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 [1].

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 [2].
- **SubgraphX** applies Monte Carlo Tree Search and Shapley values to measure subgraph importance [3].

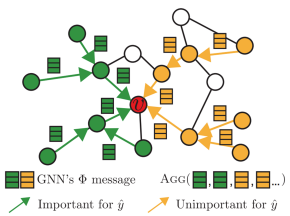

Figure 1: A computation graph for predicting the class of node v with useful and unimportant neighbors. The goal is to identify a small set of important pathways crucial to the prediction [2].

Explanations are evaluated by multiple metrics including sparsity (size of explanation), fidelity (necessary/sufficient), stability, accuracy, and inference time [4].

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

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

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

EXPERIMENT

We modify existing explanation methods for link prediction and heterogeneous graphs and quantify explanation quality.

Datasets

- **Facebook**: Homogeneous graph of 4000+ Facebook users with 170,000+ friend connections [5].
- **IMDB**: Heterogeneous graph with 11000+ movies, directors, and actors, 15000+ (actor-movie), (movie-director) edges [6]

Link Prediction Models

- **Facebook**: We trained a 2-layer GCN Encoder and Linear Decoder with 180K parameters, 89% ROC AUC on test set
- **IMDB**: We trained a 2-layer SAGE Encoder and dot-product Decoder with 320K parameters, 77% ROC AUC on test set

Restricted explanations to immediate neighbors for increased interpretability, real-world applications.

Explanation Models

- **Random Baseline**: Randomly predict importance
- **Embedding Baseline**: Similarity to non-adjacent node
- **Modified GNNExplainer**: Extended GNNExplainer to link prediction and heterogeneous graphs. Modified subgraph masking and loss function to improve performance.
- **Modified SubgraphX**: Removed MCTS step for speedup. Masks edges instead of nodes, use greedy algorithm.

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

RESULTS

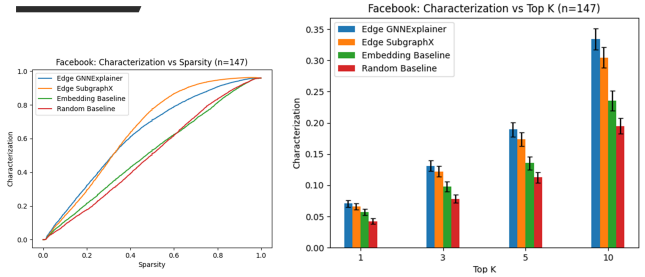

Figure 2: Facebook dataset characterization scores for four explanation methods. Left: Continuous plot with sparsity percentage. Right: Discrete plot with fixed explanation size. Modified GNNExplainer and SubgraphX outperform baselines.

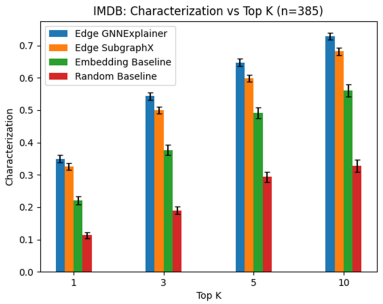

Figure 3: IMDB dataset characterization scores for four explanation methods. Modified GNNExplainer and SubgraphX outperform Embedding baseline.

CONCLUSIONS

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

> Modified GNNExplainer and SubgraphX algorithms outperform the Embedding baseline

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

NEXT STEPS

Improve scalability to enable additional results on the Facebook dataset, LastFM dataset [6]. Develop new explanation methods leveraging heterogeneous graph meta-paths.

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

REFERENCES

[1] Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. Explainability in graph neural networks: A taxonomic survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1–19, 2022.
[2] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. In Advances in neural information processing systems, volume 32. Curran Associates, Inc., 2019.
[3] Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. On explainability of graph neural networks via subgraph explorations. In International Conference on Machine Learning, pages 12241–12252. PMLR, 2021.
[4] Kenza Amara, Rex Ying, Zitao Zhang, Zhihao Han, Yinan Shan, Ulrik Brandes, Sebastian Schemm, and Ce Zhang. Graphframex: Towards systematic evaluation of explainability methods for graph neural networks. arXiv preprint arXiv:2206.09677, 2022.
[5] Jure Leskovec and Julian Mcauley. Learning to discover social circles in ego networks. Advances in neural information processing systems, 25, 2012.
[6] Xinyu Fu, Jiani Zhang, Ziqiao Meng, and Irwin King. Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding. In Proceedings of The Web Conference 2020, pages 2331–2341, 2020.