Parameterized Explainer for Graph Neural Network

Dongsheng Luo[1*] **Wei Cheng**[2*] **Dongkuan Xu**[1] **Wenchao Yu**[2] **Bo Zong**[2]
Haifeng Chen[2] **Xiang Zhang**[1]
[1]The Pennsylvania State University
[2]NEC Labs America
[1]{dul262,dux19,xzz89}@psu.edu
[2]{weicheng,wyu,bzong,haifeng}@nec-labs.com

Abstract

Despite recent progress in Graph Neural Networks (GNNs), explaining predictions made by GNNs remains a challenging open problem. The leading method independently addresses the local explanations (i.e., important subgraph structure and node features) to interpret why a GNN model makes the prediction for a single instance, e.g. a node or a graph. As a result, the explanation generated is painstakingly customized for each instance. The unique explanation interpreting each instance independently is not sufficient to provide a global understanding of the learned GNN model, leading to the lack of generalizability and hindering it from being used in the inductive setting. Besides, as it is designed for explaining a single instance, it is challenging to explain a set of instances naturally (e.g., graphs of a given class). In this study, we address these key challenges and propose PGExplainer, a parameterized explainer for GNNs. PGExplainer adopts a deep neural network to parameterize the generation process of explanations, which enables PGExplainer a natural approach to explaining multiple instances collectively. Compared to the existing work, PGExplainer has better generalization ability and can be utilized in an inductive setting easily. Experiments on both synthetic and real-life datasets show highly competitive performance with up to 24.7% relative improvement in AUC on explaining graph classification over the leading baseline.

1 Introduction

Graph Neural Networks (GNNs) are powerful tools for representation learning of graph-structured data, such as social networks [46], document citation graphs [33], and microbiological graphs [44]. GNNs broadly adopt a message passing scheme to learn node representations by aggregating representation vectors of its neighbors [49, 16]. This scheme enables GNN to capture both node features and graph topology. GNN-based methods have achieved state-of-the-art performance in node classification, graph classification, and link prediction, etc [21, 45, 56].

Despite their remarkable effectiveness, the rationales of predictions made by GNNs are not easy for humans to understand. Since GNNs aggregate both node features and graph topology to make predictions, to understand predictions made by GNNs, important subgraphs and/or a set of features, which are also known as explanations, need to be uncovered. In the literature, although a variety of efforts have been undertaken to interpret general deep neural networks, existing approaches [6, 26, 12, 31, 43, 19, 20] in this line fall short in their ability to explain graph structures, which is essential for GNNs. Explaining predictions made by GNNs remains a challenging open problem, on which few methods have been proposed. The combinatorial nature of explaining graph structures makes it difficult to design models that are both robust and efficient. Recently, the first general model-agnostic
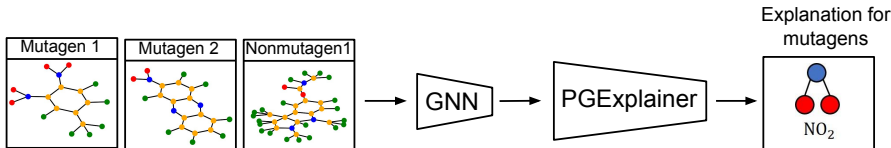
---

*Equal Contribution

Figure 1: PGExplainer provides human-understandable explanations for predictions made by GNNs. The left part shows the process of applying GNNs for graph classification on the MUTAG dataset. A GNN based model is trained to predict their mutagenic effects. As a post-hoc method, PGExplainer takes the trained GNN model as input and provides consistent explanations for predictions made by the GNN model. For the mutagen molecule graphs in the example, the explanation is the NO_2 group.

approach for GNNs, GNNExplainer [53], was proposed to address the problem. It takes a trained GNN and its predictions as inputs to provide interpretable explanations for a given instance, e.g. a node or a graph. The explanation includes a compact subgraph structure and a small subset of node features that are crucial in GNN's prediction for the target instance. Nevertheless, there are several limitations in the existing approach. First, GNNExplainer largely focuses on providing the local interpretability by generating a painstakingly customized explanation for a single instance individually and independently. The explanation provided by GNNExplainer is limited to the single instance, making GNNExplainer difficult to be applied in the inductive setting because the explanations are hard to generalize to other unexplained nodes. As pointed out in previous studies, models interpreting each instance independently are not sufficient to provide a global understanding of the trained model [19]. Furthermore, GNNExplainer has to be retrained for every single explanation. As a result, in real-life scenarios where plenty of nodes need to be interpreted, GNNExplainer would be time-consuming and impractical. Moreover, as GNNExplainer was developed for interpreting individual instances, the explanatory motifs are not learned end-to-end with a global view of the whole GNN model. Thus, it may suffer from suboptimal generalization performance. How to explain predictions of GNNs on a set of instances collectively and easily generalize the learned explainer model to other instances in the inductive setting remains largely unexplored in the literature.

To provide a global understanding of predictions made by GNNs, in this study, we emphasize the *collective* and *inductive* nature of this problem and present our method PGExplainer (Figure 1). PGExplainer is a general explainer that applies to any GNN based models in both transductive and inductive settings. Specifically, a generative probabilistic model for graph data is utilized in PGExplainer. Generative models have shown the power to learn succinct underlying structures from the observed graph data [23]. PGExplainer uncovers these underlying structures as the explanations, which is believed to make the most contribution to GNNs' predictions [35]. We model the underlying structure as edge distributions, where the explanatory graph is sampled. To collectively explain predictions of multiple instances, the generation process in PGExplainer is parameterized with a deep neural network. Since the neural network parameters are shared across the population of explained instances, PGExplainer is naturally applicable to provide model-level explanations for each instance with a global view of the GNN model. Furthermore, PGExplainer has better generalization power because a trained PGExplainer model can be utilized in an inductive setting to infer explanations of unexplained nodes without retraining the explanation model. This also makes PGExplainer much faster than the existing approaches.

Experimental results on both synthetic and real-life datasets demonstrate that PGExplainer can achieve consistent and accurate explanations, bringing up to 24.7% improvement in AUC over the SOTA method on explaining graph classification with significant speed-up.

2 Related work

Graph neural networks. Graph Neural networks (GNNs) have achieved remarkable success in various tasks, including node classification [25, 21, 45, 48], graph classification [10], and link prediction [56]. The study of GNNs was initiated in [17], and then extended in [41]. These methods iteratively aggregate neighbor information to learn node representations until reaching a static state. Inspired by the success of convolutional neural networks (CNNs) in computer vision, attempts of applying convolutional operations to graphs were derived based on graph spectral theory [4] and

graph Fourier transformation [42]. In recent work, GNNs broadly encode node features as messages and adopt the message passing mechanism to propagate and aggregate them along edges to learn node/graph representations, which are then utilized for downstream tasks [25, 10, 41, 29, 45, 34, 47]. For efficiency consideration, localized filters were proposed to reduce computation cost [21]. The self-attention mechanism was introduced to GNNs in GAT to differentiate the importance of neighbors [45]. Xu. et al. analyzed the relationship between GNNs and Weisfeiler-Lehman graph isomorphism test, and showed the express power of GNNs [49].

