Degree-Related Bias in Link Prediction

Yu Wang
Vanderbilt University
yu.wang.1@vanderbilt.edu

Tyler Derr
Vanderbilt University
tyler.derr@vanderbilt.edu

***Abstract*—Link prediction is a fundamental problem for network-structured data and has achieved unprecedented success in many real-world applications. Despite the significant progress being made towards improving its performance by characterizing underlined topological patterns or leveraging representation learning, few works have focused on the imbalanced performance among nodes of different degrees. In this paper, we propose a novel problem, degree-related bias and evaluation bias, on link prediction with an emphasis on recommender system applications. We first empirically demonstrate the performance difference among nodes with different degrees and then theoretically prove that Recall is an unbiased evaluation metric compared with F1, NDCG and Precision. Furthermore, we show that under the unbiased evaluation metric Recall, low-degree nodes tend to have higher performance than high-degree nodes in link prediction.**

***Index Terms*—Link prediction, degree-related bias, the node-centric evaluation metric**

I. Introduction

Graph-structured data is omnipresent in various fields, such as biology, chemistry, social media, and transportation [1], [2]. Link prediction, as one of the most important graph-related tasks, has become a central problem and finds its applications in predicting drug interactions, recovering knowledge graphs, and recommendations [3], [4].

As well-known in many graphs (e.g. citation graphs and social networks, etc.), node degree usually follows a power-law distribution. While the degree of major nodes is relatively small, few nodes on the long tail have significantly high-degree. Existing works [5]–[7] have shown that such power-law distributed node degree may hurt the performance of GNNs in node classification. Specifically, nodes with higher degrees are much more likely to own labeled neighbors compared with lower-degree ones and by message-passing mechanism, these nodes participate more frequently in the optimization and their learned representations are more predictive of their ground-truth labels. However, we argue that this conclusion does not hold in the task of link prediction.

On one hand, in link prediction, we are not given any golden label and hence message-passing may not cause imbalanced training/optimization between nodes of high and low degree. On the other hand, even given golden labels in link prediction, then each unique node would correspond to a unique label (we aim to correctly classify all neighbors of this node to be this unique class). Therefore, the more frequent participation of high-degree nodes in the optimization by message-passing, the more likely their representations would be optimized towards pairing with so many unrelated nodes, and hence their performance would decrease. As shown in Figure 1, Recall@20 decreases when node degree increases, which aligns with our argument. Note that because Normalized Discounted Cumulative Gain (NDCG), unlike Recall, is a biased evaluation metric (as justified later in Section II-A), we observe that NDCG@20 increases as the node degree increases.

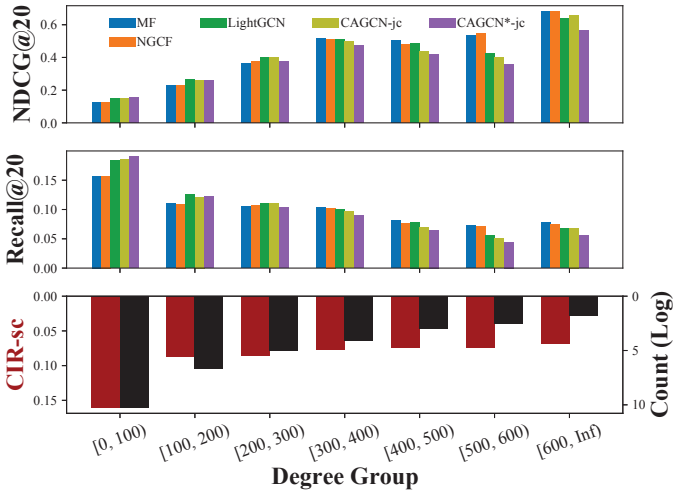

Fig. 1. The performance of each model w.r.t. node degree on Gowalla. We can clearly see that Recall decreases as node degree increases while NDCG increases as node degree increases. CIR denotes the common interacted ratio, which measures how nodes' neighbors are connected with each other through higher-order paths. See [4] for further definition.

