Explaining Graph Neural Networks via Structure-aware Interaction Index

Ngoc Bui [1] **Hieu Trung Nguyen** [2] **Viet Anh Nguyen** [3] **Rex Ying** [1]

Abstract

The Shapley value is a prominent tool for interpreting black-box machine learning models thanks to its strong theoretical foundation. However, for models with structured inputs, such as graph neural networks, existing Shapley-based explainability approaches either focus solely on node-wise importance or neglect the graph structure when perturbing the input instance. This paper introduces the Myerson-Taylor interaction index that internalizes the graph structure into attributing the node values and the interaction values among nodes. Unlike the Shapley-based methods, the Myerson-Taylor index decomposes coalitions into components satisfying a pre-chosen connectivity criterion. We prove that the Myerson-Taylor index is the unique one that satisfies a system of five natural axioms accounting for graph structure and high-order interaction among nodes. Leveraging these properties, we propose Myerson-Taylor Structure-Aware Graph Explainer (MAGE), a novel explainer that uses the second-order Myerson-Taylor index to identify the most important motifs influencing the model prediction, both positively and negatively. Extensive experiments on various graph datasets and models demonstrate that our method consistently provides superior subgraph explanations compared to state-of-the-art methods.

1. Introduction

Graph Neural Networks (GNNs) are ubiquitous thanks to their predictive power in many applications (Zhou et al., 2020; Wu et al., 2020). GNNs proliferate in various real-world applications, from natural language processing (Wu et al., 2023b), image recognition and detection (Han et al., 2022), point cloud analysis (Shi & Rajkumar, 2020; Zhang et al., 2022b) to AI for science (Sun et al., 2020). However, understanding the rationale behind the prediction of GNNs remains challenging. As GNNs are gaining popularity in high-stake domains, their (lack of) explainability becomes a growing concern (Yuan et al., 2022; Kakkad et al., 2023). Attempts towards explaining GNNs can be categorized into two main directions: *white-* and *black-box* explanability. White-box explainers (Pope et al., 2019; Feng et al., 2022b) typically necessitate access to a model's internal structures or gradients. In contrast, black-box explainers (Ying et al., 2019; Luo et al., 2020; Vu & Thai, 2020; Schlichtkrull et al., 2020; Yuan et al., 2020) only require querying the model's output; hence they are more versatile and applicable to a broader range of architectures.

Shapley value (Shapley, 1953) is a game theory concept successfully applied to explain black-box ML models in various domains, such as tabular or image data (Lundberg & Lee, 2017). However, adopting the Shapley values to models with graph input poses a significant challenge, mainly due to the combinatorial nature of graph structures. Several works (Duval & Malliaros, 2021; Yuan et al., 2021; Ye et al., 2023) have utilized the Shapley value to determine node importance in the graph input by perturbing the graph and measuring the change in the model's prediction when specific nodes are removed or ablated. The importance scores, or attribution scores, are used to identify a subset of nodes most influential to the model prediction (Zhang et al., 2022a). However, existing approaches to leveraging Shapley values for graph data encounter several challenges.

First, the Shapley value does not consider the graph structure when perturbing the graph input. This can lead to perturbed graphs that may be disconnected or pathological, which the GNN model does not observe during the training. Assessing the model on these pathological graphs may inject bias, adversely affecting the estimated attribution scores. Therefore, Shapley's attribution may not reflect the true importance of nodes (Zhang et al., 2022a).

Second, most existing methods focus on attributing importance scores for nodes or edges individually (Zhang et al., 2022a; Duval & Malliaros, 2021). They then apply a greedy algorithm to highlight a group of nodes/edges with the highest total node-wise importance. However, this approach is not suitable for applications requiring the identification of

[1]Yale University [2]VinAI Research [3]The Chinese University of Hong Kong. Correspondence to: Ngoc Bui <ngoc.bui@yale.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

multiple motifs because the sum of node-wise importance is not sufficient to capture interaction among nodes within each motif (Sundararajan et al., 2020; Masoomi et al., 2020; Zhang et al., 2021a). Therefore, the highlighted nodes might be disconnected and unintuitive to humans.

Finally, existing graph explainers only consider identifying substructures that positively affect the prediction and neglect the structures/motifs that may negatively affect the model prediction, hindering the model from giving a higher confidence score. Meanwhile, identifying negative structures can provide counterfactual reasoning, which can help practitioners avoid drawing misleading conclusions from the model's output.

Proposed work. To address the aforementioned challenges, we introduce the Myerson-Taylor interaction index, which generalizes the Myerson values (Myerson, 1977) and Shapley-Taylor index (Sundararajan et al., 2020) to capture both structure information and high-order node interactions in the graph input when assigning importance scores. The Myerson-Taylor index incorporates the structure information into the Shapley values by only allowing interactions from connected nodes, thus potentially mitigating the Out-Of-Distribution (OOD) bias of the Shapley values.

Building on this, we propose a Myerson-Taylor Structure-aware Graph Explainer (MAGE) that leverages the second-order Myerson-Taylor index to compute pair-wise interaction among nodes. These pair-wise importances are used to compute the group attribution score, accounting for the importance of a subgraph to the model prediction. MAGE then solves an optimization model to find *multiple* explanatory substructures that maximize the total absolute attribution score. Thus, MAGE can effectively identify both *positive* and *negative* motifs that contribute to the GNN output.

Extensive experiments on ten datasets and three GNN models to show MAGE's effectiveness in explaining GNN predictions. MAGE empirically outperforms seven popular, state-of-the-art baselines across diverse tasks, including molecular prediction, image, and sentiment classification. Specifically, we achieve up to 27.55% increase in the explanation accuracy compared to the best baseline.

2. Related Work

Explainability for GNN models. There are two main approaches to finding explanations for GNN models: *self-interpretable methods* and *post-hoc methods* (Chen et al., 2023; Kakkad et al., 2023; Yuan et al., 2022; Amara et al., 2022). Self-interpretable methods focus on designing model architectures that inherently generate explanations from input subgraphs (Feng et al., 2022a; Miao et al., 2022). In contrast, post-hoc explanation aims to construct explanations for existing trained models. Methods in post-hoc explainability for graph models can also be divided into two categories, *black-box* and *white-box*, depending on how they access the model information. White-box explainers usually require access to the internal structure, parameters, or gradients of the model (Feng et al., 2022b; Schnake et al., 2021; Baldassarre & Azizpour, 2019; Pope et al., 2019; Huang et al., 2024). Black-box explainers only require to query the model output to train a surrogate model (Huang et al., 2022; Zhang et al., 2021b; Pereira et al., 2023), or generative model (Chen et al., 2024; Wang & Shen, 2022; Chen & Ying, 2024; Shan et al., 2021; Yuan et al., 2020; Lin et al., 2021; Li et al., 2023) to construct explanations. Another black-box approach is perturbation-based methods (Ying et al., 2019; Schlichtkrull et al., 2020; Luo et al., 2020; Wang et al., 2021; Funke et al., 2022; Huang et al., 2024), which attribute node/edge importance by perturbing the input graphs and assess the change in model's prediction. Specifically, within this domain, (Duval & Malliaros, 2021; Yuan et al., 2021; Ye et al., 2023) propose to treat a subgraph as a supernode and other nodes of the graph as singletons. They then use the Shapley values of the supernode as its importance score. Although they leverage the graph input to compute Shapley values for L-hop neighbors around the supernode to reduce complexity, the underlying attribution score still relies on the Shapley value, which neglects the structural information (Zhang et al., 2022a). To address this, Zhang et al. (2022a) propose to use Hamiache and Navarro (HN) value (Hamiache & Navarro, 2020) to incorporate graph structure by assigning zero weight for disconnected subgraphs. However, they only focus on node-wise importance and neglect node interactions when forming multiple motif groups.

Cooperative game theory. In ML's explainability with cooperative game theory, Grabisch & Roubens (1999); Sundararajan et al. (2020); Tsai et al. (2023) propose allocation rules to analyze high-order interactions among input features of ML models. Zhang et al. (2021a); Masoomi et al. (2020) study the group attribution, where features form non-separable coalitions, acting as unified groups. However, these works neglect graph inputs, thus omitting structural information in allocating importance scores. Recently, Zhang et al. (2022a); Homberg et al. (2023) adopted the Myerson value and HN-value to explain models with graph inputs. However, they only focus on node-wise importance. In this work, we propose a generalized allocation rule that considers both graph structure and high-order node interactions.

3. Preliminaries

A graph input is denoted by $G = (V, E)$, where V is the set of nodes, and E is the set of edges. While a graph input may contain node and edge feature vectors, our framework does not exploit this information; hence, we drop the node

and edge features from the graph notation. A black-box graph neural network (GNN) is represented by a function f that takes G as input, and it outputs the probability or logit value for predicting G to be in a specific class. To simplify the notation, we use $f(V)$ to denote the output value: this notation highlights the node composition of the input, and the edges are taken implicitly. Similarly, for any subset of nodes $T \subseteq V$, we use $f(T)$ to denote the GNN output to the graph (T, E_T), where E_T is the collection of edges induced by E with both endpoints in the subset T. We refer to T as a subset of nodes or a subgraph interchangeably. To simplify notations, for any set T and nodes i and j, we use $T \cup i$ as a shorthand for $T \cup \{i\}$, and $T \cup ij$ for $T \cup \{i, j\}$.

Given a set of nodes[1] V and a function f, an attribution rule distributes the output value $f(V)$ to the members of V.

The Shapley value quantifies the potential change in the model's prediction resulting from removing or ablating a particular node. The value attributed to each node is calculated as the average of its marginal contribution over all possible coalitions it could join.

Definition 3.1 (Shapley value (Shapley, 1953))**.** *Given a function f and a set of nodes V, the Shapley value of node $i \in V$ is defined as*

$$\phi_i = \frac{1}{|V|} \sum_{T \subseteq V \setminus i} \frac{1}{\binom{|V|-1}{|T|}} (f(T \cup i) - f(T)).$$

There are several drawbacks of the Shapley values: (i) it focuses solely on node-wise importance, thus failing to illustrate the interactions among nodes; (ii) it is not suitable for graph inputs because it disregards the graph connectivity structure. We next discuss several extensions in the literature that attempt to alleviate these drawbacks.

