Factorized Explainer for Graph Neural Networks

Rundong Huang[1], Farhad Shirani[2*], Dongsheng Luo[2*]

[1] Technical University of Munich, [2] Florida International University

Objective

- **Post-hoc Instance-Level Explainability** Given model $f(\cdot)$ and input graph $\mathbf{G}$, find minimal and sufficient subgraph $\mathbf{G}_s = \Psi(\mathbf{G})$ w.r.t $f(\cdot)$.
- **Research Question 1:** How to quantify Minimality and Sufficiency?
- **Research Question 2:** How to design explainer mechanisms to optimize the resulting objective?
- **Research Question 3:** How to overcome the locality of the computation graph of GNNs to design explainers for multi-motif scenarios?

Graph Information Bottleneck

- **Graph Information Bottleneck (GIB):** Tradeoff between Minimality and Sufficiency

$$\Psi(\cdot) \triangleq \arg\min_{\Psi: G \mapsto G_s} I(G; G_s) - \alpha I(Y; G_s),$$

- Used to train various explainers such as: PGExaplainer, GSAT, and MixupExplainer.
- $I(G; G_s)$ and $I(Y; G_s)$ difficult to estimate, consequently, surrogates used in practice.

Shortcomings of GIB

- **Minimality Quantification:** $I(G; G_s)$ allows leakage of low-entropy components into explanation.
- **Sufficiency Quantification:** $I(Y; G_s)$ leads to a signaling issue.
- **Theorem 1:** For a graph classification task parametrized by $P_{G,Y}$, assume that there exists a mapping $h : \mathcal{G} \to \mathcal{Y}$ such that $G \leftrightarrow h(G) \leftrightarrow Y$ holds. Then, for any $\alpha > 0$, there exists an explanation algorithm $\Psi_\alpha(\cdot)$ such that $G' \triangleq \Psi_\alpha(G)$ optimizes the objective function in GIB and $\Psi_\alpha(G) \leftrightarrow h(G) \leftrightarrow G$ holds.

Modified Objective Function

- **Modified Objective Function:**

$$\Psi(\cdot) = \arg\min_{\Psi: G \mapsto G_s} \max(\mathbb{E}_G(|G_s|), \beta) + \alpha \mathbb{E}_G(CE(Y; f(G_s))),$$

- G_s is OOD with respect to P_G.
- Example: In MUTAG, inputs have tens of vertices, however, explanation subgraphs have few vertices.
- The OOD issue is addressed in recent follow-up works.

