Beyond Spectral Graph Theory: An Explainability-Driven Approach to Analyzing the Stability of GNNs to Topology Perturbations

Rauno Arike
Author
R.Arike@student.tudelft.nl

Elvin Isufi
Responsible Professor

CSE3000 Research Project

Maosheng Yang
Supervisor

Mohammad Sabbaqi
Supervisor

1 - Introduction

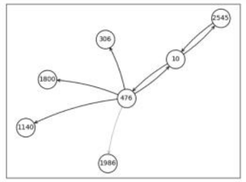
Graphs are all around us, from social networks to chemical structures. **Graph Neural Networks (GNNs)** are machine learning models that leverage the relational structure of graph data for efficient learning on graphs.

The above graphs represent the friend relationships of 4 people. Suppose you have trained a GNN to predict the favourite hobby of each person based on their relationships, but suddenly, Joe and Anna are no longer friends and Anna and Kelly become friends — the graph on the left becomes the graph on the right.

Can you still trust the predictions of the GNN model that you trained on the original graph after this change to the topology of the graph? This is an example of a question researched in the field of **GNN stability to topology perturbations**.

3 - Methodology

1. **Train the model**
2. **Get an explanation subgraph for each node**



3. **Find the number of explanation subgraphs each node appears on**

node #10: 31
node #476: 24
node #1800: 18
node #1140: 16
node #2545: 3
node #306: 2
...

4. **Sort the nodes into 10 frequency groups**

24+ nodes: node #10, node #476, ...
16-23 nodes: node #1800, node #1140, ...
...
1-3 nodes: node #2545, node #306, ...

5. **Randomly sample half of the nodes from each bin**
6. **Remove the sampled nodes from the original graph and measure the stability. Repeat steps 5 and 6 for *m* times.**

Stability measure (RRMSE):

$$\Delta y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i)^2}}$$

Figure 1. The algorithm used in our experiments.

2 - Research Question and Background

For some GNNs, the stability properties can be precisely characterised using the mathematical tools of spectral graph theory. For GNNs not amenable to these tools, a different approach is needed. Motivated by that, this project asks: **can we characterise and explain the stability properties of GNNs using tools from the field of Explainable AI?**

We explore this question using two types of GNNs:
- Graph Attention Networks (GATs) [1]
- Graph Convolutional Networks (GCNs) [2]

... and for three types of topology perturbation:
- Node removal perturbations
- Edge removal perturbations
- Edge weight perturbations

The explainability tools (i.e., **explainers**) we use are ones that output a small compact subgraph (i.e., an **explanation subgraph**) to explain the model's prediction for each individual node. A single explanation subgraph displays the nodes and edges that most strongly influence the model's prediction for a specific node. We use two different explainers:
- GNNExplainer [3]
- Integrated Gradients [4]

To generate perturbations informed by the outputs of the explainers, we categorise nodes and edges based on the number of explanation subgraphs they appear on.

4 - Results

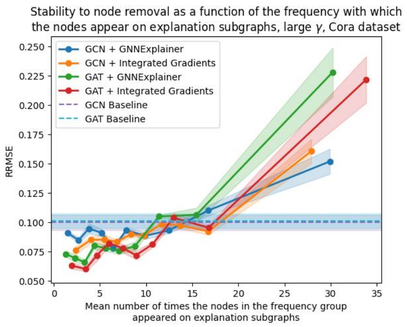


Figure 2. Results for **node removal perturbations.**

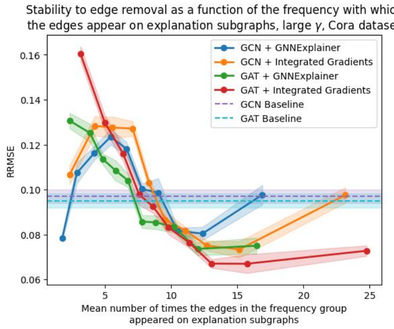


Figure 3. Results for **edge removal perturbations.**

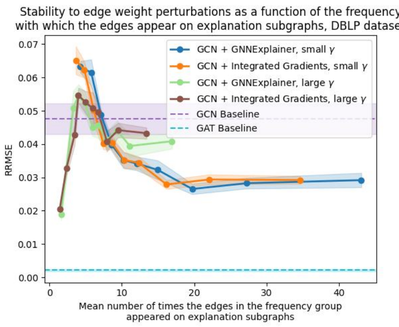


Figure 4. Results for **edge weight perturbations.**

5 - Discussion

The results for node removal have an intuitive explanation: if a node appears on various different explanation subgraphs, it must strongly impact the model's predictions for many other nodes. If we suddenly remove many such nodes from the graph, the model predictions for many other nodes are affected and the impact on stability is high. In contrast, if we remove the nodes that appear only on a few explanation subgraphs, the predictions for fewer nodes are affected and impact on stability is lower.

We expected the edge removal results to follow the same intuitive story, but this isn't the case, as the graphs for edge removal display an opposite trend to node removal.

We found that the reason behind the unintuitive edge removal results is that **perturbations involving edges from smaller frequency groups increases the number of connected components in the graph**, which appears to have a strong impact on stability.

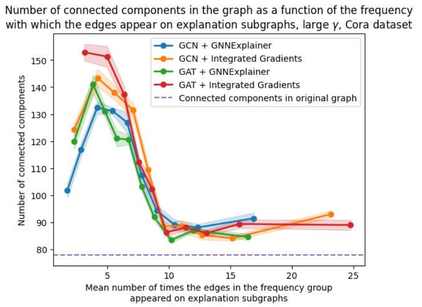


Figure 5. The number of connected components per frequency group.

6 - Conclusion

- Explainers can be a valuable addition to the stability analysis toolkit: we can identify nodes and edges which have a stronger impact on stability than others using the algorithm we introduced, and this can lead to interesting insights, such as the connection between the number of connected components in the graph and stability to edge removal perturbations that we found.
- Nevertheless, our work has multiple limitations that should be addressed in the future: e.g., our algorithm doesn't currently scale to large real-world graphs, and it could be tested on a wider variety of models and perturbation types.

References
[1] P. Velickovic, G. Cucurull, et al., *Graph attention networks*, 2018.
[2] T. N. Kipf and M. Welling, *Semi-supervised classification with graph convolutional networks*, 2017
[3] R. Ying, D. Bourgeois, et al., *GNNExplainer: Generating explanations for graph neural networks*, 2019.
[4] M. Sundararajan, A. Taly, and Q. Yan, *Axiomatic attribution for deep networks*, 2017.