GGUN: Global Graph Understanding via Graph Neural Network Explanation for Identification of Influential Nodes in Healthcare

Tongtong Su [1] **Beilun Wang** [1 2]

Abstract

Investigating the globally influential nodes in a network is a fundamental task for healthcare analysis. Recent approaches omit either discrete graph structure or node attributes, providing incomprehensible or unreliable results. In this work, we introduce global graph understanding framework GGUN, specifically leveraging the explanatory power of graph neural networks. Following the path of the perturbation-based explanation, GGUN fills the gap between the continuous node features and discrete graph structure. Moreover, it obtains an efficient solution by relaxing the primal combinatorial objective function to overcome the local optimum and low efficiency drawbacks. Experimental evaluations show that GGUN outperforms baselines on both quantitative metrics and human-intelligible visualization.

1. Introduction

Graphs are widely used to model interacting objects, such as social networks, biomedical interactions, and cerebral biological neuronal network. One of the fundamental tasks in graph analysis is influential nodes detection, i.e., identifying a set of the most important nodes in a graph(12; 19). For example, focusing on infectious individuals in social network for epidemic prevention(8), or locating critical cerebral neurons for understanding brain diseases(7).

Many graph theory-based metrics have been introduced in the past to evaluate node importance in complex networks from different perspectives, such as Degree Centrality, Closeness Centrality, Eigenvector Centrality(3). However, these metrics ignore the attributive information in the graph data, namely, holding an oversimplified assumption that all nodes are identical and homogeneous. Such simplification weakens useful node information, e.g., the infectious disease test results, or the activity scope. One infected person with large range of activities deserves more attention than a healthy and indoorsy person even with higher degree in social network. To make full use of available data, we resort to a data-driven approach. Specifically, we consider investigating the explanatory power of graph neural networks (GNNs) for node classification task, which jointly model the topological and node attributive information of a graph.

There have been abundant researches on explaining deep neural networks (DNNs) (13; 1; 11). However, the strength of those approaches, beyond the high prediction power of DNN, relies on the capability of precise operations (e.g., gradient calculation) on the continuous data representation (13; 11). If applied to the adjacency matrix, which is the discrete representation of the graph topology, they will produce results that cannot be restored to structural information. Breaking the discreteness of adjacency matrix always results in useless explanations(18): giving everyone in the social network an soft importance score cannot reach targeted epidemic-control. Therefore, directly applying those explanation methods to GNN is inappropriate.

In this work, we aim to close the gap and investigate a proper explanation method that can effectively detect important nodes for graph understanding. The first thought should be adopting explanation methods that do not require to compute gradients over the features. To this end, we consider perturbation-based explanation (15; 10; 5; 17), which interprets a model by perturbing data features and observing its influence to prediction. However, the well-established works(17; 6) provide explanation for specific nodes, and thus unable to find globally important nodes. Specifically, GNNExplainer(17) can only remind someone that which of his friends have high risk of infection, however, cannot provide global strategy to epidemic prevention and control from management's view.

To provide global explanation, we consider selecting fixed number of nodes that, if removed, would change overall node predictions the most. The problem is formulated as a computational challenging combinatorial optimization prob-

[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China [2]Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China. Correspondence to: Beilun Wang <beilun@seu.edu.cn>.

Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML).

lem in Section 3.1. To address this challenge, we adopt two strategies in Section 3.2. Our method outperforms baselines on two case studies related to healthcare from both quantitative analysis and visualization.

2. Preliminaries

Node classification employing GNN. We consider the task of node classification in a single graph with node set $\mathcal{V}$. Let $G = (A, X)$ be an attributed graph, where $A \in \{0,1\}^{N\times N}$ is the adjacency matrix and $X^{N\times D}$ represents the node features. Without loss of generality, node-ids are denoted as $\mathcal{V} = 1, \cdots, N$ and feature-ids are denoted as $\mathcal{F} = 1, \cdots, D$. We focus on node classification employing graph convolution layers. Since our goal is to provide explanations rather than pursuing the state-of-the-art classification accuracy, we employ the well-established GCN (4). The hidden layer $l+1$ is defined as

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}), \tag{1}$$

where $\tilde{A} = A + I_N$ is the adjacency matrix adding self-loops. $\sigma(\cdot)$ is an activation function(usually ReLU). $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ is the degree matrix. $H^{(l)}$ is the node representation of layer l, in the first layer we have $H^{(0)} = X$. $W^{(l)}$ is a trainable parameter of layer l. Following (4), we consider a GCN with a single hidden layer:

$$Y = f_\theta(A, X) = \text{softmax}(\hat{A}\sigma(\hat{A}XW^{(1)})W^{(2)}), \tag{2}$$

where $\hat{A} = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$ is the symmetric normalized adjacency matrix. The model parameter $\theta = \left(W^{(1)}, W^{(2)}\right)$ is optimized by minimizing the cross-entropy loss towards health-related predictions.

