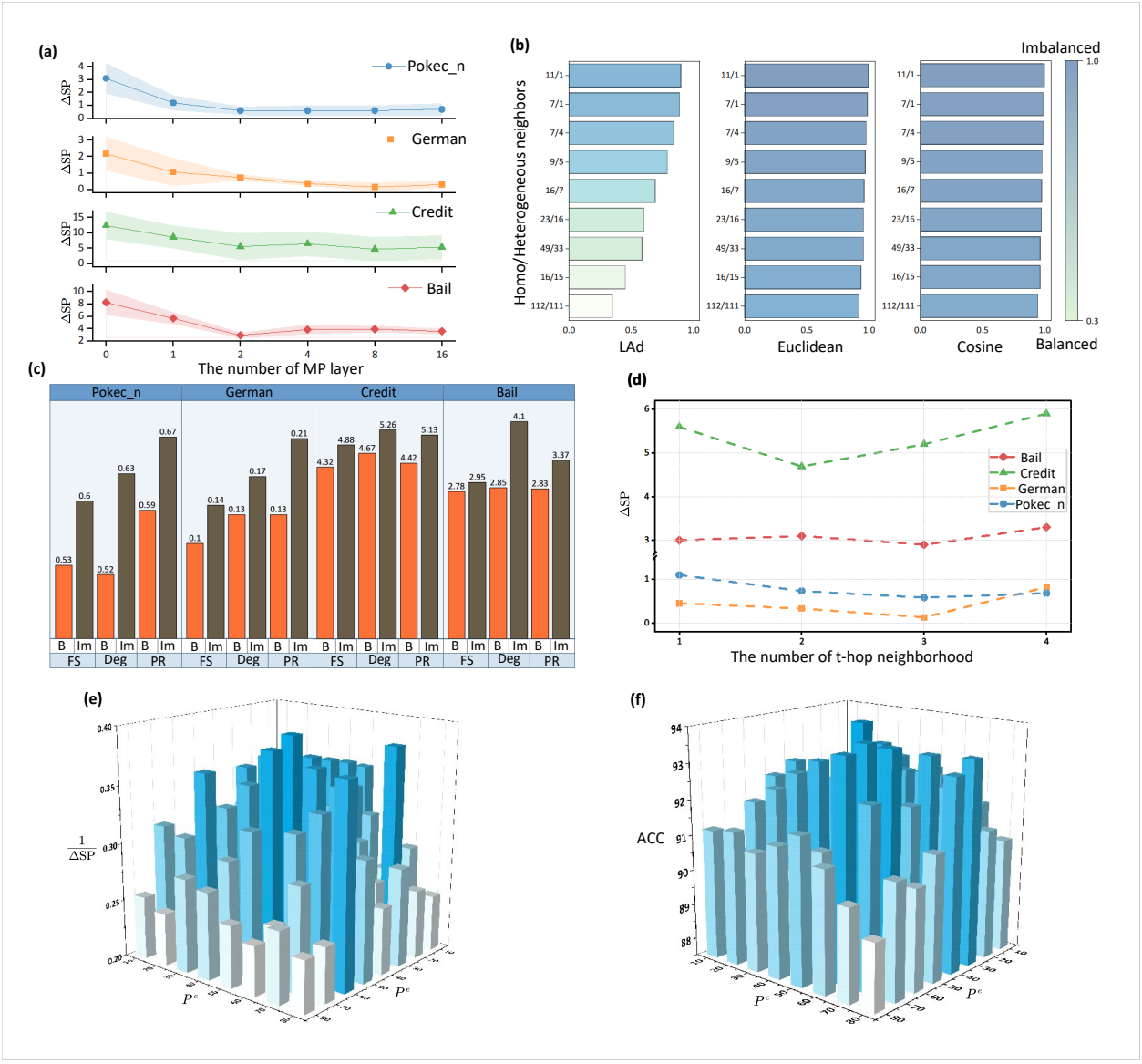

Figure C1 Experiments for FairDHP: (a) the experimental support theoretical proof, (b) ablation study for fairness sensitivity with different metrics, (c) ablation study for fairness sensitivity with different calculation formulas, (d) the impact of number of t-hop neighborhood, (e) the impact of number of regions on fairness, and (f) the impact of number of regions on accuracy.

can only achieve a coarse-grained comparison between embeddings. However, LAd enables fine-grained comparisons across different nodes, thereby revealing the neighborhood balance properties of distinct nodes.

