



**ICLR**

Twelfth International Conference  
on Learning Representations

# ***GOAt: Explaining GNNs via Graph Output Attribution***

---

Shengyao Lu, Keith G. Mills, Jiao He, Bang Liu, Di Niu  
Presenter: shengyao@ualberta.ca

# Background: Instance-level GNN Explainability

**Instance-level GNN Explainability:** it focuses on identifying important nodes, edges or subgraphs behind a GNN model's specific predictions, these explanations are generated per individual data instance.

**Problem Formulation:** Our goal is to generate a faithful explanation for each graph instance  $G = (\mathcal{V}, \mathcal{E})$  by identifying a subset of edges  $\mathcal{S} \subseteq \mathcal{E}$ , which are important to the predictions, given a GNN  $f(\cdot)$  pretrained on a set of graphs  $\mathcal{G}$ .

Most existing methods train an auxiliary model to explain GNNs, which causes the “Explain a black box with another black box” problem, making the explanations less transparent or reliable.

Our approach GOAt has the following advantages:

- *Transparent and Faithful:* It forwardly determines the attribution of each edge at the output. It avoids the training of auxiliary models and directly computes the attribution of each edge to the GNN prediction, which allows GOAt to be more faithful to the GNN itself, hence have better discriminative capability and stability across the same-class samples.
- *Handling discrete inputs:* The design nature of GOAt offers to handle the discrete inputs more effectively, compared with methods like Integrated Gradients that always consider the inputs as continuous. For example, the elements in the adjacency matrix can be  $\{0, 1\}$  indicating the absence or presence of edges between pairs of nodes. Any values between  $\{0,1\}$  are not relevant to the problem, as they do not carry any accumulating meaning.
- *Statistically convincing:* GOAt passes the sanity check that the attribution scores for input features add up to the difference in the GNN's output with and without those features, whereas most of the search-based approaches or learning-based approaches cannot.

# Toy Example of GOAt

MOM



IF YOU FLIP THE COIN THREE TIMES. AND  
RECORD THE RESULTS AS X, Y, Z. I WILL  
GIVE YOU  $10 * X * Y * Z$  BUCKS.

1      0  
HEAD    TAIL  
①      ②

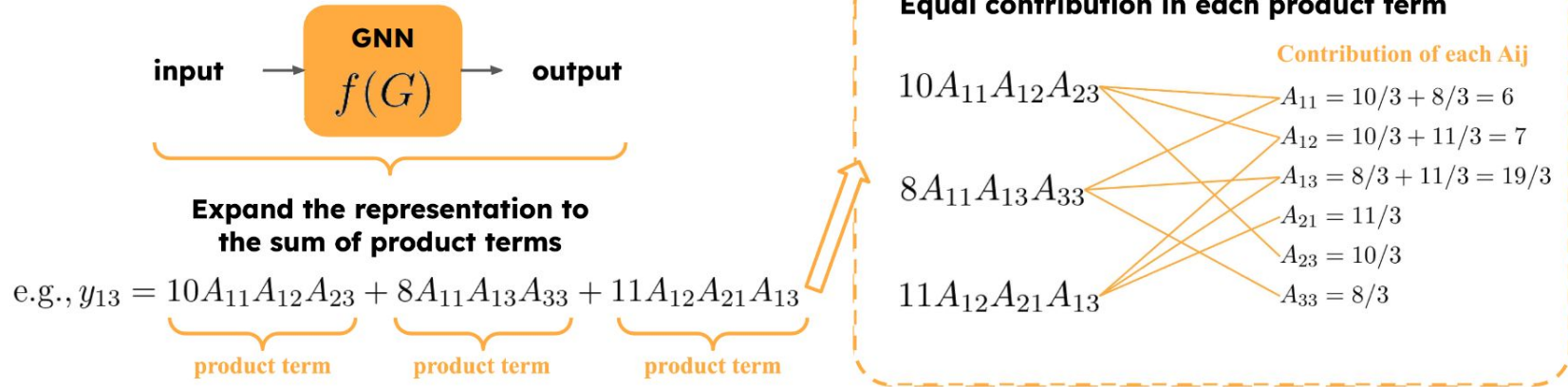
ILLUMI

OKAY.



There are only two possible outcomes: **\$10** or **\$0**.  
Any of X, Y, Z being 0 will result in **\$0**. Therefore,  
“X=1”, “Y=1”, “Z=1” is equally important to the  
outcome “**\$10**”.