For other GNN models like GAT (14) and GraphSAGE (2), the message passing functions are different, while the idea of aggregate neighbor information is consistent. According to (20), we use surrogate linear convolutional model for two reasons. On one hand, the model becomes tractable with only linear manipulation while preserving the idea of graph convolutions. On the other hand, the substitution encourages the transferability of selected important nodes.

Perturbation-based explanation A large fraction of existing explanation methods are based on sensitivity analysis, which aims to assign importance to input features given a prediction. Perturbation-based explanation is a possible way to study feature attribution (10). Through measuring the prediction difference by perturbing the input, the attributing factors can thus be located. The perturbation-based explanation ensure the output explanation has a precise meaning by restricting the perturbing. For instance, LIME (10) learns a local linear model to obtain the contributing score of each superpixel, GRACE (5) restrict the counterfactual sample conforming to features domain constraints of the dataset

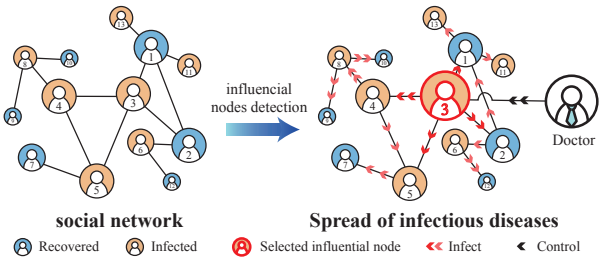


Figure 1. An example of influential node detection. Detecting influential individuals in social network to locating those highly infectious person.

by using projection operator. For our influential nodes detection task, we believe perturbation-based explanation can reserve the discreteness property of the graph structure.

3. Method

Overview. We aim to select a set of important nodes from an attributed graph dataset. We design the proposed method as per three properties. 1) Global explanation method in a data-driven manner leveraging the node feature and graph structure information simultaneously. We utilize the explanatory prediction power of GNN by first training a model on origin graph. Following the idea of *perturbation-based* explanation, i.e. final predictions will change significantly once important features are perturbed, we formulate the primal combinatorial optimization problem in Section 3.1. 2) Relaxing the infeasible bi-level combinatorial optimization into trainable convex optimization problem with two strategies in Section 3.2. 3) Two structural regularizers jointly encourage the approximate solution of the convex optimization close to the space of adjacency matrix, while a projection operator can easily recover the solution given the number of nodes.

3.1. Global graph understanding through perturbation-based explanation for GNN

We formulate our objective function by following the perturbation-based explanation scheme. In detail, we first minimize the rate of invariant predictions by removing a fixed size subset of node $\mathcal{S}$. Let $\mathbf{1}(\cdot)$ be the indicator function. We then have the score function as follows:

$$\underset{\mathcal{S}\subset\mathcal{V}}{\text{argmin}} \sum_{v\in V} \mathbf{1}(y_v^G = y_v^{G'}), \tag{3}$$

where G' is the perturbed graph after removing a set of nodes S from the original graph G, y_v^G and y_v^G are predicted labels of node v on origin and perturbed graph. However, the objective is non-differentiable and the output is insensitive

to minor revision to graph. Therefore, we convert the hard prediction result in Eq. (3) to a continuous probability form. This loss is so-called the Cumulative Classification Margin, which can be seen as maximizing the potential of flipping prediction of the whole graph:

$$\underset{\mathcal{S}\subset\mathcal{V}}{\operatorname{argmin}} \sum_{v\in\mathcal{V}} (P'_{v,c_{\text{old}}} - \max_{c\neq c_{\text{old}}} P'_{v,c}), \tag{4}$$

where $P' = [A'^2XW]^{N\times C}$ is the output probabilistic matrix on perturbed graph and A' is the adjacency matrix for the perturbed graph. c_{old} denotes the predicted class for v based on the original graph. σ and the softmax activation function are removed here for tractability and transferability.

