Evaluating the Explainability of Graph Neural Networks for Disease Subnetwork Detection

Sucharitha Rajesh

Supervisor(s): Dr. Megha Khosla, Dr. Jana Weber

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2024

Name of the student: Sucharitha Rajesh
Final project course: CSE3000 Research Project
Thesis committee: Dr.Megha Khosla, Dr.Jana Weber, Dr.Thomas Abeel

Abstract

Graph neural networks (GNNs), while effective at various tasks on complex graph-structured data, lack interpretability. Post-hoc explainability techniques developed for these GNNs in order to overcome their inherent uninterpretability have been applied to the additional task of detecting important subnetworks in graphs. For example, the GNN-SubNet program uses explanations of protein-protein interaction networks to detect the most important disease subnetworks for specific types of cancer. However, when using a post-hoc explanation for such additional tasks, evaluating the quality of the explanation becomes critical.

This study implements four explainability evaluation metrics to provide a fast and accurate way of evaluating explainability, using the GNN-SubNet program as a case study of explainable GNNs for subnetwork detection. Fidelity and sparsity metrics are implemented as defined in existing literature, while validity+ and validity- are newly defined. The results show that GNN-SubNet finds robust and faithful but highly dense explanations.

1 Introduction

Graphs are a powerful way of capturing rich, non-linear information as a collection of nodes and edges. Deep-learning models which operate on graphs, called Graph Neural Networks (GNNs), have been used to analyse such complex data in many domains. However, the complex non-linear mechanisms which make GNNs powerful also make them inherently difficult to interpret. Explainability is one of the four key principles of trustworthy AI, and the lack of interpretability limits the application of GNNs in high-stake real-world domains such as healthcare [1].

Multiple explainability techniques have been proposed for GNNs [2], such as counterfactual explanations, self-interpretable models, and factual post-hoc explanations.

Post-hoc explanations highlight the "important" nodes and/or edges of the input graph, which were used by the GNN to arrive at its decision. Apart from aiding human understanding, post-hoc explanations have also been used for the task of detecting important subnetworks in a graph, in applications such as drug discovery using molecular substructures [3] and detecting regions of interest in the brain [4].

Using post-hoc explanations for such downstream tasks makes it critical to evaluate the quality of the explanation. The BAGEL benchmark [5] proposes four general metrics that can be used to evaluate an explanation:

- Faithfulness measures how well the explanation approximates the model's behaviour.
- Sparsity measures the size of the explanation - a smaller explanation can be more easily understood by a human.
- Correctness measures the explanation's ability to detect correlations that are injected into the graph.
- Plausibility measures to what extent the model uses a decision-making process similar to human rationale.

In addition to providing empirical measures, these metrics are also a fast and accurate tool to support domain experts in validating the results of subnetwork detection. A quick empirical evaluation strategy can also aid the development and testing of new GNNs and explainer techniques.

This study focuses on the GNN-SubNet program [6] as a case study for subnetwork detection using explainable GNNs. GNN-SubNet operates on protein-protein interaction (PPI) networks, where nodes represents proteins and edges represent interactions between proteins. A GNN is trained to classify PPI graphs that have been enriched with data from cancer patients. The post-hoc explanation of the GNN is then used to detect disease subnetworks. These subnetworks suggest novel proteins that may be key to the activation and progression of a certain type of cancer.

This study aims to answer the question, **"How do different explainability evaluation metrics evaluate GNN-SubNet?"**. The following research questions are used to guide the overall goal:

- **RQ1:** How well do the explanations of GNN-SubNet on synthetic data perform when assessed with the most suitable metrics from the BAGEL benchmark?
- **RQ2:** How well do the explanations of GNN-SubNet on KIRC data perform when assessed with the most suitable metrics from the BAGEL benchmark, taking into account the nature of the data, the GNN and the explainer used?

The GNN-SubNet program is further detailed in Section 2. Section 3 presents various explainability metrics and their application to the case of GNN-SubNet, while Section 4 presents and discusses the results of the evaluation. Section 5 reflects on responsible research. Finally, Section 6 presents the conclusions drawn and discusses the limitations and future scope of work.

2 Background - GNN-SubNet

The GNN-SubNet [6] project uses an explainable GNN to detect novel disease subnetworks from a dataset of enriched PPI networks. PPI networks model the interaction between proteins as the edges of a graph. GNNs trained on PPI networks have been used for tasks such as predicting protein function, identifying essential proteins and predicting protein interfaces [7].