Explaining GNNs. Interpretability and feature selection have been extensively addressed in neural networks. Methods demystifying complicated deep learning models can be grouped into two main families, whitebox and blackbox [19, 20]. Whitebox mechanisms mainly focus on yielding explanations for individual predictions. Forward and backward propagation based methods are used routinely in whitebox mechanisms. Forward propagation based methods broadly perturb the input and/or hidden representations and check the corresponding updating results in the downstream task [8, 14, 28]. The underlying intuition is that the outputs of the downstream task are likely to significantly change if important features are occluded. Backward propagation based methods, in general, infer important features from the gradients of the deep neural networks. They compute weights of features by propagating the gradients from the output back to the input. Blackbox methods generate explanations by locally learning interpretable models, such as linear models, and additive models to approximate the predictions [32, 5].

Following the line of forward propagation methods, GNNExplainer initiates the research on explaining predictions on graph-structured data [53]. It excludes certain edges and node features to observe the changes in node/graph classification. Explanations (subgraphs/important features) are extracted by maximizing the mutual information between the distribution of possible subgraphs and the GNN's prediction. However, similar to other forward propagation methods, GNNExplainer generates customized explanations for single instance prediction independently, making it insufficient to provide a global understanding of the trained GNN model. Besides, it is naturally difficult to be applied in the inductive setting and provide explanations for multiple instances. XGNN provides model-level explanations without preserving the local fidelity [55]. Thus, the generated explanation may not be a substructure of the real input graph. On the other hand, PGExplainer can provide an explanation for each instance with a global view of the GNN model, which can preserve the local fidelity.

Graph generation. PGExplainer learns a probabilistic graph generative model to provide explanations for GNNs. The first model generating random graph is the Erdős–Rényi model [15, 11]. In the random graph proposed by Gilbert, each potential edge is independently chosen from a Bernoulli distribution. Some approaches generate graphs with certain properties reflected, such as pairwise distances betweenness [7], node degree distribution [27], and spectral properties [22, 2]. In recent years, deep learning models have shown great potential to generate graphs with complex properties preserved [54, 18, 35]. However, these methods mainly aim to generate graphs that reflect certain properties in the training graphs.

3 Background

We first describe notations, and then provide some background on graph neural networks.

Notations. Let $G = (\mathcal{V}, \mathcal{E})$ represent the graph with $\mathcal{V} = \{v_1, v_2...v_N\}$ denoting the node set and $\mathcal{E} \in \mathcal{V} \times \mathcal{V}$ as the edge set. The numbers of nodes and edges are denoted by N and M, respectively. A graph can be described by an adjacency matrix $\mathbf{A} \in \{0, 1\}^{N \times N}$, with $a_{ij} = 1$ if there is an edge connecting node i and j, and $a_{ij} = 0$ otherwise. Nodes in $\mathcal{V}$ are associated with the d-dimensional features, denoted by $\mathbf{X} \in \mathbb{R}^{N \times d}$.

Graph neural networks. GNNs adopt the message-passing mechanism to propagate and aggregate information along edges in the input graph to learn node representations [25, 21, 45, 16]. Each GNN layer includes three essential steps. First, at the propagation step of the i-th GNN layer, for each edge (i, j), GNN computes a message $\mathbf{m}_{ij}^l = \text{Message}(\mathbf{h}_i^{l-1}, \mathbf{h}_j^{l-1})$, where $\mathbf{h}_i^{l-1}$ and $\mathbf{h}_j^{l-1}$ are representations of v_i and v_j in previous layer, respectively. Second, at the aggregation step, for each node v_i, GNN aggregates messages received from its neighbor nodes, denoted by $\mathcal{N}_i$, with an aggregation function $\mathbf{m}_{i\cdot}^l = \text{aggregation}(\{\mathbf{m}_{ij}^l | j \in \mathcal{N}_i\})$. Last, at the updating step, GNN updates the vector representation for each node v_i via $\mathbf{h}_i^l = \text{update}(\mathbf{m}_{i\cdot}^l, \mathbf{h}_i^{l-1})$, a function taking the

aggregated message and the representation of itself as inputs. Hidden representation of the last GNN layer serves as the final node representation: $\mathbf{z}_i = \mathbf{h}_i^L$, which is then used for downstream tasks, such as node/graph classification and link prediction.

4 The PGExplainer

In this section, we introduce PGExplainer. Different from GNNExplainer which provides explanations on both structure and features, PGExplainer focuses on explanation on graph structures because feature explanation in GNNs is similar to that in non-graph neural networks, which has been extensively studied in the literature [1, 19, 32, 39, 8, 14, 28]. PGExplainer is flexible and applicable to interpret all kinds of GNNs. We start with the learning objective of PGExplainer (Section 4.1) and then present the reparameterization strategy for efficient optimization (Section 4.2). In Section 4.3, we specify particular instantiations to understand GNNs on node and graph classifications. Detailed algorithms can be found in the Appendix.

4.1 The learning objective

The literature has shown that real-life graphs are with underling structures [35, 37]. To explain predictions made by a GNN model, we divide the original input graph G_o into two subgraphs: $G_o = G_s + \Delta G$, where G_s presents the underlying subgraph that makes important contributions to GNN's predictions, which is the expected *explanatory graph*, and ΔG consists of the remaining task-irrelevant edges for predictions made by the GNN. Following [53], PGExplainer finds G_s by maximizing the mutual information between the GNN's predictions and the underlying structure G_s:

$$\max_{G_s} \mathrm{MI}(Y_o, G_s) = H(Y_o) - H(Y_o | G = G_s), \tag{1}$$

where Y_o is the prediction of the GNN model with G_o as the input. The mutual information quantifies the probability of prediction Y_o when the input graph to the GNN model is limited to the explanatory graph G_s. The intuition behind comes from the traditional forward propagation based methods for the whitebox explanation [8]. For example, if removing an edge (i, j) dramatically changes the prediction in the GNN, then this edge is important and should be included in G_s. Otherwise, it can be considered as irrelevant edge for the GNN model to make the prediction. Since $H(Y_o)$ is only related to the GNN model whose parameters are fixed in the explanation stage, the objective is equivalent to minimizing the conditional entropy $H(Y_o | G = G_s)$.

However, the direct optimization of the above objective function is intractable as there are 2^M candidates for G_s. Thus, we consider a relaxation by assuming that the explanatory graph is a Gilbert random graph [15], where selections of edges from the original input graph G_o are conditionally independent to each other. Let $e_{ij} \in \mathcal{V} \times \mathcal{V}$ be the binary variable indicating whether the edge is selected, with $e_{ij} = 1$ if the edge (i, j) is selected, and 0 otherwise. Let G be the random graph variable. Based on the above assumption, the probability of a graph G can be factorized as:

$$P(G) = \Pi_{(i,j) \in \mathcal{E}} P(e_{ij}). \tag{2}$$

A straightforward instantiation of $P(e_{ij})$ is the Bernoulli distribution $e_{ij} \sim Bern(\theta_{ij})$. $P(e_{ij} = 1) = \theta_{ij}$ is the probability that edge (i, j) exists in G. With this relaxation, we thus can rewrite the objective as:

$$\min_{G_s} H(Y_o | G = G_s) = \min_{G_s} \mathbb{E}_{G_s}[H(Y_o | G = G_s)] \approx \min_{\Theta} \mathbb{E}_{G_s \sim q(\Theta)}[H(Y_o | G = G_s)] \tag{3}$$

where $q(\Theta)$ is the distribution of the explanatory graph parameterized by θ's.