The Shapley-Taylor index aims to capture the interactions between nodes in the explanation task. To do this, Sundararajan et al. (2020) generalized the Shapley value to k-order explanation that attributes the model's prediction to interactions of subsets of nodes of size up to k. Denote $\delta_S(T)$ as the cooperative contribution (in terms of model output) of a subset of nodes S when joining another subset T. Specifically, we can write $\delta_S(T)$ as

$$\delta_S f(T) = \sum_{W \subseteq S} (-1)^{|S|-|W|} f(W \cup T). \tag{1}$$

Consider when $S = \{i\}$, then $\delta_i f(T) = f(T \cup i) - f(T)$, which is equivalent to the marginal contribution of i to subset T. If $S = \{i, j\}$ with $i \neq j$, then

$$\delta_{ij} f(T) = f(T \cup ij) - f(T \cup i) - f(T \cup j) + f(T),$$

which is surplus created from interaction between to i and j when both joins a subset T. Sundararajan et al. (2020) defined the Shapley-Taylor index as follows.

Definition 3.2 (Shapley-Taylor index (Sundararajan et al., 2020))**.** *Given a function f and a set of nodes V, the k-order Shapley-Taylor index of a subset $S \subseteq V$, $|S| \leq k$ is defined as follows*

$$\Phi_S^k = \begin{cases} \delta_S f(\emptyset) & \text{if } |S| < k, \\ \dfrac{k}{|V|} \displaystyle\sum_{T \subseteq V \setminus S} \frac{1}{\binom{|V|-1}{|T|}} \delta_S f(T) & \text{if } |S| = k. \end{cases}$$

We observe the resemblance between the Shapley value in Def. 3.1 and the Shapley-Taylor index in Def. 3.2: the branch $|S| = k$ of Φ_S^k has the same form with the Shapley value, except that Φ_S^k utilizes the difference function δ_S to capture the case when S is not a singleton. Further, when $k = 1$, the Shapley-Taylor index recovers the Shapley value.

The Myerson value extends the Shapley value to account for interaction restrictions in graph settings (Myerson, 1977). To formally delineate the Myerson value, let $\zeta(T)$ denote the set of connected components of the subgraph induced by $T \subseteq V$ in the graph $G = (V, E)$. We then define the interaction-restricted function as

$$f|_E(T) = \sum_{R \in \zeta(T)} f(R).$$

If T is a connected subgraph on G, thus $f|_E(T) = f(T)$. If T is disconnected, the worth of subgraph T is computed as the sum of its connected components.

Definition 3.3 (Myerson value (Myerson, 1977))**.** *Given a function f and a graph (V, E), the Myerson value of a node $i \in V$ is defined as*

$$\psi_i = \frac{1}{|V|} \sum_{T \subseteq V \setminus i} \frac{1}{\binom{|V|-1}{|T|}} (f|_E(T \cup i) - f|_E(T)).$$

The Myerson value is defined directly upon the Shapley value of the interaction-restricted function $f|_E$. Considering $f|_E$ instead of f will only allow the interaction among connected nodes; thus, the Myerson values explicitly capture the graph structure information into the score ψ_i. By this definition, the Myerson value retains all the characteristics of the Shapley value. If (V, E) is a complete graph, the Myerson value coincides with the Shapley value.

4. Graph Explainer with Multiple Motifs

The explanation task focuses on finding a subgraph $S \subseteq V$ so that the output of f on S is 'most similar' to that of f on the original input V. Cooperative game-based approaches to

[1] To adapt to the graph explanation task, we use the terminologies 'node' and 'subset' throughout. In the game theory literature, 'node' is called 'player', and 'subset' is called 'coalition'.

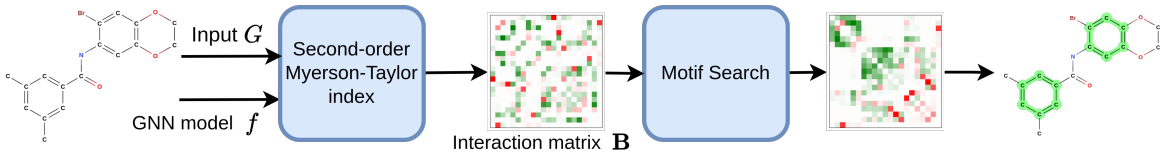


Figure 1. MAGE operates in two distinct phases: First, MAGE employs the second-order Myerson-Taylor index to calculate pairwise interactions among graph nodes, represented by an interaction matrix $\mathbf{B}$. This matrix $\mathbf{B}$ serves as the input for the motif optimization. This optimization module searches for the m most influential motifs contributing to the model's prediction.

find S often have two components: *(i)* an allocation rule that attributes the importance scores to nodes or subsets of nodes in V, *(ii)* an optimization model that takes the attribution scores and finds the optimal explanatory structures.

For component (i), we introduce the Myerson-Taylor interaction index to capture structure information and high-order interactions in graph input. For component (ii), we propose an optimization model to identify the subgraph S. Figure 1 illustrates the overall flow of our method.

4.1. Myerson-Taylor Interaction Index

We first propose the Myerson-Taylor index, which generalizes the Shapley value in both directions: capturing interactions and capturing the graph structure of the input.

Definition 4.1 (Myerson-Taylor index)**.** *Given a function f and a graph (V, E), the k-order Myerson-Taylor index of a subset $S \subseteq V$, $|S| \leq k$ is defined as*

$$\Psi_S^k = \begin{cases} \delta_S f|_E(\emptyset) & \text{if } |S| < k, \\ \dfrac{k}{|V|} \displaystyle\sum_{T \subseteq V \setminus S} \dfrac{1}{\binom{|V|-1}{|T|}} \delta_S f|_E(T) & \text{if } |S| = k. \end{cases}$$

One can contrast Definition 4.1 and 3.2 to see that we replace the original model f by the interaction-restricted function $f|_E$ that explicitly takes the graph structure of (V, E) into consideration. Figure 2a shows how an interaction-restricted function $f|_E$ differs from the original function f when evaluating a disconnected subgraph. This interaction-restricted function $f|_E$ is similar to the message-passing paradigm in GNNs in the sense that both only allow information propagation and aggregation among connected nodes. Thus, the role of $f|_E$ is to prevent the model from evaluating disconnected subgraphs, which could be pathological or OOD samples for the GNN models. In contrast, the Shapley-Taylor index Ψ^k is structure-agnostic.

In general, the Myerson-Taylor index generalizes from both the Myerson value and the Shapley-Taylor index to capture high-order interactions and structural information in the graph input. For a complete graph (V, E), the Myerson-Taylor index recovers the Shapley-Taylor index, and for $k = 1$, it recovers the Myerson value (Figure 2b).

4.2. Motif Search

This section delineates the procedure to find multiple motifs that significantly sway the model's predictions using the Myerson-Taylor interactions. It is crucial to note that our investigation extends beyond merely isolating a single motif that bolsters the model's confidence score as in prior graph explainers (Yuan et al., 2021; Zhang et al., 2022a; Ye et al., 2023). We also aim to uncover structures that obscure the model's understanding, hindering it from giving a higher confidence score.

We define the space of possible explanations for the input graph G as follows

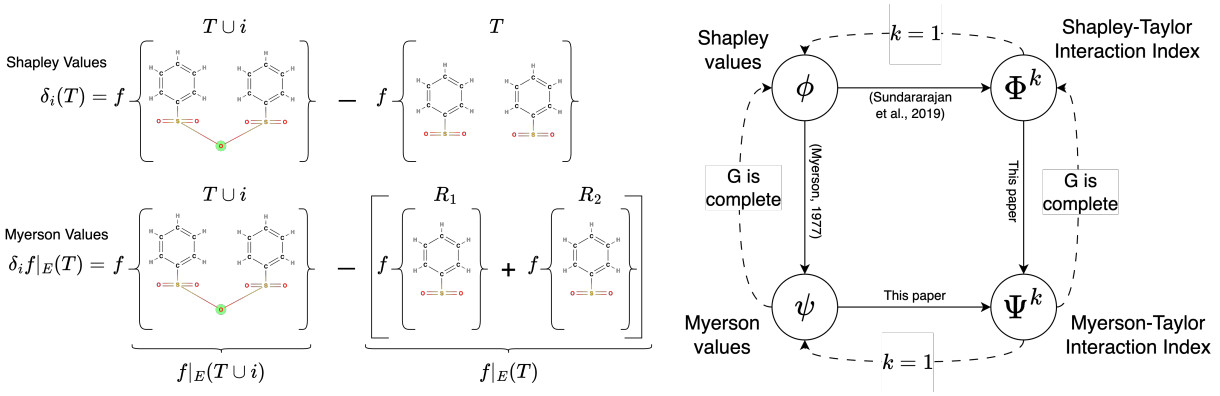
$$\mathcal{H}_{m,M} = \left\{ \begin{array}{l} (S_1, \ldots, S_m) \subseteq V^m \text{ such that :} \\ S_l \cap S_h = \emptyset \; \forall l, h, \;\; |\cup_{l=1}^m S_l| \leq M \\ S_l \text{ induces a } \textit{connected} \text{ subgraph} \end{array} \right\}. \quad (2)$$

An (m, M)-explanation for the input G is a decomposition of G into m subgraphs that are non-overlapping, of totally at most M nodes, and each subgraph is a connected subgraph. Note that we do not impose a minimum node count on each motif, allowing for the possibility that S_l could be empty and the number of highlighted motifs to fall below m. This obviates the need for users to explicitly calibrate m in order to select an appropriate explanation. Therefore, both m and M serve as complexity budgets for the explanation. A higher m allows for more dispersed explanations, and a higher M enables explanations that include more nodes.

Let $\mathbf{B} \in \mathbb{R}^{n \times n}$ be a matrix capturing the second-order Myerson-Taylor (Ψ^2) between each node, *i.e.*, $\mathbf{B}_{ij} = \Psi_{ij}^2$. Decompose $\mathbf{B} = \mathbf{B}^+ + \mathbf{B}^-$, where $\mathbf{B}^+ = \max(0, \mathbf{B})$ and $\mathbf{B}^- = \min(0, \mathbf{B})$ are matrices containing only positive and negative interactions, respectively. We define the Myerson-Taylor group attribution of a set S as

$$\text{GrAttr}(S) = \sum_{\substack{i,j \in S \\ i \leq j}} \tau \mathbf{B}_{ij}^+ + (1 - \tau) \mathbf{B}_{ij}^-,$$

where we explicitly constrain $i \leq j$ to avoid double counting. The parameter $\tau \in [0, 1]$ allows users to focus on motifs that exert positive or negative contributions or both.



