SCIENCE CHINA
Information Sciences

• LETTER •

March 2026, Vol. 69, Iss. 3, 139102:1–139102:2
https://doi.org/10.1007/s11432-025-4722-2

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*

Received 20 April 2025/Revised 5 October 2025/Accepted 28 November 2025/Published online 15 January 2026

Citation Wang Y G, Yang X B, Liu K Y, et al. Topology matters: achieving fairness in graph neural networks through heterophily propagation. Sci China Inf Sci, 2026, 69(3): 139102, https://doi.org/10.1007/s11432-025-4722-2

Graph neural networks (GNNs) have gained significant attention due to their immense potential in graph-driven applications. However, GNNs are susceptible to sensitive attributes (e.g., race, gender and religion), which may result in unfair decisions for certain sensitive groups. Meanwhile, network homophily creates a biased tendency for nodes to form homogeneous connections (edges within the same sensitive group), while lacking heterogeneous connections (edges across groups) [1]. This topology imbalance is further amplified by GNNs' message-passing mechanism: aggregation functions disproportionately reinforce patterns from majority groups, while minority groups remain underrepresented in learned node embeddings [2].

Existing debias efforts encompass different stages of GNN training and can be categorized into pre-processing, in-training, and post-processing. Pre-processing debiasing of features and topology serves as a fundamental approach for enhancing fairness in GNNs [3]. Actually, more researchers introduce constraints or regularization terms in the objective function, guiding the model to learn fair and unbiased embeddings during the training process [4]. Additionally, several approaches leverage the output embeddings of a graph neural network, incorporating filters or removing information related to sensitive attributes from these embeddings, thereby eliminating their biases [5].

Nevertheless, these strategies often neglect the nature of bias propagation inherent to message-passing mechanisms, and a preliminary experiment further quantified the inherent bias. As shown in Figure 1(a), topology diversity in neighborhoods (i.e., balanced homo/heterogeneous connections) is critical to mitigating bias propagation. However, standard aggregation functions fail to take into account such diversity, necessitating a redesigned message-passing mechanism. Therefore, (i) assessing the balance of homo/heterogeneous neighbors for nodes, and (ii) designing a novel message-passing mechanism to address this imbalance, have become critical challenges. These challenges demand a unified solution that not only quantifies topology bias at the neighborhood level but also adaptively adjusts aggregation strategies to counteract it.

Methods. To tackle these challenges, we propose a novel framework, FairDHP, that achieves fairness-aware graph learning. As shown in Figure 1(b), FairDHP consists of two core modules, the sensitivity operator module to measure the balance of homo/heterogeneous neighbors and the adaptive message-passing mechanism module to expand the input of heterogeneous information.

Sensitivity operator module. Specifically, the sensitivity operator measures neighborhood balance by comparing the original and counterfactual embeddings, both of which undergo message aggregation. The process consists of four steps for each node: counter-sensitive features are obtained by aggregating all heterogeneous node features within its t-hop neighborhood; the original and counter-sensitive embeddings are constructed by neighborhood aggregation of the original and counter-sensitive features, respectively; fairness sensitivity is computed by determining the logarithmic absolute difference across all dimensions of the two embeddings, providing measures of both balance and the node's susceptibility to sensitive attributes; the sensitivity operator is generated by diagonalizing the fairness sensitivity and normalizing with node degree, which guides the integration of homogeneous and heterogeneous information in subsequent processes.

Adaptive message-passing mechanism module. To mitigate topology bias, we introduce two supplementary message-passing mechanisms, local disentangled propagation (LDP) and global heterophily-aware propagation (GHP).

LDP. In undirected graph-based message-passing systems, information propagates bidirectionally through edges, amplifying bias under neighborhood imbalance where homogeneous signals dominate. To address this, we implement an edge disentanglement strategy that separates undirected connections into in-flow and out-flow. Additionally, we design a set of message passing strategies to adapt to these dual flows. For homogeneous connections, we apply throttling based on the sensitivity operator, while for heterogeneous connections, we regularize excessive aggregation to prevent the introduction of noise. These objectives are achieved by modifying the edge weights in the adjacency matrix.

GHP. Given that the neighborhood-centric paradigm becomes suboptimal when local heterogeneous neighbors are statistically insufficient or entirely absent, we extend the receptive field of heterogeneous nodes from local neighborhoods to the entire graph. For