Figure 1 gives an overview of the methodology of GNN-SubNet. The nodes (proteins) of the PPI network are enriched with two patient-specific features: DNA methylation and gene expression. This results in a multi-omic dataset of graphs with one graph per patient, such that all graph share the same PPI topology but have patient-specific node features.

The dataset consists of patients with different types of cancers. Focusing on one type of cancer, the graphs are labelled as either "cancer-specific", if they belong to the chosen cancer type, and as "cancer-random" if they are of a different cancer type. A GNN with the Graph Isomorphism Network (GIN) architecture [8] is then trained to classify the enriched patient-specific graphs as either cancer-specific or cancer-random.

A modified version of the GNNExplainer technique [9] is then used to optimize a global explanation in the form of a soft node mask. A node mask contains a single importance value for each node - the higher the value, the more important the node. The traditional method of GNNExplainer optimizes a node mask on a single input graph, to create a "local" explanation. For the purpose of GNN-SubNet, the authors optimize the node mask over a sample of graphs from the dataset, resulting in a "global" explanation. Such an explanation aims to obtain a node mask that explains the GNN as a whole, rather than the GNN's action on a specific input graph.

GNNExplainer results in soft node masks. Soft masks assign a value between 0 and 1 to each node, with higher values indicating higher importance. On the other hand, hard mask explanations assign a binary value, 0 or 1, to each node. The type of node mask used influences the implementation of evaluation metrics, as further discussed in Section 3.5.

Finally, having found the global explanation as a soft node mask, the edges of the PPI network are weighted using the importance values from the mask, and a community detection algorithm is applied. The communities with the highest average importance values represent the most important disease subnetworks.

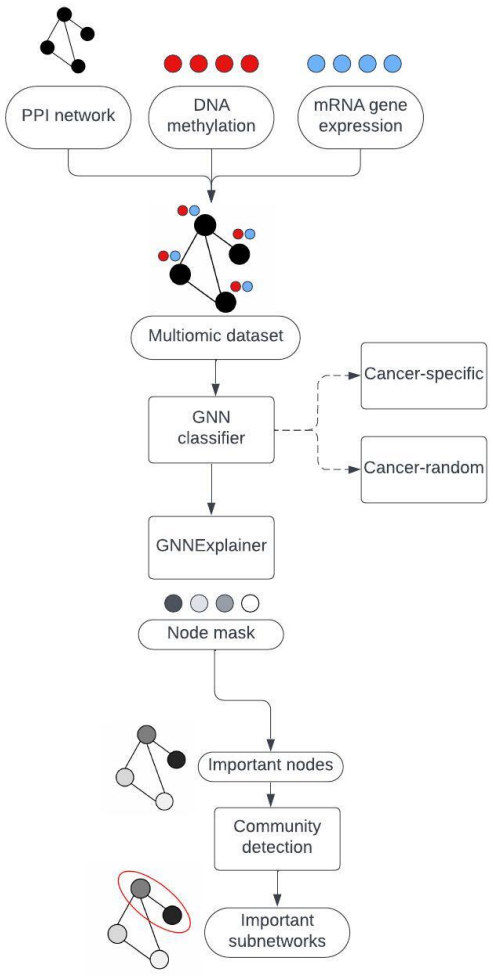

Figure 1: Existing workflow of GNN-SubNet: constructing a multiomic dataset, training a GNN classifier, and generating a global explanation in the form of a node mask. The node mask is further used to construct edge importances and detect potential disease subnetworks.

3 Methods

Metrics from the BAGEL [5] benchmark and from other existing literature [10], [11] are presented and analysed here to determine their applicability to GNN-SubNet.

3.1 Faithfulness

The fidelity of an explanation quantifies how faithful an explanation is to the behaviour of the GNN model. BAGEL [5] proposes two measures to quantify faithfulness: rate-distortion based fidelity (RDT-fidelity), and comprehensiveness and sufficiency.