# High-level Idea of GOAt



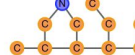


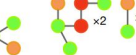
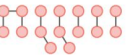
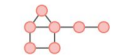




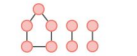
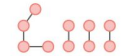






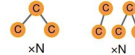
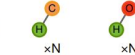




**Expansion form of the output representation:** The output matrix of a GNN can be described as the outcome of a linear transformation involving the input matrices ( $A, X$ ) and the GNN parameters ( $W, B$ ). As a result, each element within the output matrix can be represented as the sum of scalar products that involve entries from both the input matrices and the GNN parameters. Each edge appears in only some of the scalar products. Consequently, we can determine the attribution of an edge, such as  $A_{11}$ , by summing its contribution across all the scalar products in which it participates.

**Equal Contribution:** Consider a scalar product term  $z = 10A_{11}A_{12}A_{23}$ .  $z=10$  only when  $A_{11} = A_{12} = A_{23} = 1$ , otherwise  $z=0$ . This implies that the presence of all edges is equally essential for the resulting value of  $z = 10$ . Therefore, each of the three edges contributes  $1/3$  to the output, resulting in an attribution of  $10/3$ .

# Qualitative Results

Table 1: Qualitative results of the top motifs of each class produced by PGExplainer, SubgraphX, RCExplainer and *GOAt*. The percentages indicate the coverage of the explanations.

	BA-2Motifs		Mutagenicity		NCI1	
	Class0	Class1	Class0	Class1	Class0	Class1
PGExplainer	 4.8%	 1.8%	 1.2%	 1.3%	 0.1%	 0.5%
SubgraphX	 0.4%	 12.8%	 0.2%	 0.2%	 0.2%	 0.1%
RCExplainer	 6.4%	 6.2%	 0.4%	 0.5%	 0.05%	 0.1%
<i>GOAt</i>	 3.8%	 3.4%	 3.5%	 2.2%	 3.5%	 4.0%

Based on the explanations from *GOAt*, we have observed that the GNN effectively recognizes the "house" motif that is associated with Class 1. In contrast, other approaches face difficulties in consistently capturing this motif. The Class 0 motifs in the Mutagenicity dataset generated by *GOAt* represent multiple connected carbon rings. This indicates that the presence of more carbon rings in a molecule increases its likelihood of being mutagenic (Class 0), while the presence of more "C-H" or "O-H" bonds in a molecule increases its likelihood of being non-mutagenic (Class 1). Similarly, in the NCI1 dataset, *GOAt* discovers that the GNN considers a higher number of carbon rings as evidenced of chemical compounds being active against non-small cell lung cancer. Other approaches, on the other hand, fail to provide clear and human-understandable explanations.

# Experimental Results: (Stability)

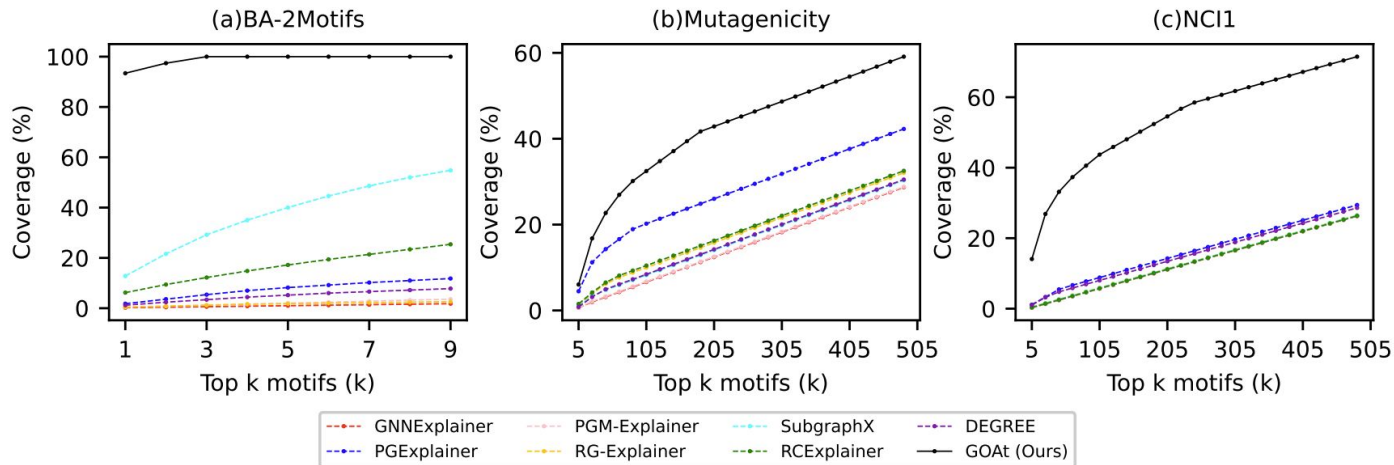


Figure 4: Coverage of the top- $k$  explanations across the datasets.