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

© Science China Press 2026

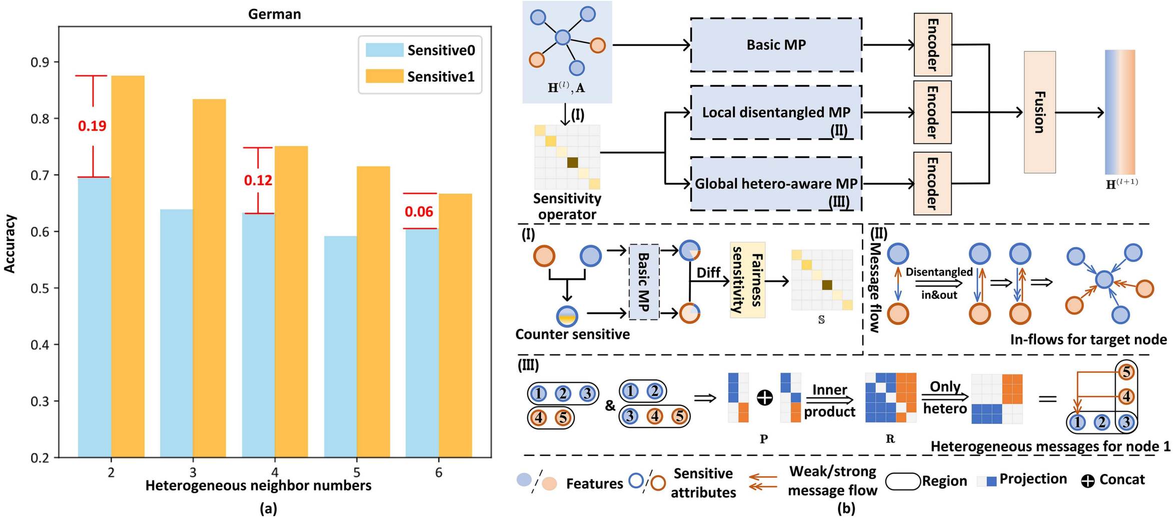

Figure 1 (Color online) Preliminary experiment (a) and overall framework (b) of the proposed FairDHP.

this purpose, we utilize feature-driven approaches multiple times to establish latent neighborhood relationships among nodes. The resulting projection matrix assigns nodes to multiple sets, with all nodes within each set having the potential to become neighbors. By concatenating projection matrices generated at multiple times, intersections are formed between sets, with shared nodes acting as bridges to facilitate information flow between sets. To further supplement heterogeneous information, we only retain heterogeneous nodes as neighbors and reconstruct the adjacency matrix. In practice, the refined adjacency matrices from LDP and GHP are deployed in the GNN framework to generate node embeddings, with further details provided in Appendix A.

Theoretical analysis. Under reasonable assumptions of finite-valued embeddings and connection homophily, we theoretically prove that the message-passing layer composed of LDP and GHP can reduce the statistical parity.

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

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

where L is the Lipschitz constant of nonlinear activation, $\sigma_{max}(\cdot)$ is the largest singular value of parameter matrix.

The absolute value accumulation term, serving as the disparity coefficient at the l-th layer, fundamentally governs the propagation trajectory of disparity metrics at the $(l+1)$-th layer.

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

$$\begin{aligned}&|(1-\alpha^+)(1-\eta)+[1-\min(R_0^\chi,R_1^\chi)]\eta\alpha^+\\ &-(R_0^\chi+R_1^\chi)\eta'\alpha^+|\in[0,1].\end{aligned} \quad (2)$$

The proofs of Lemma 1 and Theorem 1 are presented in Appendix B. In addition, the parameter matrices undergo spectral normalization of singular values, formally constrained as $\sigma_{max}(\mathbf{W}^l) \leqslant 1$ through singular value truncation. Furthermore, we integrate Softmax operators into the output layers, thereby theoretically guaranteeing the boundedness of Δ_c^{l+1} and Δ_z^{l+1}. Overall, through the analysis of these three items, we can conclude that LDP and GHP can effectively reduce statistical parity between the embeddings of different sensitive groups.

Results. Experiments were conducted on a criminal records dataset Bail, a social network dataset Pokec, and two banking financial datasets, Credit and German. On different GNN encoders, compared to other baseline methods, our model achieved the best results in terms of fairness-utility trade-off. In addition, we conducted ablation and parameter sensitivity experiments to demonstrate the effectiveness of the model components and their generalization ability. A dedicated experiment on the message-passing layer was implemented to further support our theoretical analysis. Details on these experiments can be found in Appendix C.

Discussion. The framework we proposed, FairDHP, further promotes the development of GNN in the field of fairness. Compared to other debiased methods, FairDHP makes two major contributions: (i) it can work well to measure the balance of the neighborhood; (ii) it can adaptively supplement heterogeneous information for message aggregation. Overall, FairDHP achieves a trade-off between fairness and prediction accuracy, and can be regarded as an end-to-end framework adaptable to various GNN encoders.

There is still room for improvement in the study. In future research, we intend to explore more robust and stable solutions within this fairness framework. Specifically, we will investigate advanced techniques such as generative neighbor synthesis or adaptive sampling.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62506145, 42402239).

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Köse Ö D, Shen Y. FairGAT: fairness-aware graph attention networks. ACM Trans Knowl Discov Data, 2024, 18: 164
2 Chen A, Rossi R A, Park N, et al. Fairness-aware graph neural networks: a survey. ACM Trans Knowl Discov Data, 2024, 18: 1–23
3 Kose O D, Shen Y. FairWire: fair graph generation. In: Advances in Neural Information Processing Systems 37, 2024. 124451–124478
4 Yang C, Liu J, Yan Y, et al. FairSIN: achieving fairness in graph neural networks through sensitive information neutralization. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2024. 9241–9249
5 Zhang G, Yuan G, Cheng D, et al. Disentangled contrastive learning for fair graph representations. Neural Netws, 2025, 181: 106781