RDT-Fidelity

The approach taken by RDT-fidelity perturbs the feature values in the nodes of the input graph in proportion to the importance of the node. Nodes with a higher importance are perturbed to a smaller extent. This perturbation approach is suitable for GNN-SubNet, as the feature values are meaningful representations of the DNA methylation and mRNA expression values. More formally, the perturbation Y_S of an input graph X by explanation $\mathcal{S}$ is defined by [5] as follows:

$$Y_{\mathcal{S}} = X \odot M(\mathcal{S}) + Z \odot (\mathbb{1} - M(\mathcal{S})), Z \sim \mathcal{N} \quad (1)$$

where $M(\mathcal{S})$ is the node mask corresponding to explanation $\mathcal{S}$, containing one value for each node, and $\mathcal{N}$ is a noise distribution taken to be the same as the global distribution of the dataset.

To measure RDT-fidelity, 10 sample perturbations of each test graph are created using the soft node mask of the global explanation found by GNN-SubNet. 10 samples were chosen in order to have a reasonable number of random perturbations while still being computationally feasible within the scope of the study. The proportion of these samples that have the same prediction as the original input is reported as the RDT-fidelity score. A higher score indicates a better, more robust explanation, where the perturbations do not affect the classifier's prediction.

Comprehensiveness and Sufficiency

Comprehensiveness answers the question, "whether all nodes/edges needed to make a prediction were selected", while sufficiency answers the question "whether the selected nodes/edges are sufficient to come up with the original prediction" [5].

However, these involve testing the model's output on graphs where the important nodes are removed. If applied to GNN-SubNet, this would amount to removing proteins and interactions between proteins, which render the PPI topology invalid. These metrics are thus not implemented in the evaluation of GNN-SubNet.

3.2 Validity+ and Validity-

Fidelity+ and Fidelity- have been proposed by [10] and are similar to comprehensiveness and sufficiency. They describe removing either the important nodes or the unimportant nodes from the graph and observing the change in the GNN's prediction.

Rather than removing nodes from the PPI network, we adopt the approach of setting the feature values of some nodes to a baseline. This approach is described by [11] in their definition of the explainability metric "validity". Within GNN-SubNet, this is done by setting the DNA methylation and mRNA expression (features) of a protein (node) to the average value of that protein's DNA methylation and mRNA expression over the entire dataset.

Two metrics, Validity+ and Validity- are thus defined. These use hard node masks to define whether the features of a specific node are set to average or are left unchanged. Such hard masks are obtained using the transformation described in Section 3.5.

Validity- sets the features of unimportant nodes in a graph to average values. If the explanation has selected the important nodes used by the model, this will not lead to a change in the GNN's prediction on the altered graph. The Validity- score is reported as the proportion of graphs for which the prediction does not change. The higher this score, the better.

With N the number of graphs in the test dataset, and f a function representing the GNN (such that $f(\mathcal{G}_i)$ gives the decision of the GNN on the i-th input graph of the dataset), Validity- is defined as:

$$Validity- = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{f(\mathcal{G}_i)=f(\mathcal{G}_i^{m_i})}, \tag{2}$$

where the altered graph $\mathcal{G}_i'$ is given by:

$$\mathcal{G}_i' = \mathcal{G}_i \odot M(\mathcal{S}) + \mathcal{A} \odot (\mathbb{1} - M(\mathcal{S})) \tag{3}$$

Here $M(\mathcal{S})$ is the node feature mask corresponding to explanation $\mathcal{S}$. This mask is over both node and feature values, thus has dimension $n * d$ for a graph with n nodes and d features per node. $\mathcal{A}$ contains the average node feature values for each node over the whole dataset and is also of the dimension $n * d$.

Validity+ sets the features of important nodes to average values. A good explanation should select the nodes with the most discriminative power, and averaging the features of important nodes should lead to a loss in the accuracy of the GNN model. The Validity+ score is reported as the proportion of graphs for which the prediction is different from the original prediction. The higher this score, the better.

Using the same notation described above, Validity+ is formally defined as:

$$Validity+ = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{f(\mathcal{G}_i)!=f(\mathcal{G}_i')}, \tag{4}$$

where the altered graph is given by:

$$\mathcal{G}_i' = \mathcal{A} \odot M(\mathcal{S}) + f(\mathcal{G}_i) \odot (\mathbb{1} - M(\mathcal{S})) \tag{5}$$

3.3 Sparsity

A meaningful explanation should be sparse in order to be useful - i.e. it should clearly select a small number of nodes as being important. Formally, this is defined as entropy in the BAGEL benchmark [5] and by [11] as follows:

Let M be a node mask, containing one value per node. $\mathrm{mask}(n)$ represents the value of node n in the given mask. The normalised node mask p is then computed:

$$p = \frac{\mathrm{mask}(n)}{\sum_{n' \in M} \mathrm{mask}(n')} \tag{6}$$

and from this, the entropy is computed:

$$H(p) = -\sum_{f \in M} p \log p \tag{7}$$

Explanations with lower entropy are considered more sparse. A uniform node mask of length N (having all values identical) lacks any significant meaning as an explanation. Such a node mask has the highest possible entropy, $max\_entropy = -log(1/|N|)$. To convert the entropy into a clear sparsity score between 0 and 1 such that a higher sparsity score indicates a better explanation, we define the sparsity score as:

$$1 - H(p)/max\_entropy \tag{8}$$

.

GNN-SubNet finds a single global explanation over the entire dataset, which results in a single value of the sparsity metric. For a better representation, the average sparsity score over ten runs of the explainer is reported.

3.4 Other metrics

Correctness and plausibility are two metrics defined in BAGEL which are not applied to the GNN-SubNet task.

Correctness tests whether the explanation can detect externally injected biases in the model. This is suitable for explanations of a node classification task and cannot be applied to the graph classification task that is found in GNN-SubNet.

Plausibility measures how similar the model's decision-making process is to human rationale. This requires data from human experts, which is neither readily available nor feasible to collect within the scope of this study.

3.5 Transforming Soft Masks to Hard Masks

The validity metrics require explanations to be hard node masks that contain binary values. However, GNNExplainer gives the output in the form of a soft mask, containing values in the range between 0 and 1. The approach proposed by [11] is implemented as follows: using a threshold k, the top $k\%$ of entries with the highest soft mask value are set to 1 in the hard mask, and the rest are set to zero. [11] report metrics by using the top-30% and top-50% masks, which are denoted as S-0.5 and S-0.7.

Table 1: Average metric scores over 10 iterations (training, explanation and evaluation) on the synthetic dataset. Validity metrics are evaluated using two different hard masks obtained by different thresholds, RDT-fidelity and sparsity use soft masks.

Metric	Threshold	Average	Std.deviation
RDT-fidelity	NA	1.0	0.0
Sparsity	NA	0.058	0.024
Validity+	S-0.7	0.508	0.021
	S-0.5	0.526	0.039
Validity-	S-0.7	1.0	0.0
	S-0.5	1.0	0.0

Table 2: Average metric scores over 10 iterations (training, explanation and evaluation) on the KIRC dataset. Validity metrics are evaluated using two different hard masks obtained by different thresholds, RDT-fidelity and sparsity use soft masks.

Metric	Threshold	Average	Std.deviation
RDT-fidelity	NA	0.826	0.11
Sparsity	NA	0.040	0.02
Validity+	S-0.7	0.232	0.19
	S-0.5	0.292	0.20
Validity-	S-0.7	0.840	0.15
	S-0.5	0.843	0.01

4 Results and Discussion

RQ1: How well do the explanations of GNN-SubNet on synthetic data perform when assessed with the most suitable metrics?

The four explainability metrics chosen following the analysis in Section 3 are evaluated on a synthetic dataset as a sanity check. This dataset of Barabasi graphs was created and used by [6]. It consists of 1000 graphs with 30 nodes each and a single feature for each node.

Two connected nodes in each graph are assigned discriminative values that determine to which class the graph belongs. For half the graphs, these two nodes are assigned values from $N(\mu = 1, \sigma)$, and for the other half $N(\mu = -1, \sigma)$.The rest of the nodes are assigned values from $N(\mu = 0, \sigma)$. Figure 2 visualises this spread of values.

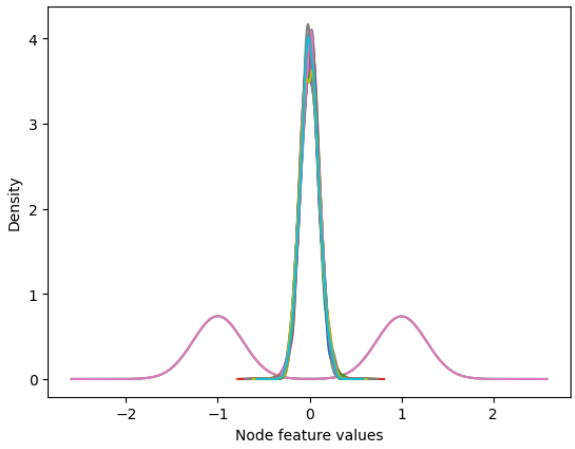

Figure 2: Distribution of the node feature values of the synthetic dataset: each colour represents one node. All nodes are assigned values from $N(\mu = 0, \sigma)$ except for the two "important" nodes whose values are taken from either $N(\mu = 1, \sigma)$ for one class or from $N(\mu = -1, \sigma)$ for the other.