4.2 The reparameterization trick

Due to the discrete nature of G_s, we relax edge weights from binary variables to continuous variables in the range $(0, 1)$ and adopt the reparameterization trick to efficiently optimize the objective function with gradient-based methods [24]. We approximate the sampling process $G_s \sim q(\Theta)$ with a determinant function of parameters Ω, temperature τ, and an independent random variable ϵ: $G_s \approx \hat{G}_s = f_\Omega(G_o, \tau, \epsilon)$. The temperature τ is used to control the approximation. Here we utilize the
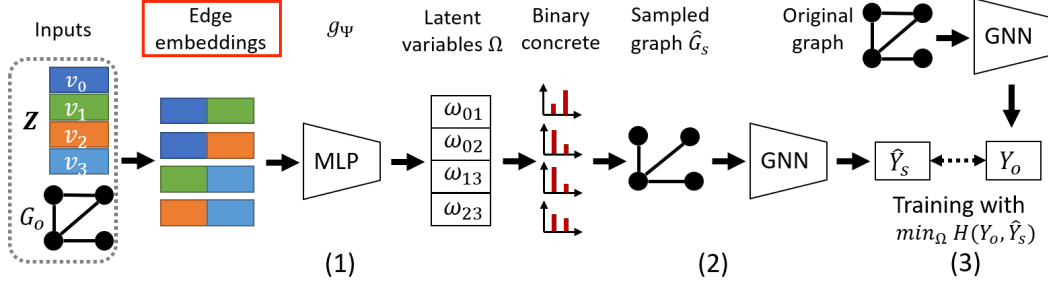
Figure 2: Illustration of PGExplainer for explaining GNNs on graph classification. (1) The left part demonstrates the explanation network. It takes node representations $\mathbf{Z}$ as well as the original graph G_o as inputs to compute Ω, the latent variables in edge distributions. Edge distributions are severed as the explanation. In case that an explanatory subgraph is wanted, we select <u>top-ranked edges according to latent variables Ω.</u> (2) A random graph $\hat{G}_s$ is sampled from edge distributions and then feed to the trained GNN model to get the prediction $\hat{Y}_s$. (3) Parameter Ψ in the explanation network is optimized with cross-entropy between the original prediction Y_o and the updated prediction $\hat{Y}_s$.

binary concrete distribution as the instantiation [36]. Specifically, the weight $\hat{e}_{ij} \in (0,1)$ of edge (i,j) in $\hat{G}_s$ is calculated by:

$$\epsilon \sim \text{Uniform}(0,1), \quad \hat{e}_{ij} = \sigma((\log \epsilon - \log(1-\epsilon) + \omega_{ij})/\tau), \tag{4}$$

where $\sigma(\cdot)$ is the Sigmoid function, and $\omega_{ij} \in \mathbb{R}$ is the parameter. When $\tau \to 0$, the weight $\hat{e}_{ij}$ is binarized with $\lim_{\tau \to 0} P(\hat{e}_{ij} = 1) = \frac{\exp(\omega_{ij})}{1+\exp(\omega_{ij})}$. Since $P(e_{ij} = 1) = \theta_{ij}$, by choosing $\omega_{ij} = \log \frac{\theta_{ij}}{1-\theta_{ij}}$, we have $\lim_{\tau \to 0} \hat{G}_s = G_s$. This demonstrates the rationality of using binary concrete distribution to approximate the Bernoulli distribution. Moreover, with temperature $\tau > 0$, the objective function is smoothed with a well-defined gradient $\frac{\partial \hat{e}_{ij}}{\partial \omega_{ij}}$. Thus, with reparameterization, the objective in Eq. (3) becomes:

$$\min_{\Omega} \mathbb{E}_{\epsilon \sim \text{Uniform}(0,1)} H(Y_o | G = \hat{G}_s) \tag{5}$$

Considering efficient optimization, we follow [53] to modify the conditional entropy with cross-entropy $H(Y_o, \hat{Y}_s)$, where $\hat{Y}_s$ is the prediction of the GNN model with $\hat{G}_s$ as the input. With the above relaxations, we adopt Monte Carlo to approximately optimize the objective function:

$$\min_{\Omega} \mathbb{E}_{\epsilon \sim \text{Uniform}(0,1)} H(Y_o, \hat{Y}_s) \approx \min_{\Omega} -\frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C} P(Y_o = c) \log P(\hat{Y}_s = c)$$

$$= \min_{\Omega} -\frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C} P_\Phi(Y = c | G = G_o) \log P_\Phi(Y = c | G = \hat{G}_s^{(k)}), \tag{6}$$

where Φ denotes the parameters in the trained GNN, K is the total number of sampled graph, C is the number of labels, and $\hat{G}_s^{(k)}$ is the k-th sampled graph with Eq. (4), parameterized by Ω.

4.3 Explanation of graph neural networks with a global view

Although explanations provided by the leading method GNNExplainer [53] preserve the local fidelity, they do not help to understand the general picture of the model across a population [50]. Furthermore, various GNN based models have been applied to analyze graph data with millions of instances [52], the cost of applying local explanations one-by-one can be prohibitive with such large datasets in practice. On the other hand, explanations with a global view of the model ascertain users' trust [39]. Furthermore, these models can generalize explanations to new instances without retraining, making it more efficient to explain large scale datasets.

To have a global view of a GNN model, our method collectively explains predictions made by a trained model on multiple instances. Instead of treating Ω in Eq. (6) as independent variables, we

utilize a parameterized network to learn to generate explanations from the trained GNN model, which also applies to unexplained instances. In general, GNN based models first learn node representations and then feed the vector representations to downstream tasks [21, 25, 45]. We denote these two functions by $\text{GNNE}_{\Phi_0}(\cdot)$ and $\text{GNNC}_{\Phi_1}(\cdot)$, respectively. For GNNs without explicit classification layers, we use the last layer instead. As a result, we can represent a GNN model with:

$$\mathbf{Z} = \text{GNNE}_{\Phi_0}(G_o, \mathbf{X}), \;\; Y = \text{GNNC}_{\Phi_1}(\mathbf{Z}). \tag{7}$$

$\mathbf{Z}$ is the matrix of node representations encoding both features and structure of the input graph, which is used as an input in the explanation network to calculate the parameter Ω:

$$\Omega = g_\Psi(G_o, \mathbf{Z}). \tag{8}$$

Ψ denotes parameters in the explanation network, which is shared by all edges among the population. Therefore, PGExplainer can be utilized to collectively provide explanations for multiple instances. Specifically, in the collective setting with instance set $\mathcal{I}$, the objective of PGExplainer is:

$$\min_{\Psi} -\sum_{i \in \mathcal{I}} \sum_{k=1}^{K} \sum_{c=1}^{C} P_\Phi(Y = c | G = G_o^{(i)}) \log P_\Phi(Y = c | G = \hat{G}_s^{(i,k)}), \tag{9}$$

where $G^{(i)}$ is the input graph and $\hat{G}_s^{(i,k)}$ is the k-th sampled graph with Eq. (4 ,8) for i-th instance. We consider both graph and node classifications and specify an instantiation for each task. Solutions for other tasks, such as link prediction, are similar and thus omitted.

Explanation network for node classification. Considering that explanations for nodes in a graph may appear diverse structures [53], especially for nodes with different labels. For example, an edge (i, j) is important for the prediction of node u, but not for another node v. Based on this motivation, we implement the $\Omega = g_\Psi(G_o, \mathbf{Z})$ to explain the prediction of node v with:

$$\omega_{ij} = \text{MLP}_\Psi([\mathbf{z}_i; \mathbf{z}_j; \mathbf{z}_v]). \tag{10}$$

MLP_Ψ is a multi-layer neural network parameterized with Ψ and $[\cdot; \cdot]$ is the concatenation operation.