Adding sparsity constraint. Since we only care about the top K influential nodes, it is naturally to add a sparsity constraint to restrict number of selected nodes. In addition, perturbing certain node is equivalent to isolating this node from others. Combining above constraints, we reformulate the problem by adding constraints on (3):

$$A'_{ij} = \begin{cases} 0, & i \in \mathcal{S} \text{ or } j \in \mathcal{S} \\ A_{ij}, & \text{otherwise} \end{cases}, \quad |\mathcal{S}| < K. \tag{5}$$

Here the operation of removing nodes in $\mathcal{S}$ is equivalent to setting the corresponding values in the adjacency matrix to 0. $|\cdot|$ refers to the cardinality of a set. Score function in (4) together with constraints in (5) compose our GGUN framework. However, there exists two critical issues to be settled. For one thing, the objective remains non-differentiable due to the inner maximum operator. For another thing, the framework is a combinatorial optimization problem with $O(N^K)$ possible solutions. Thus, the expensive search is infeasible for the real-world graph with even only thousands of nodes. To address above challenges, we propose two strategies to relax the non-differentiable combinatorial optimization problem into a differentiable convex optimization problem.

3.2. Two approximating strategies

Strategy I: Approximating the loss function with the original second-best label. In the works of targeted adversarial attack (20; 16), the runner-up, the label output with the second largest probability, is always considered to be an ideal choice for invasion target. This assumption can prominently increase the efficiency of the attacks. We apply the similar idea to alleviate first non-differential problem. In detail, given the original probabilistic matrix P, for each node v, the second-best class $\operatorname{argmax}_{c\neq c_{\text{old}}} P_{v,c}$ is assumed to be the second-best on the perturbed graph with probabilistic matrix P':

$$\underset{c\neq c_{\text{old}}}{\operatorname{argmax}} P'_{v,c} = \underset{c\neq c_{\text{old}}}{\operatorname{argmax}} P_{v,c} = c^*. \tag{6}$$

Note that P is able to be recorded during the training phase. We then simplify the solution to $\max_{c\neq c_{\text{old}}}[A'^2XW]_{v,c}$ in (4) as $[A'^2XW]_{v,c^*}$. We then rewrite the objective in (4) as the following differentiable formulation:

$$\underset{\mathcal{S}\subset\mathcal{V}}{\operatorname{argmin}}\, \mathcal{L} := \sum_{v\in\mathcal{V}} (P'_{v,c_{\text{old}}} - P'_{v,c^*}). \tag{7}$$

Strategy II: Relaxing discrete constraint In this work, we tend to utilize the convex relaxation to overcome the computational bottleneck of the combinatorial optimization. We jointly use two well-designed structural regularizers to preserve discrete property of A'.

The first constraint in Eq. (5) enforces symmetric sparsity of A' (diagonal crossing structure), since this structure indicates the removal of the corresponding node. To relax this constraint, we apply regularizer $||A'||_{2,1}$ to encourage column-wise sparsity, namely zero entries of A' to accumulate in one same column.

The left thing we need to consider is the similarity between the optimal A' and original A. Put differently, for the rest of nodes, the entry values of A' should be consistent with the original adjacency matrix A. Therefore, we choose Frobenius norm $||A - A'||_F$ to satisfy the above assumptions.

Therefore, the final formulation of our model GGUN can be formally written as the following:

$$\underset{A'}{\operatorname{argmin}}\, \mathcal{L}(A') + \lambda_1||A'||_{2,1} + \lambda_2||A - A'||_F, \tag{8}$$

where λ_1 and λ_2 are hyperparameters to control the number of selected nodes.

After obtaining the optimal A^*, we need to recover a set of selected nodes from A^* under budget K. Since a full-zero row or column refers to the removal of the corresponding node, we calculate $\mathbf{z}_i = ||A_{i:}||_2^2 + ||A_{:i}||_2^2$ for each node i and sort to get the smallest-K as the selected nodes.

4. Experiment

We conduct experiments on two synthetic datasets. Dataset generation is inspired by two health-related cases. The experiments aim to support two claims: **Claim 1:** The nodes selected by GGUN influence the predictions in the graph more significantly than baseline methods. **Claim 2:** Explanations provided by GGUN are closer to the ground-truth knowledge.