(a) (b)

Figure 2. (a) Examples of how Shapley and Myerson values evaluate a disconnected coalition. The set T is not connected, and in the Myerson value, the function $f|_E(T)$ becomes the sum of output over two connected sets R_1 and R_2. (b) The relations between the four allocation methods in this paper: solid arrows indicate the generalization direction, and dashed arrows indicate the recovery direction. Conditions for recovery are written on the dashed arrows.

We propose to extract the motifs from the solution of

$$\max_{(S_1,\ldots,S_m)\in\mathcal{H}_{m,M}} \sum_{l=1}^{m} \left|\mathrm{GrAttr}(S_l)\right|. \tag{3}$$

Problem (3) maximizes the sum of *absolute* group attribution values of identified motifs. The absolute operator ensures that negative interactions are also considered in the maximization. Ideally, nodes in the same motifs should strongly interact with each other either positively or negatively, while interactions of nodes from different motifs should be negligible.

Problem (3) is a variant of the quadratic multiple knapsack problem (Hiley & Julstrom, 2006) with absolute values. One strategy to solve (3) is by linear relaxations and then using off-the-shelf MILP solvers such as MOSEK (ApS, 2019) or GUROBI (Gurobi Optimization, LLC, 2023). The detailed discussions are provided in Appendix D.

4.3. Complexity Analysis

The Myerson-Taylor index is easier to compute than the Shapley-Taylor index. The Shapley-Taylor index needs to evaluate f for all possible subgraphs of V; however, the Myerson-Taylor index needs to evaluate f only for all possible *connected* subgraphs of V. While the number of connected subgraphs is still exponential in $|V|$, the number of queries can be significantly reduced for sparse graphs. This is an advantage of the Myerson-Taylor index when explaining large, sparse inputs or deep architectures. As a trade-off, the Myerson-Taylor index requires computing the connected components for evaluated subsets, which can be done in $\mathcal{O}(|V|)$ by the standard depth-first search algorithm. Similar to other game-based explainers (Lundberg & Lee, 2017; Yuan et al., 2021; Ye et al., 2023; Sundararajan et al., 2020), we also use Monte Carlo sampling to approximate the Myerson-Taylor index.

Finally, MAGE is more computationally tractable compared to other cooperative-based graph explainers because it decomposes the attribution computation and subgraph search into two distinct phases. Thus, we only need to compute the interaction matrix $\mathbf{B}$ once for each input instance, and this can be done in parallel. In contrast, methods based on Monte Carlo Tree Search, like SubgraphX (Yuan et al., 2021) and SAME (Ye et al., 2023), encounter a bottleneck due to the need for recalculating attribution scores for each motif candidate that is explored.

5. Axiomatic Justification

The Shapley value is theoretically attractive because it is unique under a specific set of axioms, ensuring a consistent scoring allocation as we change the model and the input. This property is thus desirable for its extensions (Sundararajan et al., 2020; Myerson, 1977; Grabisch & Roubens, 1999). We now provide a theoretical justification underpinning the Myerson-Taylor interaction index introduced in Section 4.1. Let us first introduce a system of five axioms, which are inspired by the axioms that support the Shapley-Taylor interaction index (Sundararajan et al., 2020) and the Myerson value (Selçuk & Suzuki, 2014). We recite the axioms for the Shapley-Taylor index in Appendix A.

Axiom 1 (Linearity - **L**). *A k-order interaction index $\mathcal{I}^k$ is linear, i.e., for any models f_1, f_2, a graph G, and a constant α, we have $\mathcal{I}^k(f_1 + \alpha f_2, G) = \mathcal{I}^k(f_1, G) + \alpha\mathcal{I}^k(f_2, G)$, where $(f_1 + \alpha f_2)(T) = f_1(T) + \alpha f_2(T), \forall T \subseteq V$.*

Linearity is a widely accepted axiom in the solution concepts of cooperative games, imposing additive behaviors to the

allocation rules. We now define null nodes: a node $i \in V$ is a restricted null player if this node does not contribute to any coalitions it joins to form a *connected* subgraph, *i.e.*, $f(T \cup i) = \sum_{R \in \zeta(T)} f(R) + f(i)$ for any connected subset $T \cup i \subseteq V$.

Axiom 2 (Restricted Null Player - **RNP**). *For a model f and a graph $G = (V, E)$, let node $i \in V$ be a restricted null player, then the k-order interaction index $\mathcal{I}^k(f, G)$ satisfies*

(i) $\mathcal{I}_i^k(f, G) = f(i)$,

(ii) for any $S \subseteq V, |S \cup i| \leq k$, we have $\mathcal{I}_{S \cup i}^k(f, G) = 0$.

This axiom resembles the *dummy axiom* in the Shapley-Taylor interaction. However, instead of considering every possible subset of V, **RNP** only focuses on *connected* subgraphs of V dictated by the edge information E. The axiom also implies that isolated nodes are inherently categorized as null players, suggesting they should not integrate with others to form motifs larger than their singular selves. The following axiom replaces *the symmetry axiom* in the Shapley-Taylor index.

Axiom 3 (Coalitional Fairness - **CF**). *A k-order interaction index $\mathcal{I}^k$ is coalitional fair for a graph G if for any connected coalition T, i.e., $|\zeta(T)| = 1$, and two models f_1 and f_2 such that $f_1(R) = f_2(R)$ for all $R \neq T$, we then have*

$$\mathcal{I}_{S_1}^k(f_1, G) - \mathcal{I}_{S_1}^k(f_2, G) = \mathcal{I}_{S_2}^k(f_1, G) - \mathcal{I}_{S_2}^k(f_2, G),$$

for any $S_1, S_2 \subseteq T$ such that $|S_1| = |S_2|$.

Coalitional fairness dictates that a change in the value of a connected coalition should result in an equitable redistribution of interaction levels across all subsets of equivalent size within that coalition (Selçuk & Suzuki, 2014).

The next axiom requires the definition of unanimity functions. A function u_T for a $T \in V$ is *unanimity* function if the formation of the coalition T is necessary and sufficient for u_T to have non-zero value:

$$u_T(S) = \begin{cases} 1 & \text{if } S \supseteq T, \\ 0 & \text{otherwise.} \end{cases}$$

Axiom 4 (Interaction Distribution - **ID**). *A k-order interaction index $\mathcal{I}^k$ satisfies **ID** if for any unanimity function u_T, and a graph (V, E) in which $T \subseteq V$ is connected, we have*

$$\mathcal{I}_S^k(u_T, G) = 0,$$

for all $S \subsetneq T$ such that $|S| < k$.

Similar to (Sundararajan et al., 2020), the axiom **ID** ensures the lower interaction orders ($l < k$) cannot be captured by k-th order interactions and vice versa. Meanwhile, k-th order interaction of a set S with $|S| = k$ will capture interactions of S and its supersets ($\forall T \supsetneq S$). **ID** is used to introduce the efficiency axiom for the Shapley-Taylor interaction index, arguably the main advantage of the Shapley-Taylor index compared to the classical Shapley interaction index (Grabisch & Roubens, 1999). We also expect a similar axiom for the Myerson-Taylor index.

Axiom 5 (Component Efficiency - **CE**). *A k-order interaction index $\mathcal{I}^k$ is component efficient if, for any graph $G = (V, E)$ and any model f, we have*

$$\sum_{S \subseteq C, |S| \leq k} \mathcal{I}_S^k(f, G) = f(C) - f(\emptyset) \quad \forall C \in \zeta(V).$$

The **CE** axiom ensures that the confidence score of the model is fully and fairly distributed among its interacting components. In case the graph G is connected, **CE** coincides with the efficiency axiom of the Shapley-Taylor interaction, *i.e.*, $\sum_{S \subseteq V, |S| \leq k} \mathcal{I}_S^k(f, G) = f(V) - f(\emptyset)$.

To justify the Myerson-Taylor index, we show that it is a unique construction that can satisfy the above five axioms.

Theorem 5.1 (Uniqueness). *The Myerson-Taylor index is the unique interaction allocation rule that satisfies **L**, **RNP**, **CF**, **ID**, and **CE** axioms.*

This result emphasizes the importance of the Myerson-Taylor index, as it uniquely extends the Myerson value and Shapley-Taylor index to adhere to the five outlined axioms that account for structural information and high-order node interactions. The proof is relegated to Appendix B.

It is worth noting that the notion of coalition fairness in our axiom system aligns with the four-axiom system of the Myerson value proposed in (Selçuk & Suzuki, 2014) instead of the fairness notion in the original work (Myerson, 1977). In Appendix C, we generalize the classical fairness axiom (Myerson, 1977) to higher-order interactions and show that the Myerson-Taylor index also complies with this extended fairness criterion.

6. Experiments

We evaluate our method, Myerson-Taylor Structure-Aware Graph Explainer (MAGE)[2], on ten datasets and three GNN models and compare it with eight baselines to show the effectiveness of MAGE in identifying explanatory structures.