Effectiveness analysis of counter-sensitive feature. Furthermore, we broke through the framework of embedding comparison and focus on directly using metrics that characterize neighborhood properties to measure the balance of node neighborhood. Specifically, we use 1/degree and 1/pagerank as alternative for fairness sensitivity and sample subgraphs with relatively balanced and imbalanced neighborhoods for experimentation. As shown in Figure C1 (c), the two replacement methods have significant differences in their effectiveness under different neighborhood conditions. Our proposed FS can better adapt to imbalanced neighborhoods.

Appendix C.8 RQ3: Parameter sensitivity study

To comprehensively evaluate FairDHP's performance, we investigate the impact of three crucial parameters: the number of neighbor order t, the number of regions P^e and P^c. In addition, we present the optimal hyperparameters used in the classification experiment in Table C4.

We conduct comprehensive experiments across four datasets (Pokec_n, German, Credit, Bail) using GCN as the backbone architecture, with neighborhood orders varied within the range [1, 2, 3, 4]. As quantified in Figure C1 (d), the fairness curves for Pokec_n, German, and Bail exhibit marginal sensitivity to neighborhood depth, achieving peak fairness at 3-hop neighborhoods. Notably, the Credit dataset demonstrates distinct behavior: optimal performance emerges at 2-hop

Table C4 Hyperparameters by encoder and dataset.

	Encoder	epochs	lr	MP	hidden	t	P^e	P^c
German	GCN	50	0.001	1	32	2	50	60
	GIN	50	0.01	1	32	2	60	60
	SAGE	100	0.001	2	32	3	40	40
Credit	GCN	100	0.01	2	64	3	60	70
	GIN	100	0.01	2	64	3	60	60
	SAGE	100	0.01	2	32	3	60	60
Bail	GCN	50	0.01	1	32	2	50	50
	GIN	50	0.01	1	32	2	40	40
	SAGE	50	0.01	1	32	2	50	60
Pokec_n	GCN	100	0.001	1	64	2	60	60
	GIN	200	0.001	2	32	3	60	40
	SAGE	100	0.001	1	64	2	70	70

(ΔSP = 4.69), while 4-hop configurations induce significant degradation (ΔSP = 5.9). This phenomenon is attributed to Credit's high edge density, where 3-hop neighborhoods already encompass the entire graph, so higher-order neighborhood would not bring more heterogeneous nodes.

To investigate the synergistic effects of hyperparameters P^e and P^c on FairDHP, we conduct experiments on the Bail dataset using GCN as the backbone architecture. Both P^e and P^c are selected from the set [10, 20, 30, 40, 50, 60, 70 ,80]. As depicted in Figure C1 (e) and (f), we analyze the impact of the number of region on both accuracy and statistical parity.

In order to better compare the impact of different numbers of regions on fairness, we present statistical parity in reciprocal form. From (e), we observe that our FairDHP has a sensitive range for the number of regions. It is worth noting that larger or smaller P^e and P^c values are not suitable for our method. This phenomenon can be attributed to the interaction mechanism between region granularity and information propagation. When employing smaller P^e and P^c values, nodes are clustered into coarser regions containing substantial heterogeneous neighbors, inevitably introducing noise during message passing. Meanwhile, larger parameter values can lead to excessive homophily within class clusters, making it difficult to generate cross-region shared nodes and resulting in insufficient heterophily propagation. From (f), the selection of these values also has an impact on producing high-quality embeddings of downstream tasks. The sensitivity of fairness and classification performance to the number of regions is basically consistent, which proves that our proposed FairDHP can achieve the fairness-utility trade-off.

