Paper Summary:

GRAPHTRAIL: Translating GNN Predictions into Human-Interpretable Logical Rules

Afreen Alam

CS7064: Trustworthy Machine Learning

January 9, 2026

Summary of the Paper

This paper introduces GRAPHTRAIL, an end-to-end, post-hoc, global explainer for message-passing graph neural networks (GNNs). It aims to translate black-box graph classification predictions into human-interpretable logical rules over subgraph level concepts.

Existing work on interpretability of GNNs rely on local or instance-level explainers, which are unable to uncover patterns used by GNNs at a global level across multiple graphs to make predictions. Furthermore, the authors claim that the reliance of existing global explainers on instance-level explainers leads to a lack of global understanding, interpretability, and robustness.

To address these challenges and capture the combinatorial reasoning of GNNs across the entire training dataset, the authors cleverly utilize the property that message-passing GNNs decompose a graph into a set of computation trees which maps the information flow within GNNs. This reduces the exponential candidate space to a linear one, substantially reducing the complexity. Next, using Shapely Value the global impact of the computation trees is assessed and the top-k computation trees with the highest Shapely values are selected. Lastly, Symbolic Regression is performed over the top-k computation trees to generate simple Boolean functions that can be understood by humans.

The authors carried out extensive experiments on three GNN architectures across four diverse datasets and different pooling layers using GLGEXPLAINER as the baseline. The results demonstrate the superiority of GRAPHTRAIL in terms of fidelity, interpretability, robustness, and data efficiency compared to the baseline method.

In conclusion, GRAPHTRAIL is able to effectively and efficiently capture the logic within GNN models in a human-understandable form which makes the decision-making process of black-box GNNs more transparent than existing methods.

Strengths and Weaknesses

Strengths

1. The method proposed is an end-to-end pipeline which is post-hoc. This means that retraining of the GNN model is not required.

2. The utilization of the insight of message-passing GNNs being able to decompose a graph into a set of computation trees is novel which not only contributes to the significant reduction of the search space but also aligns much better with how information flows within GNNs.

3. The authors propose an efficient way of computing Shapely values on computation trees of GNNs. This is novel since there has been no prior studies related to application of Shapely values on GNNs.

4. Experiments show that the proposed method is able to maintain high fidelity even in low-resource settings which makes this method suitable for applications where there is a lack of large training data.

Weaknesses

1. The authors address that computing Shapely values can become computationally burdensome for datasets with larger numbers of graphs and propose GRAPHTRAIL-S, which samples from the training dataset and uses a subset of it. Although they show the fidelity of GRAPHTRAIL-S, I would also have liked to see some comparisons of how computationally efficient it is compared to GRAPHTRAIL to be able to judge the trade-off between fidelity and complexity between the two proposed methods.

2. The authors mentioned in the Limitations section of the paper that the proposed method lacks the ability to capture the multiplicity of a concept's occurrence within a graph since their method relies on Boolean logic. It would have been great if they could elaborate a little on that and maybe suggest some future directions that they might employ to mitigate this limitation.

Questions

1. Can you explain what the authors mean by multiplicity of a concept's occurrence and how would knowing that increase the interpretability of GNNs?

2. The authors limit the logic to simple boolean formulas. Do you think more complex boolean logic could improve the fidelity? Or would that add too much complexity to the calculation?

3. How can we separate a failure of the GNN from a failure of the explainer?

4. The authors addressed the dilemma between faithfulness and interpretability. Could you elaborate on that and do you have any thoughts on how this challenge could be addressed?

5. Do you think any other method could have been used instead of Symbolic Regression?