Datasets. We use ten datasets commonly used in the graph explainability literature, including synthetic data, biological, text, and image data. For *synthetic datasets*, we use Ba-2Motifs (Luo et al., 2020), BA-HouseGrid (Amara et al., 2023), and SPMotif (Wu et al., 2022) for classification tasks

[2]Our implementation is available at: https://github.com/ngocbh/MAGE/

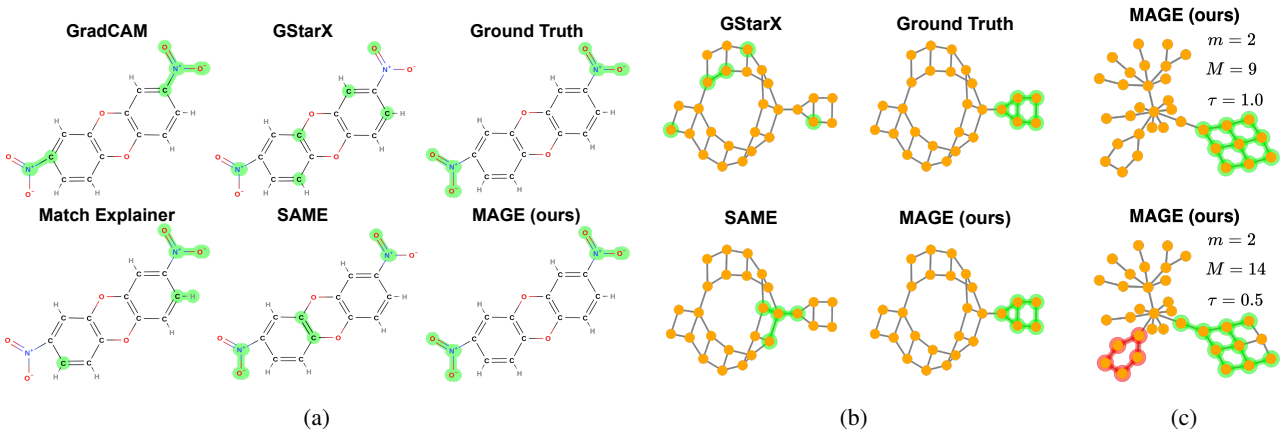


Figure 3. (a) An example in the Mutagenic dataset. Only MAGE correctly highlights the two -NO2 groups. (b) An example in the SPMotif dataset. Only MAGE can identify the house motif in the input graph. (c) An example in BA-HouseGrid shows MAGE's ability to highlight negative motifs. Green indicates positive motifs, and red indicates negative motifs. (model prediction: grid)

Table 1. Results for single motif GCN & multiple motifs GIN. On average, MAGE achieves a 59.29% improvement in F1 on single motif datasets, a 28.11% improvement in AMI on multi-motif datasets, and a 12.61% improvement in AUC across all datasets.

Method	Single Motif - GCN								Multiple Motifs - GIN							
	BA-2Motifs		BA-HouseGrid		SPMotif		MNIST75SP		BA-HouseAndGrid		BA-HouseOrGrid		Mutagenic		Benzene	
	F1↑	*AUC*↑	*F1*↑	*AUC*↑	*F1*↑	*AUC*↑	*F1*↑	*AUC*↑	*AMI*↑	*AUC*↑	*AMI*↑	*AUC*↑	*AMI*↑	*AUC*↑	*AMI*↑	*AUC*↑
GradCAM	0.634	0.753	0.459	0.485	0.491	0.616	0.193	0.492	0.825	**0.994**	0.931	**0.997**	-0.001	0.514	0.789	**0.964**
GNNExplainer	0.222	0.440	0.297	0.546	0.185	0.465	0.220	0.531	0.275	0.533	0.148	0.532	0.228	0.679	0.178	0.487
PGExplainer	0.042	0.498	0.057	0.434	0.066	0.097	0.236	0.607	0.100	0.088	0.170	0.002	0.099	0.573	0.186	0.042
Refine	0.144	0.474	0.191	0.398	0.164	0.508	0.153	0.459	0.254	0.429	0.123	0.488	0.210	0.623	0.207	0.529
MatchExplainer	0.586	0.706	0.587	0.712	0.190	0.513	0.162	0.483	0.537	0.810	0.521	0.788	0.216	0.576	0.318	0.545
SubgraphX	0.620	0.720	0.480	0.700	**0.542**	0.680	0.170	0.501	0.494	0.697	0.526	0.767	0.595	0.784	0.731	0.832
GStarX	0.180	0.480	0.267	0.544	0.203	0.498	0.280	0.517	0.203	0.494	0.130	0.484	-0.018	0.462	0.122	0.505
SAME	0.630	0.730	0.474	0.693	0.410	0.610	0.272	0.531	0.497	0.681	0.606	0.796	0.480	0.709	0.617	0.791
MAGE (ours)	**0.858**	**0.890**	**0.832**	**0.849**	**0.547**	**0.699**	**0.634**	**0.716**	**0.998**	**0.999**	**0.998**	**0.999**	**1.000**	**1.000**	**0.917**	**0.959**
Improvement (%)	35.33	18.19	41.74	19.24	0.92	2.79	126.43	17.96	20.97	0.50	7.20	0.20	68.07	27.55	16.22	-0.52

Table 2. Fidelity evaluation on sentiment classification and GCN. $\#Q$ denotes the number of GNN queries needed.

Method	GraphSST2			Twitter		
	Fid_α↑	Fid↑	#Q↓	Fid_α↑	Fid↑	#Q↓
GradCAM	0.169	0.253	N/A	0.268	0.363	N/A
SubgraphX	0.141	0.225	255K	0.238	0.32	312K
GStarX	0.161	0.273	17K	0.276	0.425	20K
SAME	0.141	0.216	24K	0.293	0.397	31K
MAGE (ours)	**0.200**	**0.337**	**2K**	**0.317**	**0.471**	**3K**

involving Barabási base structures with distinct motifs. For *molecular property prediction*, we use Mutagenic (Kazius et al., 2005) and *Benzene* (Sanchez-Lengeling et al., 2020). Molecular graphs are labeled based on their property, and the chemical fragments (-NO2 and -NH2 for Mutagenic and benzene rings for Benzene) are identified as ground-truth explanations. For *image classification*, we use MNIST75SP (Monti et al., 2017), where each image in MNIST is transformed into a graph of superpixels, with edges defined by the spatial neighborhood of the superpixels. And for *sentiment classification*, we employ two datasets *GraphSST2* and *Twitter* (Yuan et al., 2022) where each node corresponds to one word in the text, edges are constructed by the Biaffine parser (Gardner et al., 2018).

Notably, BA-2Motifs, BA-HouseGrid, SPMotif, and MNIST75SP have only one explanatory structure within a graph, while graphs in BA-HouseAndGrid, BA-HouseOrGrid, Mutag, and Benzene may have multiple explanatory structures. GraphSST2 and Twitter do not have ground truth explanations. Full descriptions of datasets are provided in Appendix E.1.1.

Models. We use three popular GNNs: GCN (Kipf & Welling, 2016), GIN (Xu et al., 2018), and GAT (Veličković et al., 2017). We report the accuracy and hyperparameters in the Appendix E.1.2. As GAT performs poorly on synthetic data, we only explain for GAT on real-world data.

Baselines. We use seven common baselines in perturbation-based graph explainability, including *GNNExplainer* (Ying et al., 2019), *PGExplainer* (Luo et al., 2020), *Refine* (Wang et al., 2021), *MatchExplainer* (Wu et al., 2023a), *SubgraphX* (Yuan et al., 2021), *GStarX* (Zhang et al., 2022a),

SAME (Ye et al., 2023). SubgraphX, GStarX, and SAME are cooperative game-based explainers; thus, they are in the same category as our method. Moreover, we also compare MAGE against *GradCAM* (Pope et al., 2019), a white-box gradient-based explainer adapted to explain GNNs.

Metrics. We use the standard metrics for explanation tasks:

- For datasets with ground truth explanations, we evaluate the accuracy of explanations using the *F1 score*, *Adjusted Mutual Information (AMI)*, and *Area Under the Curve (AUC)*. The F1 score reports the overlap of the nodes highlighted by the explainers compared to the ground truth. For datasets with multiple motifs, we use the AMI score, a widely used metric in clustering tasks, to measure the explainer's ability to identify different motifs in the graph structure. Following practice in (Ying et al., 2019; Luo et al., 2020), we also report the AUC score by comparing edge masks generated by explainers against the ground truth edge masks.
- For datasets with*out* ground-truth explanations, we utilize the *Fidelity* (Fid) (Yuan et al., 2022) to measure the faithfulness of explanations to the model's prediction. Because Fid is sensitive to OOD samples, thus favoring OOD explanations (Zheng et al., 2023; Amara et al., 2023), we also measure Fid_α proposed in (Zheng et al., 2023) to alleviate the OOD problem of Fid.

Appendix E.1.3 provides details for the above metrics.

Setup and implementation details. We split the dataset into training, validation, and test subsets with respective ratios of 0.8, 0.1, and 0.1. We train GNN models to a reasonable performance and then run the explainers for graph instances in the test datasets. We report the average metrics over instances in the test dataset.

Regarding hyperparameter settings, we set the number of explanatory nodes M and components m according to the ground truth explanations for all the baselines if they are available. For the datasets without ground truth (sentiment classification), we set M to be 30% of the number of nodes in the graph. We set $\tau = 1$ for our method as ground-truth explanations, and baselines are only for motifs with positive contributions. The number of permutations used to compute the Myerson-Taylor index is set to 200, and we use MOSEK (ApS, 2019) with default parameters for the motif search. All the results are averaged over five times tests with different random seeds.

6.1. Quantitative Results

We report the results for datasets with ground truth explanations in Table 1. Our method demonstrates superior performance to all baselines by achieving a 12.51% improvement in the AUC metric. MAGE improves the F1 score by 58.64% for datasets with a single motif. For multi-motif datasets, MAGE also improves the AMI score by 28.11% compared to the baselines, indicating more faithful explanations in datasets with multiple explanatory structures. Notably, MAGE accurately identifies all -NO2 and -NH2 chemical groups contributing to a molecule's mutagenic property for the GCN model. Compared to GradCAM, a white-box gradient-based method, MAGE's explanations are better than GradCAM by 17% in the AUC and 65% in the F1 score. For sentiment classification tasks without ground truth (Table 2), MAGE achieves a 14% higher fidelity score than the best baselines. Moreover, Table 2 also shows that MAGE is much more query-efficient than other game-based methods. Additional results for other models and running time analysis are provided in the appendix.

Table 3. Ablating Shapley-Taylor (Φ^2) and Myerson-Taylor (Ψ^2) indices and connectivity constraints in the problem (3).

Method	BA-2Motifs		BA-HouseGrid	
	F1 ↑	*AUC*↑	*F1* ↑	*AUC*↑
MAGE (Φ^2) w/o connectivity	0.699	0.773	0.634	0.735
MAGE (Φ^2) w/ connectivity	0.709	0.787	0.636	0.734
MAGE (Ψ^2) w/o connectivity	**0.854**	**0.885**	0.819	0.838
MAGE (Ψ^2) w/ connectivity	**0.858**	**0.890**	**0.832**	**0.849**

6.2. Qualitative Results

We show qualitative results for different tasks in Figure 9, and Figures 3a-3c. More examples are in Appendix E.2.