Using the known truth that these two nodes are important, [6] shows that the correct explanation is uncovered in close to 100% of the cases. Therefore, the RDT-fidelity scores, as well as the Validity+ and Validity- scores, are expected to be very high for the explanations of this dataset.

Table 1 shows the results of the evaluation, taking the average value of the metric over 10 iterations. Each iteration involves training the GNN, finding a single global explanation using 10 runs of the explainer and calculating the metrics on this explanation. The RDT-fidelity and Validity- scores are 1.0, indicating a highly robust explanation where the prediction does not change if the unimportant nodes are perturbed or averaged.

The Validity+ score of 0.5 indicates that 50% of graphs change prediction after averaging the important nodes. Specifically, since the important nodes are sampled from $N(\mu = 1, \sigma)$ for one class and $N(\mu = -1, \sigma)$ for the second class, averaging these values across the dataset gives values centered around 0. With no way for the classifier to distinguish between the two classes, a 50-50 random classification is seen as the result.

This demonstrates that it is essential to interpret the results of Validity+ in comparison with a baseline score of 0.50, rather than with the maximum possible score of 1.0. An ideal explanation that highlights all the nodes that are important to the GNN's decision-making would result in all those nodes being averaged out over both classes in the dataset, nullifying their discriminative power and reducing the GNN to a blind, random classifier.

RQ2: How well do the explanations of GNN-SubNet on KIRC data perform when assessed with the most suitable metrics?

The kidney renal clear cell carcinoma (KIRC) cancer dataset used by the GNN-SubNet program [6] is reused here. It consists of 506 graphs sharing the same PPI topology and having two patient-specific node features. The graph topology is very complex, containing 2049 nodes and 13588 edges.

Table 2 shows the average results of evaluation metrics on the KIRC dataset over 10 iterations. Each iteration involves training the GNN for 20 epochs, finding a single global explanation using 10 runs of the explainer and calculating the metrics on this explanation. A GIN architecture with a modified GNNExplainer was used for training. Following a train-validation-test split, 80 test graphs were used in evaluating the explanation.

RDT-fidelity and sparsity GNN-SubNet achieves an RDT-fidelity score of 0.826, indicating that the explanations found are quite robust to perturbations. The sparsity score of 0.04, which was observed consistently across all iterations with little

variation, is quite low and indicates a very dense explanation. This sparsity is equivalent to that of a hard mask which selects 1500 nodes out of the 2049 nodes present in the PPI network as important. Figure 3 shows that node masks have frequent occurences of high importance values, confirming that the explanation is very dense. Such an explanation that selects a large number of nodes as important is considered poor, since they are less interpretable for humans.

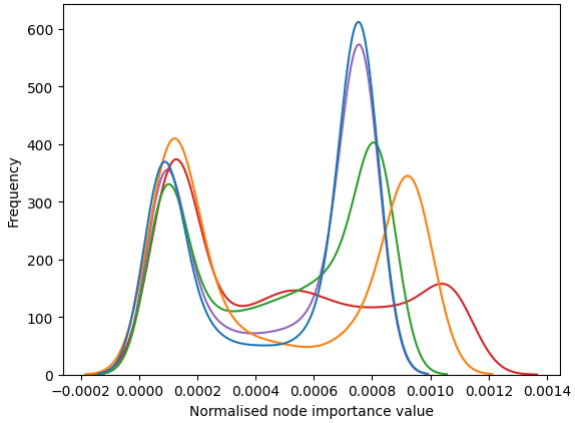

Figure 3: Distribution of normalised node importance values, based on the node masks obtained from 5 iterations of training and explanation. Each coloured line represents a different iteration.

Validity- The high validity- score of 0.84 shows that the important nodes highlighted by the explanation have high discriminative power. This is in accordance with the findings of the authors of GNN-SubNet [6], who show that by re-training the GNN based only on the selected proteins from the most important disease subnetwork, a median classification accuracy of 79% is obtained, implying that the nodes highlighted by the explanation have high discriminative power.