Explanation network for graph classification. For graph level tasks, each graph is considered as an instance. The explanation of the prediction of a graph is not conditional to a specific node. Therefore, we specify the $\Omega = g_\Psi(G_o, \mathbf{Z})$ for graph classification as:

$$\omega_{ij} = \text{MLP}_\Psi([\mathbf{z}_i; \mathbf{z}_j]). \tag{11}$$

With the graph classification as an example, the architecture of PGExplainer is shown in Figure 2. The algorithms of PGExplainer for node and graph classification can be found in Appendix.

Computational complexity. PGExplainer is more efficient than GNNExplainer for two reasons. First, PGExplainer learns a latent variable for each edge in the original input graph with a neural network parameterized by Ψ, which is shared by all edges in the population of input graphs. Different from GNNExplainer, whose parameter size is linear to the number of edges, the number of parameters in PGExplainer is irrelevant to the size of the input graph, which makes PGExplainer applicable to large scale datasets. Further, since the explanation is shared among a population, a trained PGExplainer can be utilized in the inductive setting to explain new instances without retraining the explainer. To explain a new instance with $|\mathcal{E}|$ edges in the input graph, the time complexity of PGExplainer is $O(|\mathcal{E}|)$. As a comparison, GNNExplainer has to retrain for the new instance, leading to the time complexity of $O(T|\mathcal{E}|)$, where T is the number of epochs for retraining.

4.4 Regularization

The framework of PGExplainer is flexible with various regularization terms to preserve desired properties on the explanation. We now discuss the regularization terms as well as their principles.

Size and entropy constraints. Following [53], to obtain compact and succinct explanations, we impose a constraint on the explanation size by adding $||\Omega||_1$, the l_1 norm on latent variables Ω, as a regularization term. Besides, element-wise entropy is also be applied to further achieve discrete edge weights [53].

6

Next, We provide more regularization terms that are compatible with PGExplainer. Note that for a fair comparison, the following regularization terms are not utilized in experimental studies in Section 5.

Budget constraint. To obtain a compact explanation, the l_1 norm on latent variables Ω is introduced, which penalizes all edge weights to sparsify the explanatory graph. In cases that a predefined budget B is available, for example, $|G_s| \leq B$, we could modify the size constraint to budget constraint:

$$R_b = \text{ReLU}(\sum_{(i,j)\in\mathcal{E}} \hat{e}_{ij} - B). \tag{12}$$

When the size of the explanatory graph is smaller than the budget B, the budget regularization $R_b = 0$. On the other hand, it works similarly to the size constraint when out of budget.

Connectivity constraint. In many real-life scenarios, determinant motifs are expected to be connected. Although it is claimed that GNNExplainer empirically tends to detect a small connected subgraph, the explicit constraints are not provided [53]. We propose to implement the connected constraint with the cross-entropy of adjacent edges, which connect to the same node. For instance, (i,j) and (i,k) both connected to the node i. The motivation is that is (i,j) is selected in the the explanatory graph, then its adjacent edge (i,k) should also be included. Formally, we design the connectivity constraint as:

$$H(\hat{e}_{ij}, \hat{e}_{ik}) = -[(1-\hat{e}_{ij})\log(1-\hat{e}_{ik}) + \hat{e}_{ij}\log\hat{e}_{ik}]. \tag{13}$$

5 Experimental study

In this section, we evaluate our PGExplainer with a number of experiments. We first describe synthetic and real-world datasets, baseline methods, and experimental setup. Then, we present the experimental results on explanations of both node and graph classification. With qualitative and quantitative evaluations, we demonstrate that our PGExplainer can improve the SOTA method up to 24.7% in AUC on explaining graph classification. At the same time, with a trained explanation network, our PGExplainer is significantly faster than the baseline when explaining unexplained instances. We also provide extended experiments to show deep insights of PGExplainer in the appendix. The code and data used in this work are available [2].

5.1 Datasets

We follow the setting in GNNExplainer and construct four kinds of node classification datasets, BA-Shapes, BA-Community, Tree-Cycles, and Tree-Grids [53]. Furthermore, we also construct a graph classification datasets, BA-2motifs. Illustration of synthetic datasets is shown in Table 2. (1) BA-Shapes is a single graph consisting of a base Barabasi-Albert (BA) graph with 300 nodes and 80 "house"-structured motifs. These motifs are attached to randomly selected nodes from the BA graph. After that, random edges are added to perturb the graph. Nodes features are not assigned in BA-Shapes. Nodes in the base graph are labeled with 0; the ones locating at the top/middle/bottom of the "house" are labeled with 1,2,3, respectively. (2) BA-Community dataset consists of two BA-Shapes graphs. Two Gaussian distributions are utilized to sample node features, one for each BA-Shapes graph. Nodes are labeled based on their structural roles and community memberships, leading to 8 classes in total. (3) In the Tree-Cycles dataset, an 8-level balanced binary tree is adopted as the base graph. A set of 80 six-node cycle motifs are attached to randomly selected nodes from the base graph. (4) Tree-Grid is constructed in the same way as TREE-CYCLES, except that 3-by-3 grid motifs are used to replace the cycle motifs. (5) For graph classification, we build the BA-2motifs dataset of 800 graphs. We adopt the BA graphs as base graphs. Half graphs are attached with "house" motifs and the rest are attached with five-node cycle motifs. Graphs are assigned to one of 2 classes according to the type of attached motifs.

We also include a real-life dataset, MUTAG, for graph classification, which is also used in previous work [53]. It consists of 4,337 molecule graphs. Each graph is assigned to one of 2 classes based on its mutagenic effect [53, 40]. As discussed in [53, 9], carbon rings with chemical groups NH_2 or NO_2

---

are known to be mutagenic. We observe that carbon rings exist in both mutagen and nonmutagenic graphs, which are not discriminative. Thus, we can treat carbon rings as the shared base graphs and NH_2, NO_2 as motifs for the mutagen graphs. There are no explicit motifs for nonmutagenic ones. Table 1 shows the statistics of synthetic and real-life datasets.

Table 1: Dataset statistics

	Node Classification				Graph Classification	
	BA-Shapes	BA-Community	Tree-Cycles	Tree-Grid	BA-2motifs	MUTAG
#graphs	1	1	1	1	1,000	4,337
#nodes	700	1,400	871	1,231	25,000	131,488
#edges	4,110	8,920	1,950	3,410	51,392	266,894
#labels	4	8	2	2	2	2

5.2 Baselines and experimental setup

Baselines. We compare with the following baseline methods, GNNExplainer [53], a gradient-based method (GRAD) [53], graph attention network (ATT) [45], and Gradient [38]. (1) GNNExplainer is a post-hoc method providing explanations for every single instance. (2) GRAD learns weights of edges by computing gradients of GNN's objective function w.r.t. the adjacency matrix. (3) ATT utilizes self-attention layers to distinguish edge attention weights in the input graph. Each edge's importance is obtained by averaging its attention weights across all attention layers. (4) We first adopt Gradient in [38] to calculate the importance of each node, then calculate the importance of an edge by average the connected nodes' importance scores.

Experimental setup. We follow the experimental settings in GNNExplainer [53]. Specifically, for post-hoc methods including ATT, GNNExplainer, and PGExplainer, we first train a three-layer GNN and then apply these methods to explain predictions made by the GNN. Since weights in attention layers are jointly optimized with the GNN model in ATT, we thus train another GNN model with self-attention layers. We follow [1] to tune temperature τ. We refer readers to the Appendix for more training details.