Positive motifs. For topology-based tasks, as shown in Figure 3b, MAGE accurately identifies the house motif that represents the graph class. For molecular classification tasks, including Mutagenic and Benzene, MAGE can highlight all chemical groups that exist in the molecular structure (e.g., -NO2 and -NH2 in Mutagenic shown in Figure 3a, and carbon rings in Benzene in Figure 9). Thus, the structures highlighted by MAGE align with the ground truth explanations. In contrast, other baselines struggle to provide meaningful explanations in these cases. Moreover, Figure 4 shows an example from sentiment classification where MAGE adeptly highlights the main verb 'deserves' in a sentence, which is crucial for assessing the sentence's overall sentiment. Appendix E.2.4 provides more visualizations for the remaining datasets.

Negative motifs. Figure 3c shows that MAGE can identify substructures with negative contributions to the prediction. We examine a GIN model trained on the BA-HouseGrid dataset to classify house and 3×3 grid motifs. We explain the model's prediction for a grid example with a manually injected five-cycle motif into the structure. When we set $\tau = 1.0$, MAGE can correctly identify the grid structure, which causes the model prediction. As we calibrate $\tau = 0.5$ to consider both negative and positive contributions, MAGE can effectively highlight five-cycle and grid motifs. Here, a five-cycle structure may mislead the model's prediction

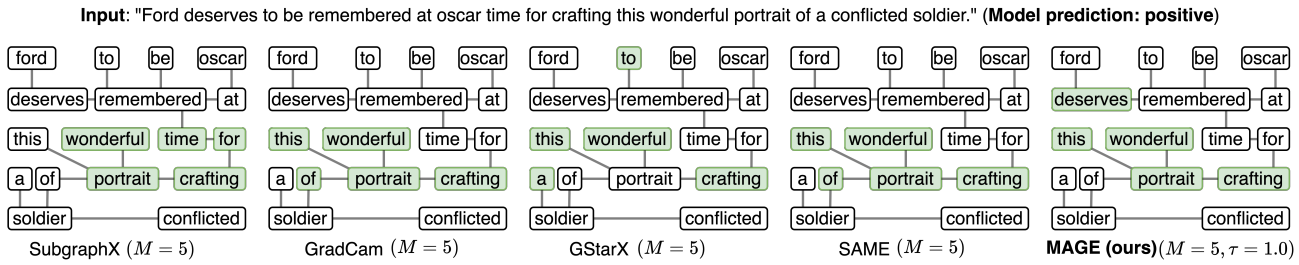


Figure 4. An example in the Graph-SST2 dataset. MAGE's explanation is more concise and correctly captures the main verb 'deserves', crucial to determining a sentence's sentiment, while other baselines fail to identify it.

towards a label for the house motif.

6.3. Ablation Study

This experiment shows the usefulness of structural information in graph explainer. *First*, we ablate the second-order Myerson-Taylor index in MAGE with the second-order Shapley-Taylor. *Second*, we run MAGE without enforcing the connectivity constraint for explanatory structures (S_l) in the optimization problem (3). Other settings are the same as in Table 1. Table 3 shows that the Myerson-Taylor index improves the MAGE's explanation significantly compared to the Shapley-Taylor, and the connectivity constraint also improves the explanation accuracy. Connectivity constraint for each motif is important to avoid fragmented explanations, which are hard to interpret for humans.

We also conduct an ablation study to compare the group attribution approximation of four methods, Shapley, Myerson, second-order Shapley-Taylor, and second-order Myerson-Taylor on Imagenet with ResNet50 and ViT16 models. We leave this experiment to the appendix.

7. Conclusion

This paper introduced the Myerson-Taylor index, which captures high-order interactions and the graph structure to explain GNN models. We proposed MAGE, a graph explainer that leverages the second-order Myerson-Taylor index to compute the motifs' attributions and highlight ones that are influential to the GNN's prediction, both positively and negatively. Extensive experiments on various domains show the compelling results of MAGE compared to other baselines.

Limitation and future work. Our approach, similar to other game-based explainers, offers consistent, stable, and model-agnostic importance scores but at the expense of high computational costs. Exact computations for these methods require exponential time with respect to the number of nodes ($|V|$) in a graph. In practice, we usually need to approximate the importance scores using Monte Carlo sampling. Notably, MAGE outperforms other game-based explainers for GNNs in terms of efficiency, requiring fewer model queries and offering faster run times.

Further, our Myerson-Taylor index treats GNN models as black-box models, thus permitting unrestricted information sharing among nodes in a connected component. This feature may not align with the practical, layer-restricted information propagation in GNNs. Addressing this discrepancy remains an open question. Moreover, as the Myerson-Taylor index uniquely generalizes the Myerson and Shapley-Taylor indices to include network structures and retains all their characteristics, extensions of the Myerson-Taylor index to study the interactions of players in (weighted) network games would be a potential research direction.

Acknowledgement

This project is made possible through the generous support of Snap Inc. Viet Anh Nguyen gratefully acknowledges the generous support from the CUHK's Improvement on Competitiveness in Hiring New Faculties Funding Scheme and the CUHK's Direct Grant Project Number 4055191.

Impact Statement

This paper primarily focuses on methods for explaining graph neural network (GNN) predictions, aiming to enhance the interpretability of black-box models. The datasets we utilize are publicly available. Our work has many potential societal consequences, none of which must be specifically highlighted here.

References

Agarwal, C., Queen, O., Lakkaraju, H., and Zitnik, M. Evaluating explainability for graph neural networks. *Scientific Data*, 10(1):144, 2023. 21

Algaba, E., Fragnelli, V., and Sánchez-Soriano, J. *Handbook of the Shapley value*. CRC Press, 2019. 19

Althaus, E., Blumenstock, M., Disterhoft, A., Hildebrandt,

A., and Krupp, M. Algorithms for the maximum weight connected-induced subgraph problem. In *International Conference on Combinatorial Optimization and Applications*, pp. 268–282. Springer, 2014. 21

Amara, K., Ying, R., Zhang, Z., Han, Z., Shan, Y., Brandes, U., Schemm, S., and Zhang, C. Graphframex: Towards systematic evaluation of explainability methods for graph neural networks. In *Learning on Graph Conference*, 2022. 2

Amara, K., El-Assady, M., and Ying, R. Ginx-Eval: Towards in-distribution evaluation of graph neural network explanations. *arXiv preprint arXiv:2309.16223*, 2023. 6, 8, 22

ApS, M. *The MOSEK optimization toolbox for MATLAB manual. Version 9.0.*, 2019. URL http://docs.mosek.com/9.0/toolbox/index.html. 5, 8, 21

Baldassarre, F. and Azizpour, H. Explainability techniques for graph convolutional networks. *arXiv preprint arXiv:1905.13686*, 2019. 2

Béal, S. and Navarro, F. Necessary versus equal players in axiomatic studies. *Operations Research Letters*, 48(3): 385–391, 2020. 18

Chen, J. and Ying, R. Tempme: Towards the explainability of temporal graph neural networks via motif discovery. *Advances in Neural Information Processing Systems*, 36, 2024. 2

Chen, J., Amara, K., Yu, J., and Ying, R. Generative explanations for graph neural network: Methods and evaluations. *IEEE Data Engineering Bulletin*, 2023. 2

Chen, J., Wu, S., Gupta, A., and Ying, R. D4explainer: In-distribution explanations of graph neural network via discrete denoising diffusion. *Advances in Neural Information Processing Systems*, 36, 2024. 2

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 25

Duval, A. and Malliaros, F. D. Graphsvx: Shapley value explanations for graph neural networks. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II 21*, pp. 302–318. Springer, 2021. 1, 2

Feng, A., You, C., Wang, S., and Tassiulas, L. Kergnns: Interpretable graph neural networks with graph kernels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 6, pp. 6614–6622, 2022a. 2

Feng, Q., Liu, N., Yang, F., Tang, R., Du, M., and Hu, X. DEGREE: Decomposition based explanation for graph neural networks. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=Ve0Wth3ptT_. 1, 2

Funke, T., Khosla, M., Rathee, M., and Anand, A. Z orro: Valid, sparse, and stable explanations in graph neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 2

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N., Peters, M., Schmitz, M., and Zettlemoyer, L. Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*, 2018. 7, 22

Grabisch, M. and Roubens, M. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28:547–565, 1999. 2, 5, 6, 14

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2023. URL https://www.gurobi.com. 5, 21

Hamiache, G. A value with incomplete communication. *Games and Economic Behavior*, 26(1):59–78, 1999. 18

Hamiache, G. and Navarro, F. Associated consistency, value and graphs. *International Journal of Game Theory*, 49: 227–249, 2020. 2

Han, K., Wang, Y., Guo, J., Tang, Y., and Wu, E. Vision GNN: An image is worth graph of nodes. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=htM1WJZVB2I. 1

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. 25

Hiley, A. and Julstrom, B. A. The quadratic multiple knapsack problem and three heuristic approaches to it. In *Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation*, pp. 547–552, 2006. 5, 21

Homberg, S., Janosch, M., Morris, G. M., and Koch, O. Interpreting graph neural networks with Myerson values for cheminformatics approaches. 2023. 2

Huang, Q., Yamada, M., Tian, Y., Singh, D., and Chang, Y. Graphlime: Local interpretable model explanations for graph neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 2

Huang, R., Shirani, F., and Luo, D. Factorized explainer for graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 12626–12634, 2024. 2

Kakkad, J., Jannu, J., Sharma, K., Aggarwal, C., and Medya, S. A survey on explainability of graph neural networks. *arXiv preprint arXiv:2306.01958*, 2023. 1, 2

Kazius, J., McGuire, R., and Bursi, R. Derivation and validation of toxicophores for mutagenicity prediction. *Journal of Medicinal Chemistry*, 48(1):312–320, 2005. 7, 21

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 7, 22

Li, W., Li, Y., Li, Z., HAO, J., and Pang, Y. DAG matters! GFlownets enhanced explainer for graph neural networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=jgmuRzM-sb6. 2

Lin, W., Lan, H., and Li, B. Generative causal explanations for graph neural networks. In *International Conference on Machine Learning*, pp. 6666–6679. PMLR, 2021. 2

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017. 1, 5, 25

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*, 33:19620–19631, 2020. 1, 2, 6, 7, 8, 21, 22

Mak-Hau, V. and Yearwood, J. A mixed-integer linear programming approach for soft graph clustering. *arXiv preprint arXiv:1906.04860*, 2019. 21

Masoomi, A., Wu, C., Zhao, T., Wang, Z., Castaldi, P., and Dy, J. Instance-wise feature grouping. *Advances in Neural Information Processing Systems*, 33:13374–13386, 2020. 2