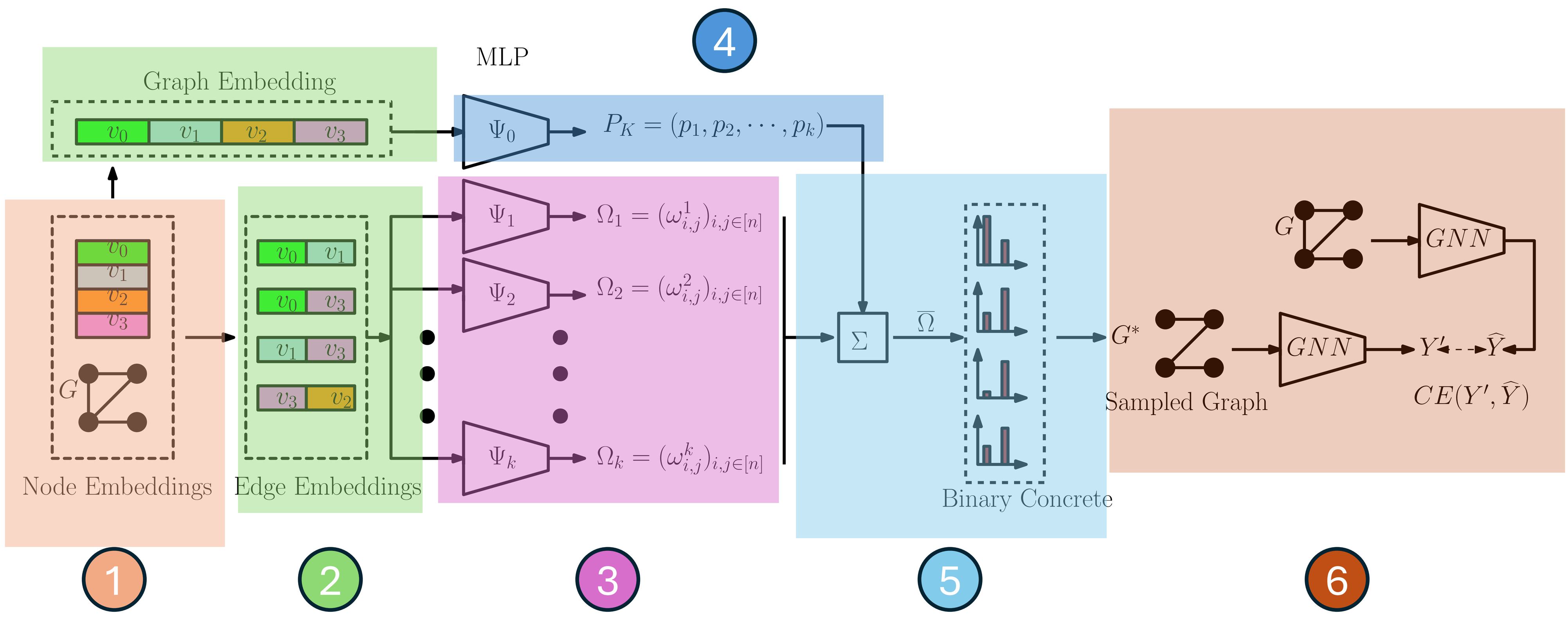
Locality of Explanation Method

- **Local Explanation Methods:** Consider a graph classification task $P_{G,Y}$, classifier $f : \mathcal{G} \to \mathcal{Y}$, a parameter $r \in \mathbb{N}$, and an explanation function $\Psi : \mathcal{G} \to \mathcal{G}$, where $\mathcal{G}$ is the set of all possible input graphs, and $\mathcal{Y}$ is the set of output labels. Let $G' = \Psi(G) = (\mathcal{V}', \mathcal{E}'; \mathbf{Z}', \mathbf{A}')$. The explanation function $\Psi(\cdot)$ is called an r-local explanation function if:
  1. The Markov chain $\mathbb{1}(v \in \mathcal{V}') \leftrightarrow G_{v,r} \leftrightarrow G$ holds for all $v \in \mathcal{V}$, where $\mathbb{1}(\cdot)$ is the indicator function.
  2. The edge (v, v') is in $\mathcal{E}'$ if and only if $v, v' \in \mathcal{V}'$ and $e \in \mathcal{E}$.
- Explanation methods which rely on GNN node embeddings, such as PGExplainer are local.

Suboptimality of Local Methods

- **Theorem 2:** Let $r, r' \in \mathbb{N}$. There exist classification problems for which:
  a) The optimal Bayes classification rule $f^*(g)$ is $\mathbb{1}(\exists i \in [s] : g_i \subseteq g)$.
  b) For any r-local explanation function, there exists $\alpha' > 0$ such that the explanation is suboptimal for f^* in the modified GIB sense for all $\alpha > \alpha'$ and β equal to maximum number of edges of $g_i, i \in [s]$.
  c) There exists $k \leq s$, a parameter $\alpha' > 0$, a collection of r'-local explanation functions $\Psi_i(\cdot), i \in [k]$, and an explanation function Ψ^*, such that for all inputs g, we have $\Psi(g) \in \{\Psi_1(g), \Psi_2(g), \cdots, \Psi_k(g)\}$ and Ψ^* is optimal in the modified GIB sense for all $\alpha > \alpha'$ and γ equal to maximum number of edges of $g_i, i \in [s]$.
- The theorem suggest that we can 'patch' together local explainers to construct optimal explanation functions.

K-FactExplainer



Step 1: The original GNN produces node embeddings. **Step 2:** Graph/edge embeddings produced by concatenation. **Step 3:** MLP sequence produce edge probabilities. **Step 4:** Ψ_0 MLP produces weights for each prediction. **Step 5:** Compute weighted average of predictions. **Step 6:** Loss calculated by comparing GNN output for subgraph

Experimental Results

	BA-Shapes	BA-Community	Tree-Circles	Tree-Grid	BA-2motifs	MUTAG
GRAD	0.882	0.750	0.905	0.667	0.717	0.783
ATT	0.815	0.739	0.824	0.612	0.674	0.765
RGExp.	$0.985_{\pm 0.013}$	$0.919_{\pm 0.017}$	$0.787_{\pm 0.099}$	$\mathbf{0.927}_{\pm 0.032}$	$0.657_{\pm 0.107}$	$0.873_{\pm 0.028}$
DEGREE	$0.993_{\pm 0.005}$	$0.957_{\pm 0.010}$	$\mathbf{0.902}_{\pm 0.022}$	$0.925_{\pm 0.040}$	$0.755_{\pm 0.135}$	$0.773_{\pm 0.029}$
GNNExp.	$0.742_{\pm 0.006}$	$0.708_{\pm 0.004}$	$0.540_{\pm 0.017}$	$0.714_{\pm 0.002}$	$0.499_{\pm 0.004}$	$0.606_{\pm 0.003}$
PGExp.	$0.999_{\pm 0.000}$	$0.825_{\pm 0.040}$	$0.760_{\pm 0.014}$	$0.679_{\pm 0.008}$	$0.566_{\pm 0.004}$	$0.843_{\pm 0.162}$
K-FactExplainer	$\mathbf{1.000}_{\pm 0.000}$	$\mathbf{0.974}_{\pm 0.004}$	$0.779_{\pm 0.004}$	$0.770_{\pm 0.004}$	$\mathbf{0.821}_{\pm 0.005}$	$\mathbf{0.915}_{\pm 0.010}$

Explanation faithfulness in terms of AUC-ROC on edges under six datasets.

Acknowledgements

This project was partially supported by NSF grants IIS-2331908 and CCF-2241057. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.