5.3 Results

The results of comparative evaluation experiments on both synthetic and real-life datasets are summarized in Table 2. In these datasets, node/graph labels are determined by the motifs, which are treated as ground truth explanations. These motifs are utilized to calculate explanation accuracy for PGExplainer as well as other baselines.

Qualitative evaluation. We choose an instance for each dataset and visualize its explanations given by GNNExplainer and PGExplainer in Table 2. In these explanations, bold black edges indicate top-K edges ranked by their importance weights, where K is set to the number of edges inside motifs for synthetic datasets and 10 for MUTAG [53]. As demonstrated in these figures, the whole motifs, such as "house" in BA-Shapes and BA-Community, cycles in Tree-Cycles and BA-2motifs, grids in Tree-Grid, and NO_2 groups in MUTAG are correctly identified by PGExplainer. On the other hand, some important edges are missing in the explanations given by GNNExplainer. For example, the explanation provided by GNNExplainer for the instance in MUTAG contains the carbon rings and part of a NO_2 group. However, the carbon rings appear frequently in both mutagen and nonmutagenic graphs, which are not discriminative. Conversely, PGExplainer correctly identifies both NO_2 groups.

Quantitative evaluation. We follow the experimental settings in GNNExplainer [53] and formalize the explanation problem as a binary classification of edges. We treat edges inside motifs as positive edges, and negative otherwise. Importance weights provided by explanation methods are considered as prediction scores. A good explanation method assigns high weights to edges in the ground truth motifs than the ones outside. AUC is adopted as the metric for quantitative evaluation. Especially, for the MUTAG dataset, we only consider the mutagen graphs because no explicit motifs exist in nonmutagenic ones. For PGExplainer, we repeat each experiment 10 times and report the average AUC scores and standard deviations here.

From the table, we have the following observations. PGExplainer achieves SOTA performances in all scenarios and the accuracy gains are up to 13.0% in node classification and 24.7% in graph

| | Node Classification | | | | Graph Classification | |
	BA-Shapes	BA-Community	Tree-Cycles	Tree-Grid	BA-2motifs	MUTAG
Base						
Motifs						
Features	None	$\mathcal{N}(\mu_l, \sigma_l)$	None	None	None	Atom types
Visualization						
Explanations by GNN-Explainer						
Explanations by PG-Explainer						
Explanation AUC						
GRAD	0.882	0.750	0.905	0.612	0.717	0.783
ATT	0.815	0.739	0.824	0.667	0.674	0.765
Gradient	-	-	-	-	0.773	0.653
GNNExplainer	0.925	0.836	0.948	0.875	0.742	0.727
PGExplainer	**0.963**±0.011	**0.945**±0.019	**0.987**±0.007	**0.907**±0.014	**0.926**±0.021	**0.873**±0.013
Improve	4.1%	13.0%	4.1%	3.7%	24.7%	11.5%
Inference Time (ms)						
GNNExplainer	650.60	696.61	690.13	713.40	934.72	409.98
PGExplainer	10.92	24.07	6.36	6.72	80.13	9.68
Speed-up	59x	29x	108x	106x	12x	42x

Table 2: Illustration of different datasets together with performance evaluation of PGExplainer and other baselines. BA-Shapes, BA-Community, Tree-Cycles, Tree-Grid are datasets for node classification [53]. Node labels are represented by their colors. BA-2motifs and MUTAG datasets are used for graph classification. Graphs with "house" motifs are labeled with 0 and the ones with cycles are with 1 in BA-2motifs dataset. NH_2, NO_2 are treated as motifs of the mutagen graphs in MUTAG. Explanations extracted by GNNExplainer and PGExplainer are also shown as case studies.

classification. Compared to GNNExplainer, which tackles instances independently thus can only achieve suboptimal explanations, PGExplainer utilizes a parameterized explanation network based upon graph generative model to collectively provide explanations for multiple instances. As a result, PGExplainer can have a global view of the GNNs, which answers why PGExplainer can outperform GNNExplainer by relatively large margins.

Efficiency evaluation. Explanation network in PGExplainer is shared across the population of instances. Thus, a trained PGExplainer can be utilized to explain new instances in the inductive setting. We denote the time to explain a new instance with a trained explanation method by inference time. Since GNNExplainer has to retrain the model, we also count the training time here. The running time comparison in Table 2 shows that PGExplainer can speed up the computation significantly, up to 108 times faster than GNNExplainer, which makes PGExplainer more practical for large-scale datasets.

Further experiments on the inductive performance and effects of regularization terms are in Appendix.

6 Conclusion

We present PGExplainer, a parameterized method to provide a global understanding of any GNN models on arbitrary machine learning tasks by collectively explaining multiple instances. We show that PGExplainer can leverage the representations produced by GNNs to learn the underlying subgraphs that are important to the predictions made by GNNs. Furthermore, PGExplainer is more efficient due to its capacity to explain GNNs in the inductive settings, which makes PGExplainer more practical in real-life applications.

Acknowledgement

Broader impact

Graph neural networks are powerful tools that have been applied in various real-world applications, including community detection, recommendation systems, computer vision, and natural language processing [3, 13, 30, 51]. Our work can not only provide interpretable explanations with local fidelity for predictions made by GNN models, but also improve the global understanding of the model.

There are several broader impacts of using our method to explain predictions made by GNNs. First, our method can increase the transparency of applying GNNs for decision-critical applications, such as drug discovery and diagnosis. As a result, our method can help alleviate safety, and fairness risks. For example, as we show in our experiments, we could correctly identify motifs that have determinant effects on the mutagenicity of molecules. On the other hand, our method also puts GNN models at a high risk of being attacked. Our method extracts subgraphs that are important to GNNs' behaviors. Disturbing these parts leads to significant changes in GNNs' predictions. Besides, increasing the interpretability of GNNs may cause automation bias, such as an undue trust on GNN models.

References

[1] Abubakar Abid, Muhammad Fatih Balin, and James Zou. Concrete autoencoders for differentiable feature selection and reconstruction. *arXiv preprint arXiv:1901.09346*, 2019.

[2] Raman Arora and Jalaj Upadhyay. On differentially private graph sparsification and applications. In *NeurIPS*, pages 13378–13389, 2019.

[3] Joan Bruna and X Li. Community detection with graph neural networks. *Stat*, 1050:27, 2017.

[4] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.

[5] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *KDD*, pages 1721–1730, 2015.

[6] Jianbo Chen, Le Song, Martin J Wainwright, and Michael I Jordan. Learning to explain: An information-theoretic perspective on model interpretation. *arXiv preprint arXiv:1802.07814*, 2018.

[7] L Paul Chew. There are planar graphs almost as good as the complete graph. *Journal of Computer and System Sciences*, 39(2):205–219, 1989.

[8] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *NIPS*, pages 6967–6976, 2017.

[9] Asim Kumar Debnath, Rosa L Lopez de Compadre, Gargi Debnath, Alan J Shusterman, and Corwin Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry*, 34(2):786–797, 1991.

[10] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *NIPS*, pages 3844–3852, 2016.

[11] P ERDdS and A R&wi. On random graphs i. *Publ. Math. Debrecen*, 6(290-297):18, 1959.

[12] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2019.

[13] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. Graph neural networks for social recommendation. In *The Web Conference*, pages 417–426, 2019.

[14] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, pages 3429–3437, 2017.

[15] Edgar N Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[16] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *ICML*, pages 1263–1272, 2017.

[17] Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. In *IJCNN*, volume 2, pages 729–734. IEEE, 2005.