Mastropietro, A., Pasculli, G., Feldmann, C., Rodríguez-Pérez, R., and Bajorath, J. Edgeshaper: Bond-centric Shapley value-based explanation method for graph neural networks. *Iscience*, 25(10), 2022. 25

Miao, S., Liu, M., and Li, P. Interpretable and generalizable graph learning via stochastic attention mechanism. In *International Conference on Machine Learning*, pp. 15524–15543. PMLR, 2022. 2

Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., and Bronstein, M. M. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5115–5124, 2017. 7, 21

Myerson, R. B. Graphs and cooperation in games. *Mathematics of Operations Research*, 2(3):225–229, 1977. 2, 3, 5, 6, 18

Owen, G. Multilinear extensions of games. *Management Science*, 18(5-part-2):64–79, 1972. 14

Pereira, T., Nascimento, E., Resck, L. E., Mesquita, D., and Souza, A. Distill n'explain: Explaining graph neural networks using simple surrogates. In *International Conference on Artificial Intelligence and Statistics*, pp. 6199–6214. PMLR, 2023. 2

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 (CVPR)*, pp. 10772–10781, 2019. 1, 2, 8, 23, 27

Sanchez-Lengeling, B., Wei, J., Lee, B., Reif, E., Wang, P., Qian, W., McCloskey, K., Colwell, L., and Wiltschko, A. Evaluating attribution for graph neural networks. *Advances in Neural Information Processing Systems*, 33: 5898–5910, 2020. 7, 21

Sato, R. A survey on the expressive power of graph neural networks. *arXiv preprint arXiv:2003.04078*, 2020. 21

Schlichtkrull, M. S., De Cao, N., and Titov, I. Interpreting graph neural networks for nlp with differentiable edge masking. *arXiv preprint arXiv:2010.00577*, 2020. 1, 2

Schnake, T., Eberle, O., Lederer, J., Nakajima, S., Schütt, K. T., Müller, K.-R., and Montavon, G. Higher-order explanations of graph neural networks via relevant walks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7581–7596, 2021. 2

Selçuk, Ö. and Suzuki, T. An axiomatization of the Myerson value. *Contributions to Game Theory and Management*, 7, 2014. 5, 6, 17, 18

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*, pp. 618–626, 2017. 27

Shan, C., Shen, Y., Zhang, Y., Li, X., and Li, D. Reinforcement learning enhanced explainer for graph neural networks. *Advances in Neural Information Processing Systems*, 34:22523–22533, 2021. 2

Shapley, L. S. A value for n-person games. In *Contributions to the Theory of Games (AM-28), Volume II*. Princeton University Press Princeton, 1953. 1, 3

Shi, W. and Rajkumar, R. Point-gnn: Graph neural network for 3d object detection in a point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1711–1719, 2020. 1

Sterling, T. and Irwin, J. J. Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling*, 55(11):2324–2337, 2015. 21

Sun, M., Zhao, S., Gilvary, C., Elemento, O., Zhou, J., and Wang, F. Graph convolutional networks for computational drug development and discovery. *Briefings in bioinformatics*, 21(3):919–935, 2020. 1

Sundararajan, M., Dhamdhere, K., and Agarwal, A. The Shapley Taylor interaction index. In *International Conference on Machine Learning*, pp. 9259–9268. PMLR, 2020. 2, 3, 5, 6, 14, 17, 18, 19

Tsai, C.-P., Yeh, C.-K., and Ravikumar, P. Faith-shap: The faithful Shapley interaction index. *Journal of Machine Learning Research*, 24(94):1–42, 2023. 2

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. 7, 22

Vinh, N. X., Epps, J., and Bailey, J. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pp. 1073–1080, 2009. 23

Vu, M. and Thai, M. T. Pgm-explainer: Probabilistic graphical model explanations for graph neural networks. *Advances in neural information processing systems*, 33: 12225–12235, 2020. 1

Wang, X. and Shen, H.-W. Gnninterpreter: A probabilistic generative model-level explanation for graph neural networks. *arXiv preprint arXiv:2209.07924*, 2022. 2

Wang, X., Wu, Y., Zhang, A., He, X., and Chua, T.-S. Towards multi-grained explainability for graph neural networks. *Advances in Neural Information Processing Systems*, 34:18446–18458, 2021. 2, 7

Wu, F., Li, S., Wu, L., Radev, D., Jiang, Y., Jin, X., Niu, Z., and Li, S. Z. Explaining graph neural networks via non-parametric subgraph matching. In *International Conference on Machine Learning*, pp. 9259–9268. PMLR, 2023a. 7

Wu, L., Chen, Y., Shen, K., Guo, X., Gao, H., Li, S., Pei, J., Long, B., et al. Graph neural networks for natural language processing: A survey. *Foundations and Trends® in Machine Learning*, 16(2):119–328, 2023b. 1

Wu, Y.-X., Wang, X., Zhang, A., He, X., and Chua, T.-S. Discovering invariant rationales for graph neural networks. *arXiv preprint arXiv:2201.12872*, 2022. 6, 21

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Philip, S. Y. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2020. 1

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018. 7, 22

Ye, Z., Huang, R., Wu, Q., and Liu, Q. Same: Uncovering GNN black box with structure-aware Shapley-based multipiece explanations. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 2, 4, 5, 8

Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., and Ravikumar, P. K. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32, 2019. 25

Ying, Z., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. Gnnexplainer: Generating explanations for graph neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2, 7, 8, 22, 25

Yuan, H., Tang, J., Hu, X., and Ji, S. Xgnn: Towards model-level explanations of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 430–438, 2020. 1, 2

Yuan, H., Yu, H., Wang, J., Li, K., and Ji, S. On explainability of graph neural networks via subgraph explorations. In *International Conference on Machine Learning*, pp. 12241–12252. PMLR, 2021. 1, 2, 4, 5, 7, 23, 24

Yuan, H., Yu, H., Gui, S., and Ji, S. Explainability in graph neural networks: A taxonomic survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (5):5782–5799, 2022. 1, 2, 7, 8, 21, 22, 24

Zhang, H., Xie, Y., Zheng, L., Zhang, D., and Zhang, Q. Interpreting multivariate Shapley interactions in dnns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 12, pp. 10877–10886, 2021a. 2

Zhang, S., Liu, Y., Shah, N., and Sun, Y. Gstarx: Explaining graph neural networks with structure-aware cooperative games. In *Advances in Neural Information Processing Systems*, 2022a. 1, 2, 4, 7

Zhang, Y., Defazio, D., and Ramesh, A. Relex: A model-agnostic relational model explainer. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 1042–1049, 2021b. 2

Zhang, Y., Hu, Q., Xu, G., Ma, Y., Wan, J., and Guo, Y. Not all points are equal: Learning highly efficient point-based detectors for 3d lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18953–18962, June 2022b. 1

Zheng, K., Yu, S., and Chen, B. Ci-gnn: A granger causality-inspired graph neural network for interpretable brain network-based psychiatric diagnosis. *Neural Networks*, pp. 106147, 2024. 22

Zheng, X., Shirani, F., Wang, T., Cheng, W., Chen, Z., Chen, H., Wei, H., and Luo, D. Towards robust fidelity for evaluating explainability of graph neural networks. *arXiv preprint arXiv:2310.01820*, 2023. 8, 23, 24, 27

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020. 1

A. Axioms of the Shapley-Taylor Interaction Index

For comparison purposes, we provide the axiomatic characterizations of the Shapley-Taylor interaction index (Sundararajan et al., 2020).

Axiom A.1 (Linearity - **L**). *$\mathcal{I}^k$ is a linear function, i.e., for any two functions f_1, f_2, we have $\mathcal{I}^k(f_1 + \alpha f_2) = \mathcal{I}^k(f_1) + \alpha\mathcal{I}^k(f_2)$, for any constant α.*

Axiom A.2 (Dummy - **D**). *If i is a dummy node, i.e., for any subset $T \subseteq V \setminus i$, $f(T \cup i) = f(T) + f(i)$, then*

$$\mathcal{I}^k_{S\cup i} = \begin{cases} f(i) & \text{if } S = \emptyset, \\ 0 & \text{if } S \subseteq V, |S| \leq k-1. \end{cases}$$

Axiom A.3 (Symmetry - **S**). *For any permutation order π on V, we have $\mathcal{I}^k_S(f) = \mathcal{I}_{\pi S}(\pi f)$ where $\pi f(T) = f(\pi T)$.*

Axiom A.4 (Interaction Distribution - **ID**). *For a unanimity function u_T, for all $S \subsetneq T, |S| < k$, we have $\mathcal{I}^k_S(u_T) = 0$.*

Axiom A.5 (Efficiency - **E**). *For any function f, we have*

$$\sum_{S\subseteq V, |S|\leq k} \mathcal{I}^k_S(f) = f(V).$$

Theorem A.1 (Shapley-Taylor uniqueness (Sundararajan et al., 2020))**.** *Shapley-Taylor index is the only interaction index that satisfies five axioms: **L**, **D**, **S**, **E**, and **ID**.*

While there exist several methods to compute interaction among players (Grabisch & Roubens, 1999), the key attraction of the Shapley-Taylor interaction index is its satisfaction with the efficiency axiom (Axiom A.5). The idea of the Shapley-Taylor index is drawn from the Taylor expansion of the multilinear extension of a cooperative game (Owen, 1972; Sundararajan et al., 2020). It posits that the interactions within a subset S of size l less than k are analogous to the l-th order term in a Taylor series, capturing only the interactions inherent to that subset. On the other hand, the k-order Shapley-Taylor indices are akin to the Lagrange remainder of the series, encompassing both the interactions of the set itself and those of higher orders.

B. Proofs

This section provides the detailed proofs for Theorem 5.1.

Let $\mathcal{F}^V$ denote a space of functions acting on the set of vertices V without restriction

$$\mathcal{F}^V = \{f : 2^V \to \mathbb{R}\},$$

and $\mathcal{F}^V|_E$ be a space of interaction-restricted functions on the graph (V, E)

$$\mathcal{F}^V|_E = \left\{ f|_E = \sum_{R\in\zeta(T)} f(R) \text{ where } f \in \mathcal{F} \right\}.$$

Thus, the space of interaction-restricted function $\mathcal{F}^V|_E$ is a subspace of $\mathcal{F}$.