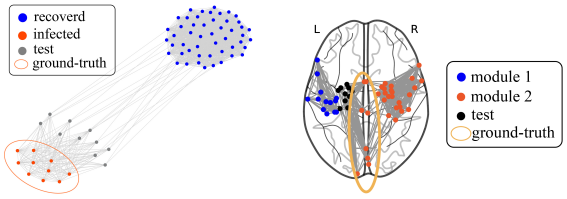
4.1. Settings

To verify **Claim 1**, we use the widely used metric Fidelity(18). Fidelity is defined as the difference of accuracy (or predicted probability) between the original predictions and the new predictions on the perturbed graph. This metric

is expected to reflect the significant change of performance when selected features are removed as these important input features identified by explanation techniques are discriminative to the model. Various versions of Fidelity have been proposed in(9), we choose **Fidelity-Prob (F-Prob):** $= \frac{1}{N}\sum_{i=1}^{N}(P_{i,y_i} - P'_{i,y_i})$, where y_i is the predicted label of node i before perturbation.

To verify **Claim 2**, the rules of dataset generating can be regarded as reasonable ground-truths. Thus, we can calculate **accuracy(Acc)**: selected node as a percentage of ground-truths. Accuracy metric can only be applied to synthetic dataset at the current stage.

Dataset generation We generate two synthetic attributed graphs for case study. The feature generation for different classes follows the multivariate normal distribution. **Case 1:** Consider a society network with 2 class of people: Infected and Recovered. There are 10 test people (grey points) fully connected with Infected. The origin predictions of these test samples are Infected. After removing those ground-truth influential Infected people, the predictions will change to Recovered. **Case 2:** Modularity in brain networks refers to internally densely connected clusters that are more weakly interconnected amongst each other. Consider a brain network with ground-truth critical hubs connect 2 modules of neurons. The origin predictions of these test samples are module 2. The origin predictions of these test neurons are module 2, since the labels on the right propagate through ground-truth influential hubs. After removing those hubs, the predictions will change to module 1.



(a) **Case 1**: Disease spread (b) **Case 2**: Brain network

Figure 2. Two healthcare cases. The rules of dataset generating can be regarded as reasonable ground-truths.

Baseline methods **1) Centrality**: Ensemble three graph theory-based node centrality metrics: Degree centrality, Eigenvector centrality, and Closeness centrality. **2) CCM-Greedy**: For our global-view score function **CCM**, even we do not find established work, greedy search can be regarded as a baseline method. We note it as CCM-Greedy in the following experiment. **3) GNNExplainer**(17): GNNExplainer learns edges and feature masks by maximizing the mutual information between the target node prediction based on the original graph and the masked graph. For fair comparison,

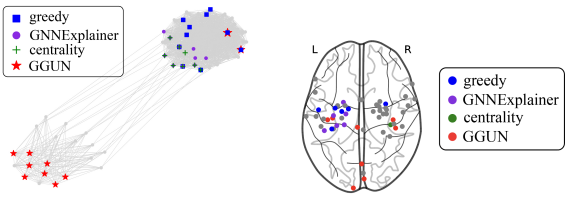
Table 1. Evaluation of F-Prob and Acc on two synthetic cases. GGUN outperforms baselines in both two metrics and is the closest to ground-truth.

	Synthetic 1		Synthetic 2	
Metrics	F-Prob	Acc	F-Prob	Acc
Ground-truth	0.7657	–	0.7443	–
Centrality	−0.6369	0/10	<u>0.5991</u>	0/8
CCM-Greedy	<u>0.1072</u>	0/10	−0.5226	0/8
GNNExplainer	−0.6396	0/10	0.5917	0/8
GGUN	**0.4288**	**8/10**	**0.6681**	**4/8**

We limit the choice to **Edge** and adopt a **Transverse** strategy. Specifically, we sum up the explanation for each node and then rank the edges and choose the top-K attached nodes.

4.2. Results

Table 1 shows the performance of GGUN compared with baselines on two synthetic dataset. GGUN outperforms three baselines on the metric of F-Prob obviously, which verifies our **Claim 1**. GGUN obtain $8/10$ and $4/10$ accuracy, while accuracy of all baselines equal to 0, which verifies our **Claim 2**. From Fig. 4.1 we can see that Centrality is attracted by fully-connected cluster, CCM-Greedy depends largely on the first selected point: once the first node has been deleted from the graph (always the nodes with the highest degree), the following selected nodes are those connected to the first one. GNNExplainer also performs poorly, indicating that the global influential nodes are not a simple summation of local results.