[18] Aditya Grover, Aaron Zweig, and Stefano Ermon. Graphite: Iterative generative modeling of graphs. In *ICML*, 2019.

[19] Wenbo Guo, Sui Huang, Yunzhe Tao, Xinyu Xing, and Lin Lin. Explaining deep learning models–a bayesian non-parametric approach. In *NeurIPS*, pages 4514–4524, 2018.

[20] Wenbo Guo, Dongliang Mu, Jun Xu, Purui Su, Gang Wang, and Xinyu Xing. Lemna: Explaining deep learning based security applications. In *CCS*, pages 364–379, 2018.

[21] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NIPS*, pages 1024–1034, 2017.

[22] Gecia Bravo Hermsdorff and Lee M Gunderson. A unifying framework for spectrum-preserving graph sparsification and coarsening. *arXiv preprint arXiv:1902.09702*, 2019.

[23] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.

[24] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[25] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[26] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Interpretable & explorable approximations of black box models. *CoRR*, abs/1707.01154, 2017.

[27] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD*, pages 177–187, 2005.

[28] Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*, 2016.

[29] Ruoyu Li, Sheng Wang, Feiyun Zhu, and Junzhou Huang. Adaptive graph convolutional neural networks. In *AAAI*, 2018.

[30] Chundi Liu, Guangwei Yu, Maksims Volkovs, Cheng Chang, Himanshu Rai, Junwei Ma, and Satya Krishna Gorti. Guided similarity separation for image retrieval. In *NeurIPS*, pages 1554–1564, 2019.

[31] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *NIPS*, 2017.

[32] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NIPS*, pages 4765–4774, 2017.

[33] Dongsheng Luo, Yuchen Bian, Yaowei Yan, Xiao Liu, Jun Huan, and Xiang Zhang. Local community detection in multiple networks. In *KDD*, 2020.

[34] Jianxin Ma, Peng Cui, Kun Kuang, Xin Wang, and Wenwu Zhu. Disentangled graph convolutional networks. In *ICML*, pages 4212–4221, 2019.

[35] Jiaqi Ma, Weijing Tang, Ji Zhu, and Qiaozhu Mei. A flexible generative framework for graph-based semi-supervised learning. In *NeurIPS*, pages 3276–3285, 2019.

[36] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

[37] Mark Newman. *Networks*. Oxford university press, 2018.

[38] Phillip E Pope, Soheil Kolouri, Mohammad Rostami, Charles E Martin, and Heiko Hoffmann. Explainability methods for graph convolutional neural networks. In *CVPR*, 2019.

[39] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *KDD*, pages 1135–1144, 2016.

[40] Kaspar Riesen and Horst Bunke. Iam graph database repository for graph based pattern recognition and machine learning. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 287–297. Springer, 2008.

[41] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE TNN*, 20(1):61–80, 2008.

[42] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine*, 30(3):83–98, 2013.

[43] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *ICML*, 2017.

[44] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 2013.

[45] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[46] Chi-Jen Wang, Seokjoo Chae, Leonid A Bunimovich, and Benjamin Z Webb. Uncovering hierarchical structure in social networks using isospectral reductions. *arXiv preprint arXiv:1801.03385*, 2017.

[47] Dongkuan Xu, Wei Cheng, Dongsheng Luo, Yameng Gu, Xiao Liu, Jingchao Ni, Bo Zong, Haifeng Chen, and Xiang Zhang. Adaptive neural network for node classification in dynamic networks. In *ICDM*, 2019.

[48] Dongkuan Xu, Wei Cheng, Dongsheng Luo, Xiao Liu, and Xiang Zhang. Spatio-temporal attentive rnn for node classification in temporal attributed graphs. In *IJCAI*, 2019.

[49] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[50] Chengliang Yang, Anand Rangarajan, and Sanjay Ranka. Global model interpretation via recursive partitioning. In *HPCC*, pages 1563–1570. IEEE, 2018.

[51] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In *AAAI*, volume 33, pages 7370–7377, 2019.

[52] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *KDD*, pages 974–983, 2018.

[53] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. In *NeurIPS*, pages 9240–9251, 2019.

[54] Jiaxuan You, Rex Ying, Xiang Ren, William L Hamilton, and Jure Leskovec. Graphrnn: Generating realistic graphs with deep auto-regressive models. In *ICML*, 2018.

[55] Hao Yuan, Jiliang Tang, Xia Hu, and Shuiwang Ji. Xgnn: Towards model-level explanations of graph neural networks. In *KDD*, 2020.

[56] Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. In *NeurIPS*, pages 5165–5175, 2018.

Supplementary Material: Parameterized Explainer for Graph Neural Network

A. Explanation algorithms

The algorithms of PGExplainer for node and graph classification are shown in Algorithm 1 and 2, respectively. We first discuss the node classification in Algorithm 1. In GNNs with message passing

---

Algorithm 1: Training algorithm for explaining node classification

---

1: **Input:** The input graph $G_o = (\mathcal{V}, \mathcal{E})$, node features $\mathbf{X}$, node labels Y, the set of instances to be explained $\mathcal{I}$, a trained GNN model: $\text{GNNE}_{\Phi_0}(\cdot)$ and $\text{GNNC}_{\Phi_1}(\cdot)$.
2: **for** each node $i \in \mathcal{I}$ **do**
3:      $G_o^{(i)} \leftarrow$ extract the computation graph for node i.
4:      $\mathbf{Z}^{(i)} \leftarrow \text{GNNE}_{\Phi_0}(G_o^{(i)}, \mathbf{X})$.
5:      $Y_o^{(i)} \leftarrow \text{GNNC}_{\Phi_1}(\mathbf{Z}^{(i)})$.
6: **end for**
7: **for** each epoch **do**
8:      **for** each node $i \in \mathcal{I}$ **do**
9:          $\Omega \leftarrow$ latent variables calculated with Eq. (10)
10:          **for** $k \leftarrow 1$ **to** K **do**
11:              $\hat{G}_s^{(i,k)} \leftarrow$ sampled from Eq. (4).
12:              $\hat{Y}_s^{(i,k)} \leftarrow \text{GNNC}_{\Phi_1}(\text{GNNE}_{\Phi_0}(\hat{G}_s^{(i,k)}, \mathbf{X}))$
13:          **end for**
14:      **end for**
15:      Compute loss with Eq. (9).
16:      Update parameters Ψ with backpropagation.
17: **end for**

---

mechanisms, the prediction at a node v is fully determined by its local computation graph, which is defined by its L-hop neighborhoods [53]. L is the number of GNN layers. Thus, for each node i in the instance set $\mathcal{I}$ to be explained, we first extract a local computation graph $G_o^{(i)}$ (line 3). With $G_o^{(i)}$ as the input graph, the trained GNN model generates the label of node i, denoted by $Y_o^{(i)}$ (line 4-5). To train the explanation network, each time we select a node i and compute parameters Ω in edge distributions with Eq. (10) (line 9). After that, we sample K graphs as input graphs for GNN to get updated predictions for node i, with the k-th prediction denoted by $\hat{Y}_s^{(i,k)}$ (line 11-13). We compute the loss and update parameters Ψ in the explanation network in line 15-16.