For convenience, we present a definition of the Myerson-Taylor index using the Shapley-Taylor index as follows.

Definition B.1 (Myerson-Taylor index)**.** *Given a function f and a graph (V, E), the k-order Myerson-Taylor index of a subset $S \subseteq V$, $|S| \leq k$ is*

$$\Psi^k_S(f, E) = \Phi^k_S(f|_E),$$

where Φ^k_S is the Shapley-Taylor index of S with respect to the interaction-restricted function $f|_E$.

By Definition B.1, the Myerson-Taylor index is defined directly upon the Shapley-Taylor index, thus inherently retaining all characteristics of the Shapley-Taylor index for interaction-restricted functions.

To prove Theorem 5.1, we first need to show the Myerson-Taylor index satisfies five axioms, which can be done by combining definitions of the Myerson-Taylor index and interaction-restricted functions.

Proposition B.2. *The Myerson-Taylor index satisfies the linearity (**L**) axiom.*

Proof of Proposition B.2. For a graph (V, E) and two functions $f_1, f_2 \in \mathcal{F}^V$, we have

$$\begin{aligned}(f_1 + \alpha f_2)|_E(T) &= \sum_{R \in \zeta(T)} (f_1 + \alpha f_2)(R) \\ &= \sum_{R \in \zeta(T)} f_1(R) + \sum_{R \in \zeta(T)} \alpha f_2(R) \\ &= f_1|_E(T) + \alpha f_2|_E(T).\end{aligned}$$

Thus, the communication restriction preserves the linearity.

As the Myerson-Taylor index is defined upon the Shapley-Taylor index, which is also a linear function on $\mathcal{F}^V$, the Myerson-Taylor index is a linear function on $\mathcal{F}^V$. □

Proposition B.3. *The Myerson-Taylor index satisfies the restricted null player (**RNP**) axiom.*

Proof of Proposition B.3. We first show that for any restricted null node i of the graph (V, E) and function f, i is also a dummy node of the interaction-restricted function $f|_E$.

Consider any restricted null node i, we have

$$f(T \cup i) = \sum_{R \in \zeta(T)} f(R) + f(i),$$

for any connected subset $T \cup i \subseteq V$.

If $T \cup i$ is connected, we directly have $f|_E(T \cup i) = f(T \cup i) = f|_E(T) + f(i)$.

If $T \cup i$ is not connected,

$$f|_E(T \cup i) = \sum_{R \in \zeta(T \cup i)} f(R). \tag{4}$$

Because i is a restricted null node, i does not provide any profit when joining others to form a connected subgraph, hence, for a connected component $R \in \zeta(T \cup i)$ such that $R \ni i$, we have

$$\begin{aligned}f(R) &= \sum_{W \in \zeta(R \setminus i)} f(W) + f(i) \\ &= f|_E(R \setminus i) + f(i).\end{aligned}$$

Combining with (4), we have $f|_E(T \cup i) = f|_E(T) + f(i)$ in case $T \cup i$ is not connected.

We then deduce that i is also a dummy node with respect to the interaction-restricted function $f|_E$, *i.e.*,

$$f|_E(T \cup i) = f|_E(T) + f|_E(i),$$

for any $T \subseteq V$.

By the dummy feature axiom (Axiom A.2), we have

$$\Psi_S^k(f, E) = \Phi_S^k(f|_E) = \begin{cases} f(S) & \text{if } |S| = 1, \\ 0 & \text{if } S \ni i. \end{cases}$$

□

Proposition B.4. *The Myerson-Taylor index satisfies the component efficiency axiom (**CE**).*

Proof of Proposition B.4. For any graph (V, E), a function f, and a connected component $C \in \zeta(V)$, we define a characteristic game f^C such that

$$f^C|_E(T) = \sum_{R \in \zeta(T \cap C)} f(R) \quad \forall T \subseteq V.$$

Because $f^C|_E(T) = f^C|_E(T \cap C)$ for any $T \subseteq V$, C will be the carrier (the grand coalition) of $f^C|_E$, which means any nodes not in C are dummy nodes. By the dummy axiom (Axiom A.2), we have $\Psi_S^k(f^C, E) = \Phi_S^k(f^C|_E) = 0$, for all $S \not\subseteq C$ such that $|S| \leq k$.

On the other hand, we notice that every connected component of a subgraph induced by a subset T is also connected in the graph (V, E), we thus have

$$\begin{aligned} f|_E(T) &= \sum_{R \in \zeta(T)} f(R) \\ &= \sum_{C \in \zeta(V)} \sum_{R \in \zeta(T \cap C)} f(R) \\ &= \sum_{C \in \zeta(V)} f^C|_E(T). \end{aligned}$$

By the linearity axiom (Axiom A.1), for any component $C \in \zeta(V)$, we have

$$\begin{aligned} \sum_{S \subseteq C, |S| \leq k} \Psi_S^k(f, E) &= \sum_{S \subseteq C, |S| \leq k} \Phi_S^k(f|_E) \\ &= \sum_{C' \in \zeta(V)} \sum_{S \subseteq C, |S| \leq k} \Phi_S^k(f^{C'}|_E) \\ &= \sum_{S \subseteq C, |S| \leq k} \Phi_S^k(f^C|_E) \\ &= f^C|_E(C) = f(C). \end{aligned}$$

The third equality follows as any node $i \in C$ is the dummy node in the game of other component $f^{C'}|_E$, $C' \neq C$; thus, $\Phi_S^k(f^{C'}|_E) = 0$, for any $S \subseteq C$. The last equality follows from the efficiency axiom of the Shapley-Taylor index (Axiom A.5). This completes the proof. □

Proposition B.5. *The Myerson-Taylor index satisfies the coalitional fairness axiom (**CF**).*

Proof of Proposition B.5. For any graph (V, E) and a connected coalition T, consider two functions f_1, f_2 such that $f_1(R) = f_2(R), \forall R \neq T$. We define a function $g = f_1 - f_2$ so that, for any subset $R \subseteq V$, we have

$$g(R) = \begin{cases} f_1(R) - f_2(R) & \text{if } R = T, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, $g = \beta u_T$ where $\beta = f_1(T) - f_2(T)$ and u_T is the unanimity function defined on the subset T. For any $S \subseteq T, |S| \leq k$, we have $\Psi_S^k(g, E) = \beta \Phi_S^k(u_T|_E) = \beta \Phi_S^k(u_T)$. The last equation is in fact that T is connected.

By the symmetry axiom (Axiom A.3), we have $\Phi_{S_1}^k(u_T) = \Phi_{S_2}^k(u_T)$ for any $S_1, S_2 \subseteq T$ such that $|S_1| = |S_2|$. We deduce $\Psi_{S_1}^k(g, E) = \Psi_{S_2}^k(g, E)$. The proof follows from the linearity of the Myerson-Taylor index. □

Proposition B.6. *The Myerson-Taylor interaction index satisfies the interaction distribution axiom (**ID**).*

Proof of Proposition B.6. Consider unanimity function u_T and a graph (V, E) in which $T \subseteq V$ is connected.

For any $S \subsetneq T$ such that $|S| < k$, we have

$$\begin{aligned}\Psi_S^k(u_T, E) &= \Phi_S^k(u_T|_E) = \delta_S u_T|_E(\emptyset) \\ &= \sum_{W \subseteq S} (-1)^{|S|-|W|} u_T|_E(W) \\ &= \sum_{W \subseteq S} (-1)^{|S|-|W|} \sum_{R \in \zeta(W)} u_T(R) \\ &= 0.\end{aligned}$$

The last equality follows as $u_T(R) = 0, \forall R \subseteq S \subsetneq T$. Hence, $\Psi_S^k(u_T, E) = 0$ for any $S \subsetneq T, |S| < k$. □

We present several elementary results needed to show the uniqueness.

The following result extends the result of Selçuk & Suzuki (2014, Lemma 1) to k-order interactions.

Lemma B.7 (Selçuk & Suzuki (2014))**.** *If an interaction index $\mathcal{I}^k$ satisfies **L** and **RNP** axioms then, for any graph (V, E), a function f, and a subset $S \subseteq V, |S| \leq k$, we have $\mathcal{I}_S^k(f, E) = \mathcal{I}_S^k(f|_E, E)$.*

Proof of Lemma B.7. Consider the function $g = f - f|_E$. We have

$$g(T) = \begin{cases} 0 & \text{if } T \text{ is connected,} \\ f(T) - \sum_{R \in \zeta(T)} (R) & \text{otherwise.} \end{cases}$$

Notice that, by definition, $g(T \cup i) - \sum_{R \in \zeta(T)} g(R) = 0$ for any node i and connected coalition $T \cup i$. Thus, every node is a restricted null player in the game with the characteristic function g and, by **RNP** axiom, must receive zero payoffs. That is, $\mathcal{I}_S^k(g, E) = 0$ for any subset S. By **L** axiom, we deduce that $\mathcal{I}_S^k(f, E) = \mathcal{I}_S^k(f|_E, E)$. □

Proposition B.8. *Let $\mathcal{I}^k$ be a k-order interaction index that satisfies five axioms, **L**, **RNP**, **CF**, **ID**, and **CE**. For any graph (V, E) and a connected subset T, then*

$$\mathcal{I}_S^k(u_T, E) = \begin{cases} 1 & \textit{if } S = T \textit{ and } |S| < k, \\ 0 & \textit{if } S \neq T \textit{ and } |S| < k, \\ \frac{1}{\binom{|T|}{k}} & \textit{if } S \subseteq T \textit{ and } |S| = k, \\ 0 & \textit{if } S \not\subseteq T \textit{ and } |S| = k. \end{cases}$$

Proof of Proposition B.8. The procedure is similar to Sundararajan et al. (2020) in which the restricted null player (**RNP**) replaces the dummy feature (Axiom A.2), and the coalitional fairness (**CF**) plays the role of symmetry axiom (Axiom A.3).

We notice that any node $i \notin T$ is a restricted null node of u_T as T is connected. Hence, by **RNP**, $\mathcal{I}_S^k(u_T, E) = 0$ for any $S \setminus T \neq \emptyset$ (or $S \not\subseteq T$).