Validity+ Based on the interpretation that the ideal validity+ score is 0.5, as seen in the synthetic dataset, the observed validity+ score is somewhat poor. Additionally, it has a very high standard deviation. The lowest observed validity+ score was 0.037, and the highest 0.487. This suggests that the explanations found in each iteration are very different from each other.

The difference between the high validity- score and the low validity+ score shows that while the most important nodes have high discriminative power, the model is still able to rely on less important nodes to reach the correct prediction. The difference is observed even when using a hard-mask of the top 50% of important nodes, showing that the model relies on a large proportion of nodes present in the graph to make its classification. This suggests that the explanations must necessarily be very dense in order to be faithful and valid, as the model uses a large number of nodes for its prediction.

Investigating the tradeoff between RDT-fidelity and sparsity

In a dense explanation where most nodes have a relatively high importance, the perturbations done by RDT-fidelity are smaller and have less impact. Thus, a correlation is expected between RDT-fidelity and sparsity. This is investigated by evaluating the RDT-fidelity score while controlling the sparsity, as suggested by [10].

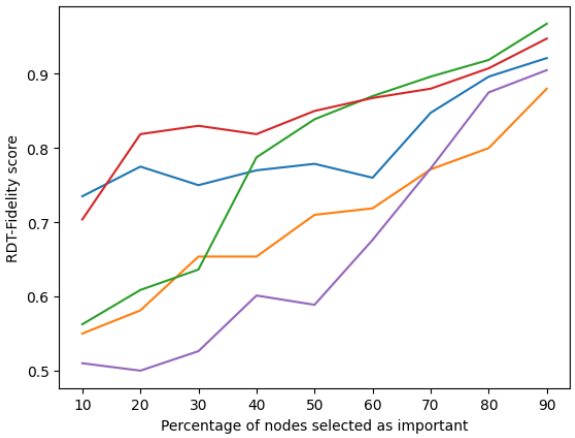

Figure 4: RDT-fidelity score calculated using hard masks where different percentages of nodes are selected as important, over 5 iterations of training and explanation. Each coloured line represents a different iteration.

Figure 4 shows how the RDT-fidelity score increases when more nodes are selected as important, and the explanation becomes more dense. Over 5 different iterations of training, finding a global explanation and evaluating the explanation, the RDT-fidelity score at low thresholds (selecting the top 10% or top 20% of nodes) is highly variable and ranges from 0.5 to 0.77. The score found has no significant correlation with model accuracy, judging by the Pearson correlation coefficient, and appears to be based solely on the explanation itself.

Investigating the size and variability of subnetworks

The high variance in the validity+ score, as well as the highly varying fidelity score when taking the top 10% and 20% of nodes (Figure 4) suggests that the explanation of GNN-SubNet is variable and highlights different nodes each iteration.

Hence the disease subnetworks found would also be highly variable with each iteration. In the ideal case, a subnetwork detection task would reveal stable subnetworks when trained on a certain dataset. Additionally, it is the important subnetworks that are directly assessed by human experts, regardless of how sparse or dense the node mask is. Therefore, small subnetworks are also desirable.

The size and variability of the disease subnetworks found by GNN-SubNet is thus investigated as an extension to RQ2. Figure 5 shows that the most important subcluster found (rank 1) varies in size between a minimum of 8 and a maximum of 49 nodes, with the average size being 31 nodes.

Over 10 iterations, where each iteration involved training the GNN, finding a global explanation and detecting important subclusters, 2 iterations found an identical most important subcluster of 43 nodes. 2 other iterations found subclusters with significant overlap (one with 49 nodes and one with 40 nodes, sharing 40 nodes in common). The remaining 6

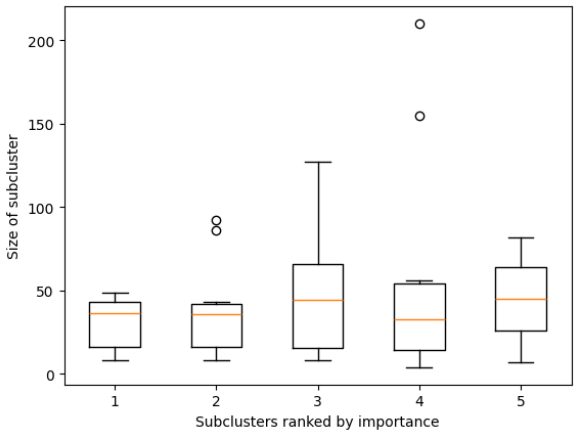

Figure 5: Sizes of the five most important subclusters found over 10 iterations (where each iteration consists of training, explaining and detecting the subclusters), with rank 1 being most important

iterations contained highly varying subclusters with no nodes in common.