Algorithm 2: Training algorithm for explaining graph classification

---

1: **Input:** A set of input graphs with i-th graph represented by $G_o^{(i)}$, node features $\mathbf{X}^{(i)}$, and a label $Y^{(i)}$, a trained GNN model: $\text{GNNE}_{\Phi_0}(\cdot)$ and $\text{GNNC}_{\Phi_1}(\cdot)$.
2: **for** each graph $G_o^{(i)}$ **do**
3:      $\mathbf{Z}^{(i)} \leftarrow \text{GNNE}_{\Phi_0}(G_o^{(i)}, \mathbf{X}^{(i)})$.
4:      $Y_o^{(i)} \leftarrow \text{GNNC}_{\Phi_1}(\mathbf{Z}^{(i)})$.
5: **end for**
6: **for** each epoch **do**
7:      **for** each graph $G_o^{(i)}$ **do**
8:          $\Omega \leftarrow$ latent variables calculated with Eq. (11)
9:          **for** $k \leftarrow 1$ **to** K **do**
10:              $\hat{G}_s^{(i,k)} \leftarrow$ sampled from Eq. (4).
11:              $\hat{Y}_s^{(i,k)} \leftarrow \text{GNNC}_{\Phi_1}(\text{GNNE}_{\Phi_0}(\hat{G}_s^{(i,k)}, \mathbf{X}^{(i)}))$
12:          **end for**
13:      **end for**
14:      Compute loss with Eq. (9).
15:      Update parameters Ψ with backpropagation.
16: **end for**

The training algorithm for explaining graph classification is shown in Algorithm 2. The algorithm is similar to the one explaining node classifications, except that computation graphs are not used, because, for graph classification, each graph is treated as an instance. Given a set of graphs $\{G_o^{(i)}\}_{i \in \mathcal{I}}$, we first compute the node embeddings $\mathbf{Z}^{(i)}$ and graph labels $Y_o^{(i)}$ with the trained GNN model (line 2-4). In each epoch, for each i-th graph, we compute the parameters Ω in its edge distributions with Eq. 11 (line 8). We then sample K subgraphs and get the updated predictions. We compute the loss with Eq. (9) and update parameters Ψ with backpropagation.

B. Hardware and implementations in experiments

All experiments are conducted on a Linux machine with an Nvidia GeForce RTX 2070 SUPER GPU with 8GB memory. CUDA version is 10.2 and Driver Version is 440.64.00. PGExplainer is implemented with Tensorflow 2.0.0. For each dataset, we first train a GNN model, which is then shared by all posthoc methods ATT, GNNExplainer, and PGExplainer. We use FC(a, b, f) to denote a fully-connected layer. a and b are the numbers of input and output neurons respectively. f is the activation function. Similarly, we denote a GNN layer with input dimension a, output dimension b, and activation function f by GNN(a, b, f). With these notations, the network structure of the GNN model for node classification is GNN(10, 20, ReLU)-GNN(20, 20, ReLU)-GNN(20, 20, ReLU)-FC(20, #label, softmax). For graph classification, we add a maxpooling layer to get graph representations before the final FC layer. Thus, the network structure is GNN(10, 20, ReLU)-GNN(20, 20, ReLU)-GNN(20, 20, ReLU)-Maxpooling-FC(20, #label, softmax). We adopt the Adam optimizer with the initial learning rate of 1.0×10^{-3}. All variables are initialized with Xavier. We follow GNNExplainer to split train/validation/test with 80/10/10% for all datasets. Each model is trained for 1000 epochs. The accuracy performances of GNN models are shown in Table 3. The results show that the designed GNN models are powerful enough for node/graph classifications on both synthetic and real-life datasets.

Table 3: Accuracy performance of GNN models

	Node Classification				Graph Classification	
Accuracy	BA-Shapes	BA-Community	Tree-Cycles	Tree-Grid	BA-2motifs	MUTAG
Training	0.98	0.99	0.99	0.92	1.00	0.87
Validation	1.00	0.88	1.00	0.94	1.00	0.89
Testing	0.97	0.93	0.99	0.94	1.00	0.87

The network structure of explanation networks in PGExplainer is FC(#input, 64, ReLU)-FC(20, 1, Linear), which is shared for all datasets. #input is 60 for node classification, and 40 for graph classification.To train PGExplainer, we also adopt the Adam optimizer with the initial learning rate of 3.0×10^{-3}. The coefficient of size regularization is set to 0.05 and entropy regularization is 1.0. The epoch T is set to 30 for all datasets. The temperature τ in Eq. (4) is set with annealing schedule [1]: $\tau^{(t)} = \tau_0 (\tau_T / \tau_0)^t$, where τ_0 and τ_T are the initial and final temperatures. A small temperature tends to generate more discrete graphs which may hinder the explanation network being optimized with backpropagation. In this task, we find that relatively high temperatures work well in practice. τ_0 is set to 5.0 and τ_T is set to 2.0.

C. Additional experiments

In this part, we conduct extensive experiments to have deep insights into our PGExplainer.

C.1 Inductive performance

As we discussed in Section 4.3, the explanation network is shared across the population. Thus, with a trained PGExplainer, we can directly infer the explanation without retraining the explanation network. As a result, our PGExplainer has better generalization power than the leading baseline GNNExplainer. Besides, our PGExplainer is more efficient in the inductive setting. In this section, we empirically demonstrate the effectiveness of PGExplainer in the inductive setting. In the inductive setting, we select α instances for training, $(N - \alpha)/2$ for validation, and the rest for testing. α is ranged from $[1, 2, 3, 4, 5, 30]$. Note that, with $\alpha = 1$, our method degenerates to the single-instance explanation

method. Recall that to explain a set of instances, GNNExplainer first detects a reference node and then computes the explanation for the reference node. The explanation is then generalized to other nodes with graph alignment [53]. We claim that it may lead to sub-optimal explanations because reference node selection and graph alignment are not jointly optimized with the explanation in an end-to-end fashion. The AUC scores of PGExplainer are shown in Figure 3. We have the following observations. 1) The testing AUC increase as more instances are trained, verifying the effectiveness of PGExplainer. Some results are higher than the reported ones in Section 5 because here we adopt validation datasets to fine-tune the hyper-parameters. 2) More training instances lead to smaller standard deviation and PGExplainer tends to globally detect shared motifs with higher robustness. 3) PGExplainer can achieve relatively good performance with a small number of trained instances, which makes PGExplainer more practical in large datasets. The results also explain why we dismiss the training time of PGExplainer and only count the inference time in Section 5.
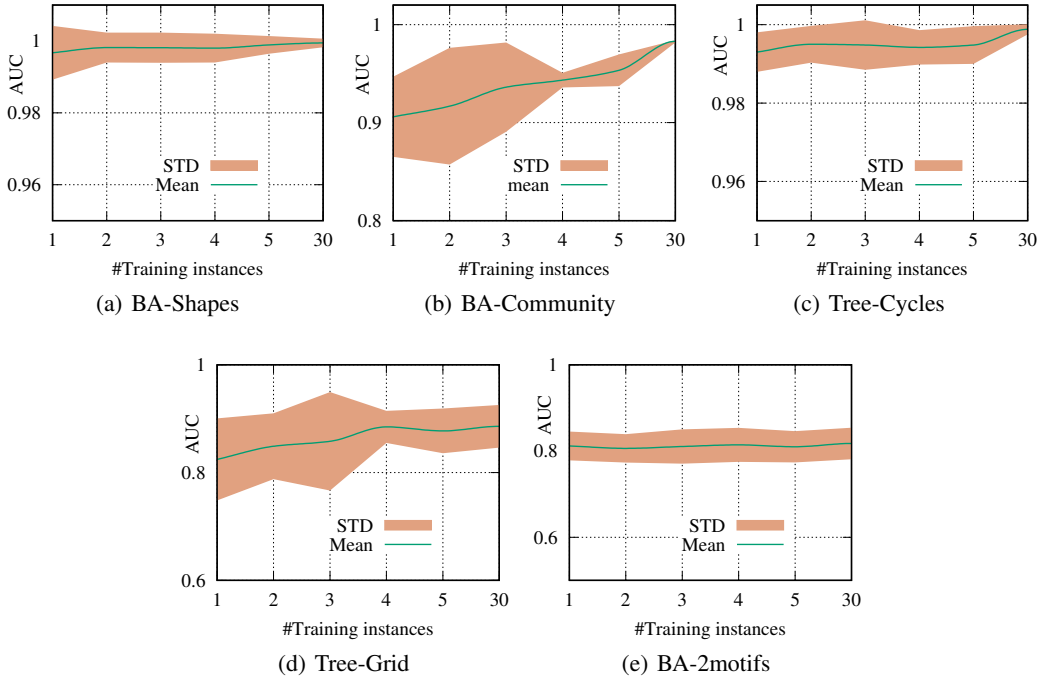

Figure 3: Evaluation of PGExplainer in the inductive setting.