To quantify the relative consistency for data samples with similar properties, we introduce the stability metric, which measures the coverage of the top- $k$  explanations across the dataset. An ideal explainer should generate explanations that cover a larger number of data samples using fewer motifs. Our approach surpasses the baselines by a considerable margin in terms of the stability of producing explanations. Specifically, GOAt is capable of providing explanations for all the Class 1 data samples using only three explanations. This explains why there are only three Class 1 scatters visible the scatter plot of GOAt on BA-2Motifs.

# Experimental Results: (Discriminability)

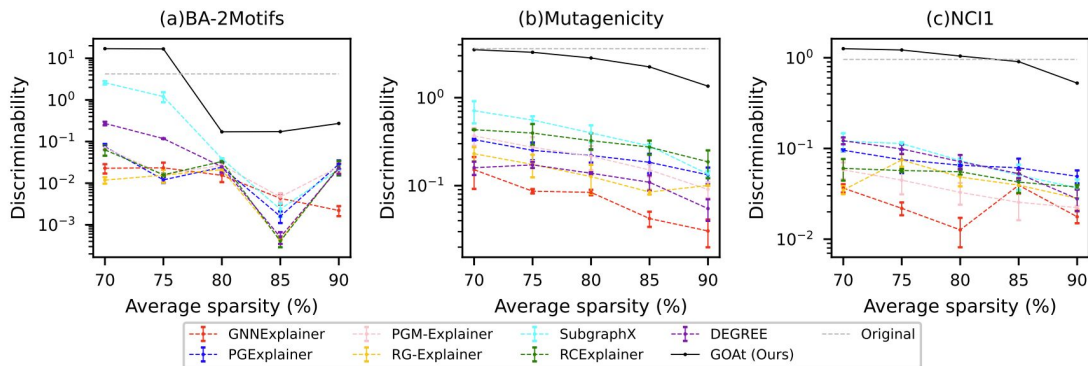


Figure 2: Discriminability performance averaged across 10 runs of the explanations produced by various GNN explainers at different levels of sparsity. "Original" refer to the performance of feeding the original data into the GNN without any modifications or explanations applied.

Discriminability refers to the ability of the explanations to distinguish between the classes. We define the discriminability between two classes  $c_1$  and  $c_2$  as the L2 norm of the difference between the mean values of explanation embeddings of the two classes. The embeddings used for explanations are taken prior to the last-layer classifier. In this procedure, only the explanation subgraph  $S$  is fed into the GNN instead of the whole graph  $G$ .

Due to the significant performance gap between the baselines and GOAt, a logarithmic scale is employed. Our approach consistently outperforms the baselines in terms of discriminability across all sparsity levels, demonstrating its superior ability to generate accurate and reliable class-specific explanations. Notably, at sparsity = 0.7, GOAt achieves higher discriminability than the original graphs on the BA-2Motifs and NCI1 datasets. This indicates that GOAt effectively reduces noise unrelated to the investigated class while producing informative class explanations.

# Experimental Results: (Discriminability)



Figure 3: Visualization of explanation embeddings on the BA-2Motifs dataset. Subfigure (i) refers to the visualization of the original embeddings by directly feeding the original data into the GNN without any modifications or explanations applied.

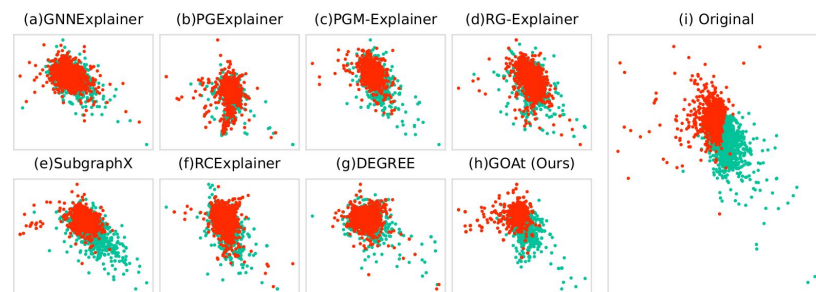


Figure I.3: Visualization of explanation embeddings on the Mutagenicity dataset. Subfigure (i) refers to the visualization of the original embeddings by directly feeding the original data into the GNN without any modifications or explanations applied.

