GXAI-Bench: Evaluating Explainers for Graph Machine Learning Algorithms

*Owen Queen[1], *Chirag Agarwal[2], Himabindu Lakkaraju[3], and Marinka Zitnik[4]
[1]University of Tennessee Knoxville, [2]Adobe Research, [3]Harvard University, [4]Harvard Medical School

Motivation

With a recent increase in explanation (XAI) methods for graph machine learning (ML) models, it becomes crucial to understand the quality of explanations. However, the evaluation of such XAI methods is often based on weak reasoning and poorly-designed synthetic datasets. GXAI-Bench introduces a rigorous pipeline for graph XAI evaluation, including metrics and a novel synthetic dataset that is robust and scalable. We present benchmarking results and introduce an open-source software package.

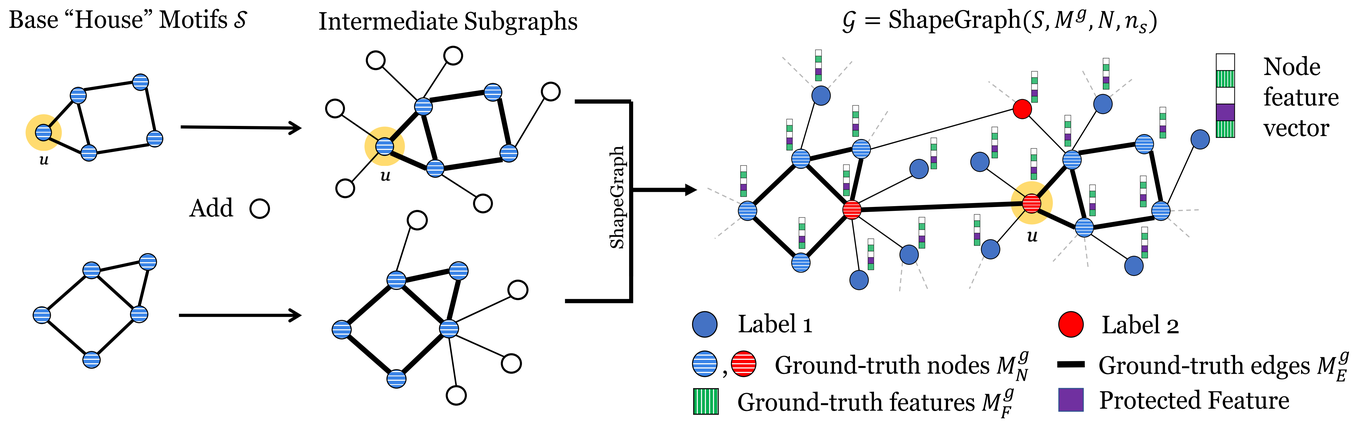

Figure 1: An enclosing subgraph within a ShapeGraph dataset. Masks are provided over important edges, nodes, and features. Labels are assigned based on number of house motifs within the 1-hop neighborhood.

Desiderata for Explainers

We adopt a desiderata for explainers from Agarwal et al., 2022[a]. Metrics and definitions is shown in Table 1.

Metric	Definition
Accuracy (GEA)	Correctness of explanation
Faithfulness (GEF)	Adherence to model behavior
Stability (GES)	Sensitivity to small changes in input data
Fairness (GECF, GEGF)	Lack of bias in generated explanations

Table 1: Metrics and definitions used within the GXAI-Bench pipeline. Abbreviations of each metric are relevant for Table 2.

ShapeGraph

Faber et al., 2021[b] outlines several pitfalls of datasets that have been used to evaluate GNN explainers. To address these pitfalls, we introduce ShapeGraph, a novel dataset generator for node classification that consists of a motif counting task with distinct ground-truth explanations. ShapeGraph also supports protected features, a feature that allows for evaluation of fairness metrics.

Through parameterization, the following components of ShapeGraph datasets can be tightly controlled:

- Homophily vs. heterophily
- Size of overall graph
- Size of ground-truth explanations
- Difficulty of classification task
- Relative importance of features and structure

Results

Method	GEA (↑)	GEF (↓)	GES (↓)	GECF (↓)	GEGF (↓)
Random	0.148±0.002	0.193±0.002	0.933±0.001	0.763±0.002	0.023±0.002
Grad	0.193±0.002	**0.086**±0.002	0.804±0.004	0.175±0.004	0.039±0.002
GradCAM	0.222±0.002	0.241±0.002	0.278±0.004	**0.031**±0.003	0.021±0.002
GuidedBP	0.190±0.001	0.177±0.002	0.444±0.004	0.087±0.003	**0.020**±0.002
IG	0.139±0.002	0.227±0.002	0.730±0.005	0.129±0.004	0.022±0.002
GNNExplainer	0.102±0.003	0.237±0.002	0.437±0.008	0.241±0.006	0.027±0.002
PGMExplainer	0.130±0.002	0.168±0.001	0.849±0.002	0.684±0.003	0.091±0.004
PGExplainer	0.194±0.002	0.241±0.002	**0.212**±0.004	0.079±0.003	0.030±0.002
SubgraphX	**0.286**±0.004	0.221±0.005	0.738±0.005	0.245±0.006	0.034±0.002

Table 2: GXAI-Bench metrics on a ShapeGraph. Best-performing methods for each metric are shown in bold.

GXAI-Bench is used to evaluate state-of-the-art graph explainers. Quantitative analysis of each metric (Table 2) shows that current explainers perform poorly on our metrics, and no method is a clear "winner".

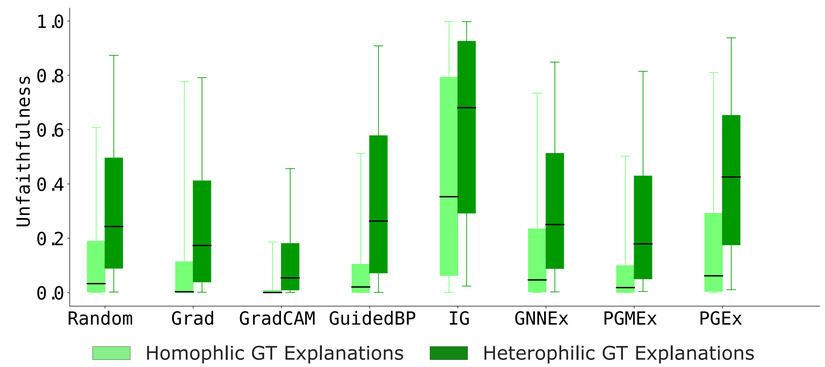

Figure 2: Evaluating faithfulness of explainers on homophilic vs. heterophilic graphs. Higher scores indicate that methods are less faithful to the predictor.

Conclusion

- GXAI-Bench establishes a rigorous methodology, including datasets and metrics, for evaluating graph XAI methods.
- We present benchmarking results on our methods.
- Our work is open-sourced for future use in the community.

[a]C. Agarwal, M. Zitnik, and H. Lakkaraju. "Probing GNN Explainers: A Rigorous Theoretical and Empirical analysis of GNN Explanation Methods." *AISTATS*, 2022.
[b]L. Faber, A. K. Moghaddam, R. Wattenhofer. "When Comparing to Ground Truth is Wrong: On Evaluating GNN Explanation Methods." *KDD* 2021.
* Please inquire for citations on explainers