This affirms the hypothesis based on the metric results that the explanations and the resulting subclusters found by GNN-SubNet are highly unstable and change from iteration to iteration on the same dataset. This variability that occurs every time the training and explanation are done afresh should be taken into account before conducting expert domain-specific analysis of the subnetworks. This could ensure that expert investigation is focused on stable, re-occuring subnetworks in the data.

5 Responsible Research

5.1 Ethical considerations of trust and safety

The use of artificial intelligence for biological and medical issues requires ethical consideration. As GNN-SubNet is a means of detecting disease subnetworks to guide potential expert research, it has an indirect impact on patients' healthcare. It is explicitly designed to be an 'expert-in-the-loop' approach, with the goal of explainability being to 'promote reliability and trust, ensuring that humans remain in control' [6].

While increasing trust in black-box models is the goal of explanable AI, this also has the potential to lead to overreliance and blind trust in models once they have been shown to be explainable. From this perspective, integrating empirical metrics into the process of using such tools acts as a clear sign for human experts on the quality and faithfulness of the explanations found. Reporting metric scores together with the results every time the tool is used, regardless of the explainer or dataset, indicates to the user to what extent the results can be trusted.

5.2 Access to code and data

The KIRC dataset used is fully anonymised and is originally sourced from TCGA (The Cancer Genome Atlas Program). The processed dataset as used in this study is available as part of the GNN-SubNet project on GitHub.[1].

The code developed for this study as an extension of GNN-SubNet is openly published on GitHub.[2]

5.3 Reproducibility and Integrity

The FAIR principles (Findable, Accessible, Interoperable, Reusable) for scientific data management [12] have been followed. Section 5.2 details how the code and datasets can be found and accessed. The code is based on standardized Python packages to ensure interoperability. To ensure reusability, documentation including instructions on how to set up and run the code, examples of usage have been provided. Within the code, clear descriptions of the functionality of each module result in a reusable and easily extendable tool.

The exact code used to run the experiments in this study are provided in the repository. While the results obtained by running them may not exactly be the same due to the randomness used in some evaluation metrics and the fluctuations in what the model learns each time it is trained, the results reported here are averaged over multiple iterations and it is expected that a similar average over iterations can reproduce this study in its entirety.

6 Conclusion

With the goal of implementing explainability evaluation metrics to assess explainable GNNs used for subnetwork detection, four metrics were implemented: RDT-fidelity and sparsity as defined in the BAGEL benchmark [5], and validity+ and validity-, newly defined in this study.

Focusing on the case study of GNN-SubNet, its explanations are found to be robust but very dense (with a high RDT-fidelity and low sparsity). They were also found to be highly variable, with different subnetworks detected each time the entire pipeline (training, explaining and evaluating) is executed.

The following findings can be more broadly applied to evaluate subnetwork detection in other domains:

- Metrics should be chosen based on their suitability to the domain at hand. For GNN-SubNet, metrics that remove edges and nodes from the graph are unsuitable, as they would make the PPI topology invalid.
- The results of the individual metrics are most useful in complement to each other:
  - A tradeoff is observed between RDT-fidelity and sparsity: the denser the explanation, the higher its RDT-fidelity. It can be useful to analyse the RDT-fidelity score at different levels of sparsity using hard-masks of different thresholds.
  - Validity+ and validity- also complement each other: in the case of GNN-SubNet, a high validity- score and low validity+ score showed that the model is able to rely on less important nodes to still reach the correct prediction.

[1] https://github.com/pievos101/GNN-SubNet

[2] https://github.com/Sucharitha-R/evaluating_gnn_subnet

- For the task of subnetwork detection, measuring the size and the stability of the subnetworks found is useful. Ideally, the subnetworks found are small (thus easy to interpret by experts) and are stable (do not vary across repeated runs).