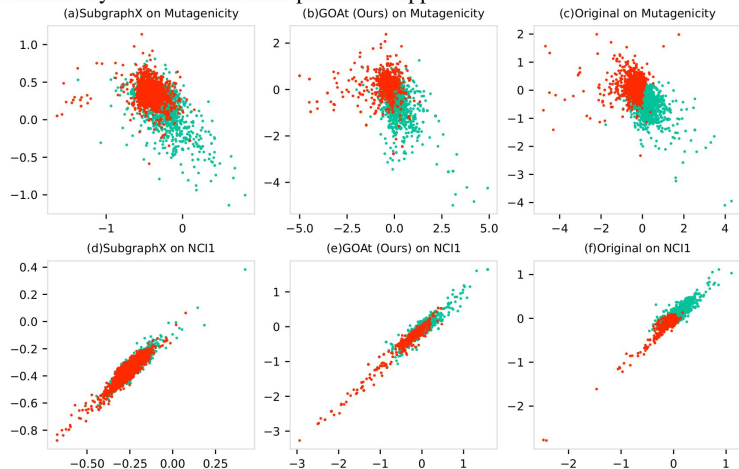


Figure I.5: Visualization of explanation embeddings on the Mutagenicity and NCI1 datasets with axes turned on.

The explanations generated by GNNExplainer fail to exhibit class discrimination, as all the data points are clustered together without any distinct separation. While some of the Class 1 explanations produced by PGExplainer, PGM-Explainer, RG-Explainer, RCExplainer, SubgraphX and DEGREE are separate from the Class 0 explanations, the majority of the data points remain closely clustered.

In contrast, GOAt provides more discriminative explanations, which exhibit greater dispersion in the scatter plot. Furthermore, compared to the original embeddings, the explanations generated from GOAt demonstrate higher confidence towards specific classes, as evidenced by the scale of the scatter plots.



# Experimental Results: (Fidelity)

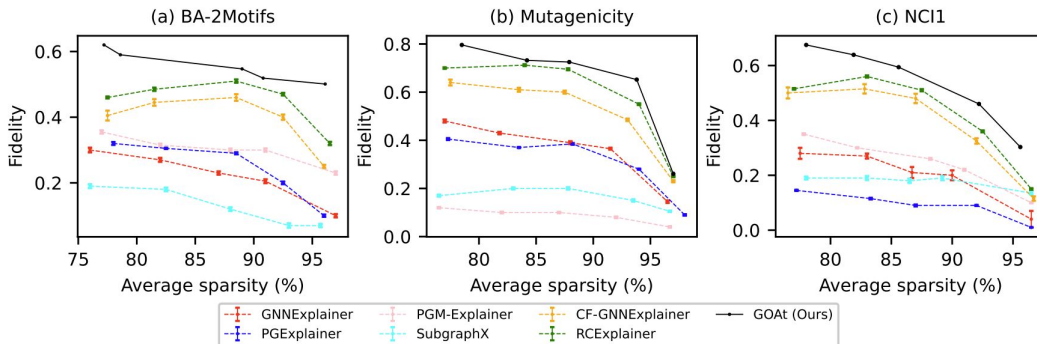


Figure 1: Fidelity performance averaged across 10 runs on the pretrained GCNs for the datasets at different levels of average sparsity.

Recall that fidelity is the decrease of predicted probability between original and new predictions after removing important edges, which are used to evaluate the faithfulness of explanations. Fidelity is compared at different sparsity levels, where sparsity is the percentage of edges that remain in  $G$  after the removal of the explanation edges.

$$fidelity(S, G) = f_y(G) - f_y(G \setminus S)$$

$$sparsity(S, G) = 1 - \frac{|S|}{|E|}.$$

Our proposed approach, GOAt, achieve the state-of-the-art fidelity performance across all sparsity levels, validating its superior performance in generating accurate and reliable faithful explanations. Among the other methods, RCExplainer exhibits the highest fidelity, as it is specifically designed for fidelity optimization. Notably, unlike the other methods that require training and hyperparameter tuning, GOAt offers the advantage of being a training-free approach, thereby avoiding any errors across different runs.

# Thanks!

Presenter: Shengyao Lu [shengyao@ualberta.ca](mailto:shengyao@ualberta.ca)

Code: <https://github.com/sluxsr/GOAt>



**UNIVERSITY  
OF ALBERTA**