(a) **Case 1** result (b) **Case 2** result

Figure 3. Visualization of selected nodes. Only GGUN captures ground-truths successfully, while baselines fall into local optimum.

5. Future Work

In this work, we propose GGUN, a global graph understanding framework for identification of influential nodes in healthcare. Experimental results with visualizations on two synthetic datasets validate the superior explanatory performance of GGUN. In the near future, we plan to conduct experiments on more real world datasets with expert-annotated influential nodes. We should verify that GGUN is trustwor-

thy before applying it to high-stakes healthcare problems.

References

[1] ETMANN, C., LUNZ, S., MAASS, P., AND SCHÖNLIEB, C.-B. On the connection between adversarial robustness and saliency map interpretability. *arXiv preprint arXiv:1905.04172* (2019).

[2] HAMILTON, W. L., YING, R., AND LESKOVEC, J. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (2017), pp. 1025–1035.

[3] JACKSON, M. O. *Social and economic networks*. Princeton university press, 2010.

[4] KIPF, T. N., AND WELLING, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[5] LE, T., WANG, S., AND LEE, D. Grace: Generating concise and informative contrastive sample to explain neural network model's prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2020), pp. 238–248.

[6] LUO, D., CHENG, W., XU, D., YU, W., ZONG, B., CHEN, H., AND ZHANG, X. Parameterized explainer for graph neural network. *Advances in neural information processing systems 33* (2020), 19620–19631.

[7] MOUSSA, M. N., VECHLEKAR, C. D., BURDETTE, J. H., STEEN, M. R., HUGENSCHMIDT, C. E., AND LAURIENTI, P. J. Changes in cognitive state alter human functional brain networks. *Frontiers in human neuroscience 5* (2011), 83.

[8] NING, Y.-Z., LIU, X., CHENG, H.-M., AND ZHANG, Z.-Y. Effects of social network structures and behavioral responses on the spread of infectious diseases. *Physica A: Statistical Mechanics and Its Applications 539* (2020), 122907.

[9] POPE, P. E., KOLOURI, S., ROSTAMI, M., MARTIN, C. E., AND HOFFMANN, H. Explainability methods for graph convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 10772–10781.

[10] RIBEIRO, M. T., SINGH, S., AND GUESTRIN, C. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (2016), pp. 1135–1144.

[11] SELVARAJU, R. R., COGSWELL, M., DAS, A., VEDANTAM, R., PARIKH, D., AND BATRA, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 618–626.

[12] SHEIKHAHMADI, A., NEMATBAKHSH, M. A., AND SHOKROLLAHI, A. Improving detection of influential nodes in complex networks. *Physica A: Statistical Mechanics and its Applications 436* (2015), 833–845.

[13] SIMONYAN, K., VEDALDI, A., AND ZISSERMAN, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[14] VELIČKOVIĆ, P., CUCURULL, G., CASANOVA, A., ROMERO, A., LIO, P., AND BENGIO, Y. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[15] WACHTER, S., MITTELSTADT, B., AND RUSSELL, C. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech. 31* (2017), 841.

[16] XU, K., CHEN, H., LIU, S., CHEN, P.-Y., WENG, T.-W., HONG, M., AND LIN, X. Topology attack and defense for graph neural networks: An optimization perspective. *arXiv preprint arXiv:1906.04214* (2019).

[17] YING, R., BOURGEOIS, D., YOU, J., ZITNIK, M., AND LESKOVEC, J. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems 32* (2019), 9240.

[18] YUAN, H., YU, H., GUI, S., AND JI, S. Explainability in graph neural networks: A taxonomic survey. *arXiv preprint arXiv:2012.15445* (2020).

[19] ZHAO, Y., LI, S., AND JIN, F. Identification of influential nodes in social networks with community structure based on label propagation. *Neurocomputing 210* (2016), 34–44.

[20] ZÜGNER, D., AKBARNEJAD, A., AND GÜNNEMANN, S. Adversarial attacks on neural networks for graph data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2018), pp. 2847–2856.