Limitations and future work

Accuracy The average accuracy of the GNN was observed to be quite low, at 66.9%. The observed min/median/max accuracy of 60/66/77 contrasts with the 79/85/91 accuracy reported by [6]. Since the GNN itself is used in the implementation of the metrics to observe whether the classification of a certain input graph changes after perturbing it, the possibility that the low accuracy of the GNN lead to a negative bias in the model scores was investigated. For all four metrics, the Pearson coefficient calculated between the model accuracy and the metric score has an absolute value lower than 0.2, showing that no significant correlation exists between them. The reasons for this difference in accuracy is unclear, as the same KIRC training data and the same training setup was used as by [6]. The factors leading to the mismatch may potentially be a limitation to this evaluation of explainability.

Metrics The validity+ and validity- metrics, newly defined in this study, would benefit from additional investigation to determine their effectiveness. Validity+ could be re-defined such that it is interpreted as a range between 0 and 1, rather than by comparing with 0.5 as an ideal value.

Domain knowledge To improve the evaluation, the perturbations for RDT-fidelity could be sampled from a known biological distribution of the DNA methylation and gene expression of proteins. Similarly, the average values taken for the validity metrics could be based on domain knowledge rather than the empirical average from the given dataset.

Subnetwork variability The high variability observed in the disease subclusters found by GNN-SubNet when the training and explanation are done multiple times merits further investigation. Defining and implementing a dedicated metric to measure the variability in subclusters can augment the evaluation of explainers that are specifically used for the task of subnetwork detection.

Acknowledgements

I would like to thank my peers Hubert Janczak and Elena-Oana Milchi for their help with the code.

References

[1] E. Dai, T. Zhao, H. Zhu, *et al.*, *A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability*, 2023. arXiv: 2204.08570 [cs.LG].

[2] J. Kakkad, J. Jannu, K. Sharma, C. Aggarwal, and S. Medya, "A survey on explainability of graph neural networks," 2023. arXiv: 2306.01958 [cs.LG].

[3] J. Jiménez-Luna, M. Skalic, N. Weskamp, and G. Schneider, "Coloring molecules with explainable artificial intelligence for preclinical relevance assessment," *Journal of Chemical Information and Modeling*, vol. 61, no. 3, pp. 1083–1094, 2021. DOI: 10.1021/acs.jcim.0c01344.

[4] H. Cui, W. Dai, Y. Zhu, X. Li, L. He, and C. Yang, *Interpretable graph neural networks for connectome-based brain disorder analysis*, 2022. arXiv: 2207.00813 [q-bio.NC].

[5] M. Rathee, T. Funke, A. Anand, and M. Khosla, "Bagel: A benchmark for assessing graph neural network explanations," *arXiv preprint arXiv:2206.13983*, 2022.

[6] B. Pfeifer, A. Secic, A. Saranti, and A. Holzinger, "Gnn-subnet: Disease subnetwork detection with explainable graph neural networks," Jan. 2022. DOI: 10.1101/2022.01.12.475995.

[7] S. Jin, X. Zeng, F. Xia, W. Huang, and X. Liu, "Application of deep learning methods in biological networks," *Briefings in Bioinformatics*, vol. 22, no. 2, pp. 1902–1917, May 2020, ISSN: 1477-4054. DOI: 10.1093/bib/bbaa043. eprint: https://academic.oup.com/bib/article-pdf/22/2/1902/36654924/bbaa043.pdf. [Online]. Available: https://doi.org/10.1093/bib/bbaa043.

[8] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" *CoRR*, vol. abs/1810.00826, 2018. arXiv: 1810.00826. [Online]. Available: http://arxiv.org/abs/1810.00826.

[9] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNN explainer: A tool for post-hoc explanation of graph neural networks," *CoRR*, vol. abs/1903.03894, 2019. arXiv: 1903.03894. [Online]. Available: http://arxiv.org/abs/1903.03894.

[10] H. Yuan, H. Yu, S. Gui, and S. Ji, *Explainability in graph neural networks: A taxonomic survey*, 2022. arXiv: 2012.15445 [cs.LG].

[11] T. Funke, M. Khosla, M. Rathee, and A. Anand, *Zorro: Valid, sparse, and stable explanations in graph neural networks*, 2022. arXiv: 2105.08621 [cs.LG].

[12] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, p. 160018, Mar. 2016, ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. [Online]. Available: https://doi.org/10.1038/sdata.2016.18.