II. Analyzing Bias in Link Prediction

Let $G = (\mathcal{V}, \mathcal{E})$ be an undirected graph, where $\mathcal{V} = \{v_1, v_2, ..., v_n\}$ is the set of nodes with $n = |\mathcal{V}|$ and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges with $m = |\mathcal{E}|$. Given the historical edges $\bar{\mathcal{E}}$ that we have observed, most link predictors expect to predict the incoming edges $\hat{\mathcal{E}}$ with $\mathcal{E} = \bar{\mathcal{E}} \cup \hat{\mathcal{E}}$ by learning a mapping $\mathcal{V} \times \mathcal{V} \rightarrow \mathbf{S} \in \mathbb{R}^{n \times n}$, where $\mathbf{S}_{ij} \in \mathbb{R}$ represents how likely a link will form between v_i and v_j. The performance of each node v_i is evaluated by comparing the level of the alignment between its ground-truth 1-hop neighbors $\hat{\mathcal{N}}_i^1$ and its predicted 1-hop neighbors $\widetilde{\mathcal{N}}_i^1$. Specifically, for each node v_i, we sort its preference scores over all nodes $\mathbf{S}_i \in \mathbb{R}^n$ and select the top-K items to form its predicted 1-hop neighbors $\widetilde{\mathcal{N}}_i^1 = \{v_{\phi_i^k}\}_{k=1}^K$ where ϕ_i^k denotes v_i's k^{th} preferred item selected according to the rank of $\mathcal{S}_i$. Assuming $K < |\hat{\mathcal{N}}_i^1|$, then we formulate four commonly-used evaluation metrics Recall(R), Precision(P), F1 and NDCG(N) as:

2022 IEEE International Conference on Data Mining Workshops (ICDMW) | 979-8-3503-4609-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICDMW58026.2022.00103

2375-9259/22/$31.00 ©2022 IEEE
DOI 10.1109/ICDMW58026.2022.00103

$$\mathrm{R@}K_i = \frac{|\hat{\mathcal{N}}_i^1 \cap \widetilde{\mathcal{N}}_i^1|}{|\hat{\mathcal{N}}_i^1|}, \qquad \mathrm{P@}K_i = \frac{|\hat{\mathcal{N}}_i^1 \cap \widetilde{\mathcal{N}}_i^1|}{K} \tag{1}$$

$$\mathrm{F1@}K_i = 2\frac{\mathrm{P@}K \cdot \mathrm{R@}K}{\mathrm{R@}K + \mathrm{P@}K} = \frac{2|\hat{\mathcal{N}}_i^1 \cap \widetilde{\mathcal{N}}_i^1|}{K + |\hat{\mathcal{N}}_i^1|} \tag{2}$$

$$\mathrm{N@}K_i = \frac{\sum_{k=1}^{K} \frac{1[v_{\phi_i^k} \in (\hat{\mathcal{N}}_i^1 \cap \widetilde{\mathcal{N}}_i^1)]}{\log_2(k+1)}}{\sum_{k=1}^{K} \frac{1}{\log_2(k+1)}} \tag{3}$$

A. Theoretical Analysis

Before analyzing bias in link prediction with the evaluation metrics defined above, we first theoretically prove that Recall is an unbiased evaluation metric while Precision, F1, and NDCG are biased ones. Assuming that $|\hat{\mathcal{N}}_i^1 \cap \widetilde{\mathcal{N}}_i^1|$ follows hyper-geometric distribution for any node v_i and $|\hat{\mathcal{N}}_i^1| = d$, the relationship between the expectation of each evaluation and the node activity d is derived as:

1) Recall:

$$E(\mathrm{R@K}|d) = \frac{K}{n}, \quad \frac{\partial E(\mathrm{R@K}|d)}{\partial d} = 0, \tag{4}$$

2) Precision:

$$E(\mathrm{P@K}|d) = \frac{d}{n}, \quad \frac{\partial E(\mathrm{P@K}|d)}{\partial d} = 1, \tag{5}$$

3) F1:

$$E(\mathrm{F1@K}|d) = \frac{2K}{n}\frac{d}{K+d}, \frac{\partial E(\mathrm{F1@K}|d)}{\partial d} = \frac{2K^2}{n}\frac{1}{(K+d)^2} \in (0,1), \tag{6}$$

4) NDCG:

$$E(\mathrm{N@K}|d) = \frac{d}{n}, \quad \frac{\partial E(\mathrm{N@K}|d)}{\partial d} = 1. \tag{7}$$

For brevity, we leave out the detailed derivations for Eq. (4)-(7). Obviously, Precision, F1, and NDCG increase as the node degree d increases and hence will lead to bias in evaluating the degree-related bias in link prediction. Note that although the node degree d is defined to be the size of the ground-truth neighborhood, the conclusion still holds since typically, nodes with high degrees in training data would also have high degrees in testing data assuming no degree distribution shift. For example, on social networks, if the users are highly-active on the platform in the last year, it is safe to assume they will maintain their high-level activity in the following year [8].

