2024 IEEE 40th International Conference on Data Engineering (ICDE)

SES: Bridging the Gap Between Explainability and Prediction of Graph Neural Networks

Zhenhua Huang[†], Kunhao Li[†], Shaojie Wang[†], Zhaohong Jia[*†], Wentao Zhu[‡], Sharad Mehrotra[§]
[†]Anhui University, Hefei, China
[‡]Amazon Research, Seattle, USA
[§]University of California Irvine, Irvine, USA
[*]Key Lab of Intelligent Computing and Signal Processing of Ministry of Education, Hefei, China
{zhhuangscut, kunhomlihf, wsj.ahu, wentaozhu91}@gmail.com, zhjia@mail.ustc.edu.cn, sharad@ics.uci.edu

2024 IEEE | DOI: 10.1109/ICDE60146.2024.00229

***Abstract*—Despite the Graph Neural Networks' (GNNs) proficiency in analyzing graph data, achieving high-accuracy and interpretable predictions remains challenging. Existing GNN interpreters typically provide post-hoc explanations disjointed from GNNs' predictions, resulting in misrepresentations. Self-explainable GNNs offer built-in explanations during the training process. However, they cannot exploit the explanatory outcomes to augment prediction performance, and they fail to provide high-quality explanations of node features and require additional processes to generate explainable subgraphs, which is costly. To address the aforementioned limitations, we propose a self-explained and self-supervised graph neural network (SES) to bridge the gap between explainability and prediction. SES comprises two processes: explainable training and enhanced predictive learning. During explainable training, SES employs a global mask generator co-trained with a graph encoder and directly produces crucial structure and feature masks, reducing**

volution networks (GCN) [17], graph attention networks (GAT) [18], GraphSAGE [19], graph isomorphism network (GIN) [20], ARMA [21], UniMP [22], FusedGAT [23], and ASDGN [24], *etc*. However, these advancements have not adequately addressed the need for explainability in the representations learned by GNNs.

To address this, instance-level or model-level approaches have been proposed to offer explanations of GNNs, which are mostly post-hoc models. A post-hoc model is a statistical or predictive model constructed after data processing, enabling retrospective analysis and interpretability of variable relationships [25]. Noteworthy instance-level post-hoc GNN explainers include GNNExplainer [26], PGExplainer [27], PGMExplainer [28], and GraphLIME [29], *etc*. instance-