References

1 Zhang G, Yuan G, Cheng D, et al. Disentangled contrastive learning for fair graph representations. Neural Networks, 2025, 181: 106781
2 Russell C, Kusner M J, Loftus J R, et al. When worlds collide: Integrating different counterfactual assumptions in fairness. In: Proceedings of Adv Neural Inf Process Syst, 2017
3 Kusner M J, Loftus J R, Russell C, et al. Counterfactual fairness. In: Proceedings of Adv Neural Inf Process Syst, 2017
4 Yang C, Liu J, Yan Y, et al. FairSIN: Achieving fairness in graph neural networks through sensitive information neutralization. In: Proceedings of AAAI Conf Artif Intell, 2024. 9241–9249
5 Chen F, Shi T, Duan S, et al. Diffusion least logarithmic absolute difference algorithm for distributed estimation. Signal Process, 2018, 142: 423–430
6 Huang C, Wang Y, Jiang Y, et al. Flow2GNN: Flexible two-way flow message passing for enhancing gnns beyond homophily. IEEE Trans Cybern, 2024, 54: 6607–6618
7 Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. In: Proceedings of Int Conf Learn Represent, 2017
8 Fu C, Liu G, Yuan K, et al. Nowhere to h^{2}ide: Fraud detection from multi-relation graphs via disentangled homophily and heterophily identification. IEEE Trans Knowl Data Eng, 2025, 37: 1380–1393
9 Creager E, Madras D, Jacobsen J, et al. Flexibly fair representation learning by disentanglement. In: Proceedings of Int Conf Learn Represent, 2019
10 Oh C, Won H, So J, et al. Learning fair representation via distributional contrastive disentanglement. In: Proceedings of ACM Conf Knowl Discov Data Min, 2022
11 Ding M, Kong K, Li J, et al. VQ-GNN: A universal framework to scale up graph neural networks using vector quantization. In: Proceedings of Adv Neural Inf Process Syst, 2021. 6733–6746
12 Seghair T, Besbes O, Abdellatif T, et al. VQ-VGAE: Vector quantized variational graph auto-encoder for unsupervised anomaly detection. In: Proceedings of IEEE International Conference on Big Data, 2024. 2370–2375
13 Aurenhammer F. Voronoi diagrams - A survey of a fundamental geometric data structure. ACM Comput Surv, 1991, 23: 345–405
14 Liu S, He D, Yu Z, et al. Beyond homophily: Neighborhood distribution-guided graph convolutional networks. Expert Syst Appl, 2025, 259: 125274
15 Chen A, Rossi R A, Park N, et al. Fairness-aware graph neural networks: A survey. ACM Trans Knowl Discov Data, 2024, 18: 138
16 Köse Ö D, Shen Y. FairGAT: Fairness-aware graph attention networks. ACM Trans Knowl Discov Data, 2024, 18: 164
17 Koszela-Kulinska J, Michalski R. The effects of the anthropological race, gender and location of verbal-pictorial stimuli on the usability of visual information conveyance. In: Proceedings of Int Conf Hum-Comput Interact, 2015. 441–451
18 Yeh I, Lien C. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Syst Appl, 2009, 36: 2473–2480
19 Dua D, Graff C, et al. Uci machine learning repository. 2017
20 Takac L, Zabovsky M. Data analysis in public social networks. In: Proceedings of Int Sci Conf Int Workshop Present Day Trends Innov, 2012
21 Dai E, Wang S. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In: Proceedings of ACM Int Conf Web Search Data Min, 2021. 680–688

22 Dong Y, Liu N, Jalaian B, et al. EDITS: Modeling and mitigating data bias for graph neural networks. In: Proceedings of ACM Web Conf, 2022. 1259–1269
23 Agarwal C, Lakkaraju H, Zitnik M. Towards a unified framework for fair and stable graph representation learning. In: Proceedings of Uncertain Artif Intell, 2021. 2114–2124
24 Wang Y, Zhao Y, Dong Y, et al. Improving fairness in graph neural networks via mitigating sensitive attribute leakage. In: Proceedings of ACM SIGKDD Conf Knowl Discov Data Min, 2022. 1938–1948
25 Luo R, Huang H, Yu S, et al. FairGT: A fairness-aware graph transformer. In: Proceedings of Int Joint Conf Artif Intell, 2024. 449–457
26 Xu K, Hu W, Leskovec J, et al. How powerful are graph neural networks? In: Proceedings of Int Conf Learn Represent, 2019
27 Hamilton W L, Ying Z, Leskovec J. Inductive representation learning on large graphs. In: Proceedings of Adv Neural Inf Process Syst, 2017
28 Guo Q, Yang X, Zhang F, et al. Perturbation-augmented graph convolutional networks: A graph contrastive learning architecture for effective node classification tasks. Eng Appl Artif Intell, 2024, 129: 107616
29 Cong H, Sun Q, Yang X, et al. Enhancing graph convolutional networks with progressive granular ball sampling fusion: A novel approach to efficient and accurate GCN training. Inf Sci, 2024, 676: 120831
30 Guo Q, Yang X, Guan W, et al. Robust graph mutual-assistance convolutional networks for semi-supervised node classification tasks. Inf Sci, 2025, 694: 121708
31 Guo Q, Yang X, Li M, et al. Collaborative graph neural networks for augmented graphs: A local-to-global perspective. Pattern Recognit, 2025, 158: 111020
32 Kose O D, Shen Y. FairWire: Fair graph generation. In: Proceedings of Adv Neural Inf Process Syst, 2024. 124451–124478
33 Dwork C, Hardt M, Pitassi T, et al. Fairness through awareness. In: Proceedings of Innov Theor Comput Sci, 2012. 214–226
34 Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: Proceedings of Adv Neural Inf Process Syst, 2016