B. Empirical Analysis

We further empirically verify the above observation by leveraging an untrained link predictor to calculate the corresponding evaluation metric. More specifically, for each node v_i, we randomly select K nodes from $\mathcal{V}$ and check whether the selected K nodes come from $\hat{\mathcal{N}}_i$. To approximate the expectation with less error, we average the results over 200 runs. Straightforwardly, any untrained model should output exactly the same performance for each individual. However, it is clearly seen in Figure 2 that when evaluating with Precision, F1, and NDCG, the performance still increases as the node degree increases, which is consistent with what we derive in Section II-A and further demonstrates the evaluation bias embedded in Precision, F1, and NDCG.

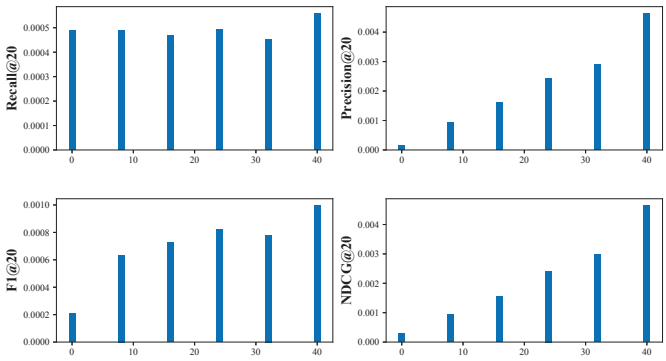

Fig. 2. Performance under each metric w.r.t. node degrees on Gowalla.

III. Conclusion

In this paper, we propose a novel issue, degree-related bias and evaluation bias, in link prediction. We first empirically demonstrate the imbalanced performance of link prediction on nodes with different degrees, which disclose the degree-related bias in link prediction. Then, we theoretically analyze the bias of different evaluation metrics and prove that NDCG, F1 and Precision are all biased towards high-degree nodes while Recall is the only unbiased evaluation metric. When evaluating under the unbiased metric Recall, we finally conclude that low-degree nodes tend to have higher performance in link prediction than high-degree nodes. In future work, we plan to focus on degree-related bias from the perspective of local clustering coefficient, and then more generally on fair graph representation learning [6], [9].

References

[1] Y. Liu, Y. Wang, O. T. Vu, R. Moretti, B. Bodenheimer, J. Meiler, and T. Derr, "Interpretable chirality-aware graph neural network for quantitative structure activity relationship modeling in drug discovery," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[2] X. Wang, Y. Ma, Y. Wang, W. Jin, X. Wang, J. Tang, C. Jia, and J. Yu, "Traffic flow prediction via spatial temporal graph neural network," in *Proceedings of The Web Conference 2020*, 2020, pp. 1082–1092.

[3] B. Rozemberczki, C. T. Hoyt, A. Gogleva, P. Grabowski, K. Karis, A. Lamov, A. Nikolov, S. Nilsson, M. Ughetto, Y. Wang, T. Derr, and B. M. Gyori, "Chemicalx: A deep learning library for drug pair scoring," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

[4] Y. Wang, Y. Zhao, Y. Zhang, and T. Derr, "Collaboration-aware graph convolutional networks for recommendation systems," *arXiv preprint arXiv:2207.06221*, 2022.

[5] X. Tang, H. Yao, Y. Sun, Y. Wang, J. Tang, C. Aggarwal, P. Mitra, and S. Wang, "Investigating and mitigating degree-related biases in graph convoltuional networks," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020.

[6] Y. Dong, J. Ma, C. Chen, and J. Li, "Fairness in graph mining: A survey," *arXiv preprint arXiv:2204.09888*, 2022.

[7] T. Zhao, X. Zhang, and S. Wang, "Graphsmote: Imbalanced node classification on graphs with graph neural networks," in *Proceedings of the 14th ACM international conference on web search and data mining*, 2021.

[8] Y. Yang, Y. Dong, and N. V. Chawla, "Predicting node degree centrality with the node prominence profile," *Scientific reports*, vol. 4, no. 1, 2014.

[9] Y. Wang, "Fair graph representation learning with imbalanced and biased data," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 2022.