C.2 Effects of regularization terms

In this part, we analyze the effects of regularization terms. In addition to the size and entropy regularizers introduced in GNNExplainer, we also have discussed regularization terms on budgets and connectivity constraints. Since the first two regularizers are used in the quantitative evaluation, we first conduct parameter studies. Visualization results on synthetic datasets show that the explanatory graph extract by PGExplainer tends to be small and compact. To verify the effectiveness of the proposed regularizer for connectivity constraint, we synthesize a noisy BA-Shapes dataset.

Effects of size and entropy constraint. We select synthetic datasets for parameter study. The coefficients of size and entropy regularizers are denoted by λ_s and λ_e, respectively. AUC scores w.r.t coefficients are shown in Figure 4. We observe that PGExplainer achieves competitive performances even without any regularization terms in all datasets except the BA-Community, which verifies the effectiveness of the model itself. For the BA-Community dataset, the entropy constraint plays an important role.

Effects of connectivity constraint. To show the effect of the connectivity constraint on the explanatory graph, we build a noisy BA-Shapes dataset with $0.2N$ noisy edges. We vary the coefficient of the connectivity regularization term λ_c from 0 to 10 and apply PGExplainer to explain a single instance.

(a) BA-Shapes

(b) BA-Community

(c) Tree-Cycles

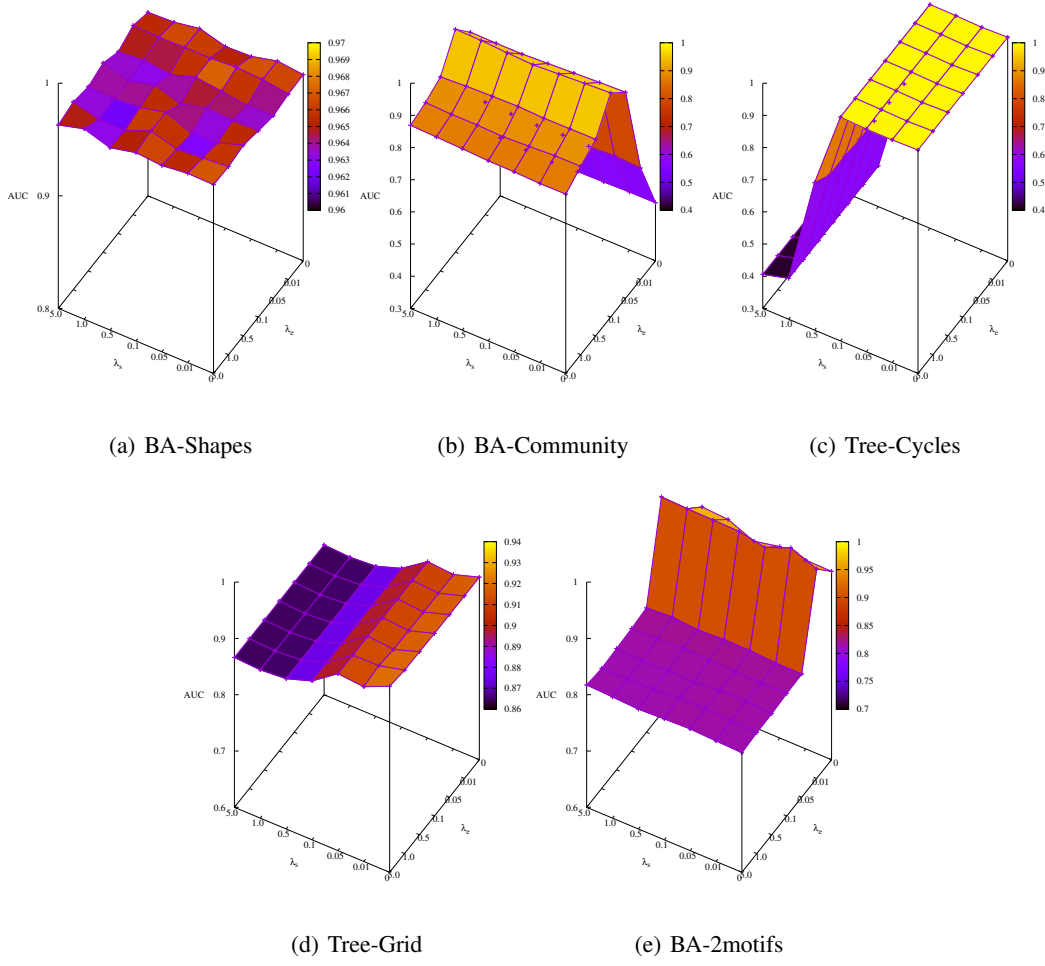(d) Tree-Grid

(e) BA-2motifs

Figure 4: Effects of size and entropy constraint

The visualization results with regard to different choices of coefficients are shown in Figure 5. The figure demonstrates that without explicit constraint, PGExplainer may detect several connected edges in the noisy input graph, although these edges are also inside motifs. With the connectivity constraint, we observe that PGExplainer tends to provide a connected subgraph as an explanation.
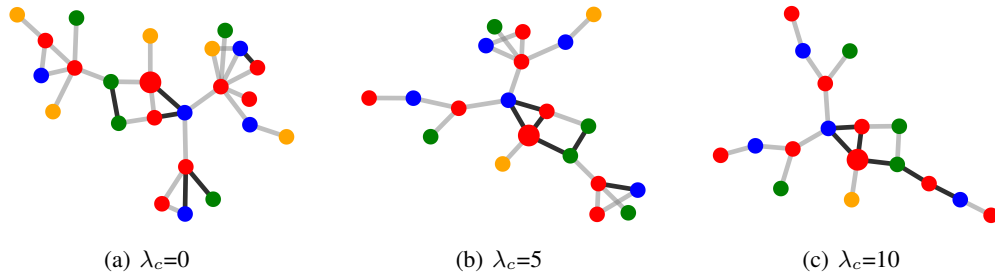

(a) λ_c=0

(b) λ_c=5

(c) λ_c=10

Figure 5: Effects of connectivity constraint

D. Selection of subset of features

In this paper, we focus on globally understanding predictions made by GNNs by providing topological explanations. To explain node features, in GNNExplainer, the authors propose to use a feature mask to select features that are important to preserve original predictions. Feature selection has been extensively studied in non-graph neural networks and can be applied directly to explain GNNs, such as the concrete autoencoder [1]. Besides, since the selected features are shared among instances across the population, feature explanation is naturally global and applicable to new instances in the inductive setting.