Link Explanation for Heterogeneous Graphs

Yale

Abhijit Gupta, Department of Computer Science, Rex Ying, Department of Computer Science

MOTIVATION

Graph data emerges in many contexts (social networks, molecules, knowledge graphs) but creates unique challenges for traditional machine learning (ML) methods.

Graph Neural Networks (GNNs) are a type of neural network that operates directly on the graph structure to perform tasks.

Explainability builds trust, promotes fairness, and can improve human-in-the-loop performance.

Just as graph structured data poses challenges to traditional ML methods, new explainability methods are required to reason about graph structured data.

GNNs have been applied in many settings, but explanation methods have lagged behind. For example, **link explanation** and **heterogeneous graph explanation** are underexplored.

BACKGROUND

GNNs are composed of multiple message passing layers. GCN, GraphSAGE, GAT, and GIN are common building blocks.

Explanations can be made during initial prediction or post hoc. We focus on post hoc model-agnostic perturbation methods.

- **GNNExplainer** optimizes soft masks to maximize mutual information with the GNN prediction.
- **SubgraphX** applies Monte Carlo Tree Search and Shapley values to measure subgraph importance.

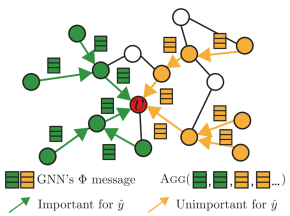

Figure 1: A computation graph for predicting the class of node v with useful and unimportant neighbors. The goal is to identify a small set of important pathways crucial to the prediction.

Explanations are evaluated by multiple metrics including sparsity (size of explanation), fidelity (necessary/sufficient), stability, accuracy, and inference time.

$$fid+^{prob} = \frac{1}{N}\sum_{i=1}^{N}(f(G_C)_{y_i} - f(G_{C\setminus S})_{y_i})$$

$$fid-^{prob} = \frac{1}{N}\sum_{i=1}^{N}(f(G_C)_{y_i} - f(G_S)_{y_i})$$

$$charact = \frac{w_+ + w_-}{\frac{w_+}{fid_+} + \frac{w_-}{1-fid_-}}$$

EXPERIMENTS

We selected the **Facebook** (homogeneous) and **IMDB** (heterogeneous; movie-actor, movie-director edges) datasets.

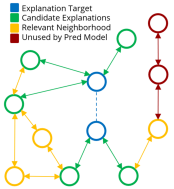

Figure 2: Explanations are restricted to **immediate neighbors** for increased interpretability and real-world use cases.

When explaining the dotted blue edge, we only allow subsets of the green nodes as possible explanations.

We measured continuous characterization score (with emphasis on necessary explanations) as a function of sparsity.

Modified GNNExplainer: New loss encourages ordering of candidate nodes, handles varying neighborhood sizes better.

Encourages smaller explanations (in # of nodes)

$$L_{\text{old}} = -H(Y|G = G_S) + \alpha \sum_{e_i \in E_S} e_i + \beta \cdot \text{CrossEntropy}(E_S)$$

Encourages discrete mask

Optimizes explanation towards target

Encourages continuous mask

$$L_{\text{new}} = -H(Y|G = G_S) + \alpha\left(\left(\frac{1}{|E_S|}\sum_{e_i \in E_S} e_i\right) - 0.5\right)^2 - \beta \cdot \text{CrossEntropy}(E_S)$$

Encourages medium explanation (in % of nodes)

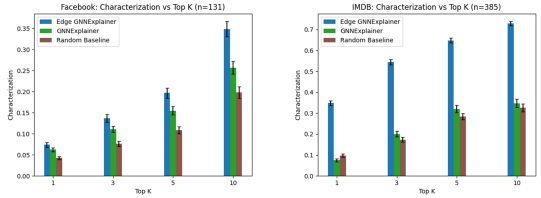

Figure 3: Modified GNNExplainer outperforms GNNExplainer on both datasets, across all explanation sizes.

Modified SubgraphX: Instead of masking removed nodes to 0, mask removed edges by deleting edge from subgraph.

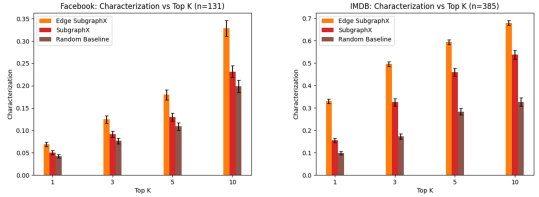

Figure 4: Modified SubgraphX outperforms SubgraphX as well.

RESULTS

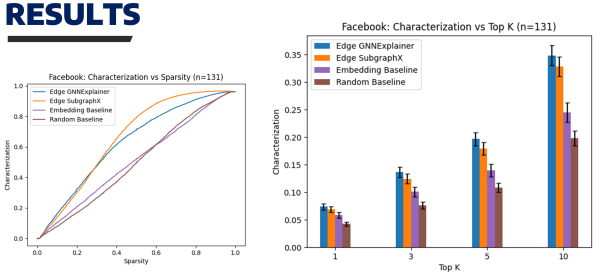

Figure 5: Facebook dataset characterization scores for four explanation methods. Left: Continuous plot with sparsity percentage. Right: Discrete plot with fixed explanation size. Modified GNNExplainer and SubgraphX outperform baselines.

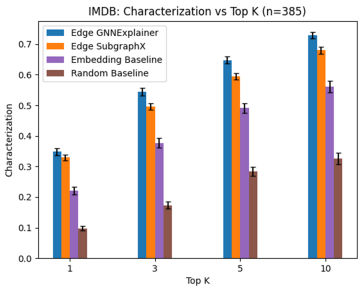

Figure 6: IMDB dataset characterization scores for four explanation methods. Modified GNNExplainer and SubgraphX outperform Embedding and Random baseline.

CONCLUSIONS

> First application of GNN explanation methods to heterogeneous graphs, link prediction task.

> Modified GNNExplainer and SubgraphX algorithms outperform the Embedding baseline.

> Modified GNNExplainer has the best combination of inference time and fidelity among methods tested.

NEXT STEPS

Improve scalability to enable additional results on the Facebook dataset, LastFM dataset. Develop new explanation methods leveraging heterogeneous graph meta-paths.

Contribute heterogenous graph explanation and link explanation to the open-source PyTorch Geometric project.