Consider the case $|T| < k$, for any subset S such that $S \subsetneq T$, $\mathcal{I}_S^k(u_T, E) = 0$ by **ID** axiom. Therefore, $\mathcal{I}_S^k(u_T, E) = 0$ for any $S \neq T$. By **CE** axiom, we have $\sum_S \mathcal{I}_S^k(u_T, E) = u_T(C) = 1$ where $C \in \zeta(V), C \supseteq T$ is a connected component containing T. As $\mathcal{I}_S^k(u_T, E) = 0$ for all $S \neq T$. We deduce $\mathcal{I}_S^k(u_T, E) = 1$ for $S = T$.

For the case $|T| \geq k$, by **RNP** and **ID** axiom, we also have $\mathcal{I}_S^k(u_T, E) = 0$ for any S such that $S \not\subseteq T$ or $S \subsetneq T, |S| < k$. Hence, by **CE** axiom, we have $\sum_{S \subseteq T, |S|=k} \mathcal{I}_S^k(u_T, E) = 1$. By the coalitional fairness axiom (**CF**), we then have $\mathcal{I}_S^k(u_T, E) = \frac{1}{\binom{|T|}{k}}$.

This completes the proof. □

Proof of Theorem 5.1. Propositions B.2-B.6 assert that the Myerson-Taylor index satisfies five axioms. The following shows its uniqueness.

Let $\mathcal{I}^k$ be a k-order interaction index that satisfies five axioms. Consider a graph (V, E) and a function $f \in \mathcal{F}^V$. By Lemma B.7, we have $\mathcal{I}^k(f, E) = \mathcal{I}^k(f|_E, E)$ as $\mathcal{I}^k$ satisfies **L** and **RNP**. Thus, we only need to show that $\mathcal{I}^k$ is a unique allocation rule in the space of interaction-restricted functions $\mathcal{F}^V|_E = \{f|_E : f \in \mathcal{F}^V\}$

By Hamiache (1999, Lemma 2), an interaction-restricted function $f|_E \in \mathcal{F}^V|_E$ can be decomposed into

$$f|_E = \sum_{\substack{T \subseteq V \\ T \text{ is connected}}} \Delta_{f|_E}(T) u_T,$$
$$\text{where} \quad \Delta_{f|_E}(T) = \sum_{\substack{R \subseteq T \subseteq \mathcal{N}(R) \\ R \text{ is connected}}} (-1)^{|T|-|R|} f(R).$$

Here, $\mathcal{N}(R)$ is the set of vertices that are adjacent to R, *i.e.*, $\mathcal{N}(R) = \{i : \exists ij \in E, j \in R\}$. Hence, the set of unanimity functions of connected subgraphs $\{u_T : T \subseteq V \text{ where } T \text{ is connected}\}$ forms a basis of the space of $\mathcal{F}^V|_E$. By the **L** axiom, we have

$$\mathcal{I}^k_S(f|_E, E) = \sum_{\substack{T \subseteq V \\ T \text{ is connected}}} \Delta_{f|_E}(T) \mathcal{I}^k_S(u_T, E).$$

Since $\mathcal{I}^k$ satisfies five axioms, $\mathcal{I}^k_S(u_T, E)$ is uniquely determined for any connected coalition T by Proposition B.8. Thus, $\mathcal{I}^k$ is uniquely determined for any interaction-restricted function $f|_E$. This completes the proof. □

C. Additional Results

C.1. Fairness of the Myerson-Taylor index

We extend the Myerson value to high-order interaction using four axioms as in (Selçuk & Suzuki, 2014) as they align with the analyses for the Shapley-Taylor index (Sundararajan et al., 2020). In what follows, we adapt the classical fairness axiom (Myerson, 1977) to higher-order interactions and show that the Myerson-Taylor index complies with this extended fairness criterion.

The following axiom extends the classical fairness axiom of the Myerson value (Myerson, 1977).

Property 6 (Fairness - **F**). *A k-order interaction index $\mathcal{I}^k$ is fair if, for any graph (V, E), characteristic function f, and $ij \in E$, we have*

$$\mathcal{I}^k_{S \cup i}(f, E) - \mathcal{I}^k_{S \cup i}(f, E \setminus ij) = \mathcal{I}^k_{S \cup j}(f, E) - \mathcal{I}^k_{S \cup j}(f, E \setminus ij),$$

for all $S \subseteq V \setminus ij$ such that $|S| \leq k - 1$.

Proposition C.1. *The Myerson interaction index satisfies the fairness properties.*

Before going to the proof of Proposition C.1, we present a generalized property from the equal treatment of equals in (Béal & Navarro, 2020) to interaction indices.

Property 7 (Equal treatment of equals). *For any function f any two equal nodes i, j such that $f(T \cup i) = f(T \cup j), \forall T \subseteq V \setminus ij$, we then have $\mathcal{I}^k_{S \cup i}(f) = \mathcal{I}^k_{S \cup j}(f), \forall S \subseteq V \setminus ij, |S| \leq k - 1$.*

It is known that equal treatment of equals is weaker than symmetry: any allocation rule satisfying symmetry also satisfies equal treatment of equals, while the reverse does not necessarily apply (Béal & Navarro, 2020).

Lemma C.2. *Symmetry implies equal treatment of equals.*

Proof. Consider a function f and two equal nodes i, j. Let π be a permutation of V such that

$$\pi(l) = \begin{cases} i & \text{if } l = j, \\ j & \text{if } l = i, \\ l & \text{otherwise.} \end{cases}$$

We have $\pi f = f$ as $f(T \cup i) = f(T \cup j)$ for all $T \subseteq V \setminus ij$. By the symmetry axiom, we have $\mathcal{I}^k_{S\cup i}(f) = \mathcal{I}^k_{\pi(S\cup i)}(\pi f) = \mathcal{I}^k_{S\cup j}(f)$, for any $S \subseteq V \setminus ij$. □

We are now ready to prove Proposition C.1. In what follows, we use $\zeta(T, E \setminus ij)$ to denote a set of connected components of the subgraph of $(V, E \setminus ij)$ induced by a subset T.

Proof of Proposition C.1. Consider any graph (V, E), a function f, and any edge $ij \in E$, we define a characteristic function $g = f|_E - f|_{E\setminus ij}$. For any subset $T \subseteq V$, we have

$$g(T) = \sum_{R\in\zeta(T,E)} f(R) - \sum_{R\in\zeta(T,E\setminus ij)} f(R).$$

Thus $g(T) = 0$ if $\{i, j\} \not\subseteq T$ as removing the edge ij does not affect the components in T. Consequently, for any $T \subseteq V \setminus ij$, $g(T \cup i) = g(T \cup j) = 0$.

In other words, i and j are equals in the game g. According to the equal treatment of equals property (Lemma C.2), we have $\Phi^k_{S\cup i}(g) = \Phi^k_{S\cup j}(g), \forall S \subseteq V \setminus ij, |S| \leq k - 1$.

Using the linearity axiom and replacing the Shapley-Taylor index with the Myerson interaction index, we deduce

$$\Psi^k_{S\cup i}(f, E) - \Psi^k_{S\cup i}(f, E \setminus ij) = \Psi^k_{S\cup j}(f, E) - \Psi^k_{S\cup j}(f, E \setminus ij).$$

This completes the proof. □

C.2. Reduction

The following result reduces the high-order interaction Shapley-Taylor into the classical Shapley value.

Proposition C.3 (Reduction)**.** *Let Φ^k is the k-order Shapley-Taylor interaction and ϕ is the Shapley value, for any function f and a player i, we have*

$$\phi_i(f) = \sum_{\substack{S\subseteq V\setminus i\\ |S\cup i|\leq k}} \frac{1}{|S \cup i|}\Phi^k_{S\cup i}(f).$$

Proof of Proposition C.3. We show Propsition C.3 for any unanimity functions u_T, $T \subseteq V$. The proof for a general function f follows by applying linear axiom (**L**) to the Shapley value and the Shapley-Taylor interaction index.

We first consider the case that $i \notin T$, thus, by the dummy axiom (**D**), we have $\phi_i(u_T) = \Phi^k_{S\cup i}(u_T) = 0$, for any $S \subseteq V \setminus i$.

If $i \in T$, by the dummy axiom, we have

$$\begin{aligned}\sum_{\substack{S\subseteq V\setminus i\\ |S\cup i|\leq k}} \frac{1}{|S \cup i|}\Phi^k_{S\cup i}(u_T) &= \sum_{\substack{S\subseteq T\setminus i\\ |S\cup i|= k}} \frac{1}{k}\Phi^k_{S\cup i}(u_T)\\ &= \sum_{\substack{S\subseteq T\setminus i\\ |S\cup i|= k}} \frac{1}{k}\frac{1}{\binom{t}{k}}\\ &= \frac{1}{k}\frac{\binom{t-1}{k-1}}{\binom{t}{k}} = \frac{1}{t}.\end{aligned}$$

The first equation follows by the **ID** axiom. The second equation follows by Sundararajan et al. (2020, Proposition 4).

It is known that $\phi_i(T) = \frac{1}{t}, \forall i \in T$ (Algaba et al., 2019). We deduce that $\phi_i(f) = \sum_{\substack{S\subseteq V\setminus i\\ |S\cup i|\leq k}} \frac{1}{|S\cup i|}\Phi^k_{S\cup i}(f)$, for $i \in T$.

This completes the proof. □

Proposition C.3 indicates that the Shapley-Taylor index distributes the importance score of i, which is the Shapley value of i, to interactions up to size k that node i can join.

Since the Myerson value is a generalization of the Shapley value and the Myerson-Taylor index is a generalization of the Shapley-Taylor index, the reduction in Proposition C.3 holds for the Myerson values and the Myerson-Taylor index.

D. Implementation Details

D.1. Myerson-Taylor Interaction Index

We present a permutation-based sampling algorithm to compute the Myerson-Taylor index for a given graph input and black-box model in Algorithm 1. The algorithm leverages the principle that the Shapley-Taylor index represents the expected value of discrete derivatives over a randomly chosen ordering of nodes in V. The main difference compared to the Shapley-Taylor is that the Myerson-Taylor uses the interaction-restricted function $f|_E$ (Line 8), instead of the vanilla f. Detailed implementation to compute the interaction-restricted value for a coalition T is provided in Algorithm 2.

Algorithm 1 Permutation-based sampling algorithm for the k-order Myerson-Taylor index.

Input : a graph $G = (V, E)$, a value function $f : 2^{|V|} \to \mathbb{R}$, order k
Output : interaction index $\mathbf{B}$

1 $\mathbf{A} \leftarrow \mathbf{0}$; /* Accumulated interactions */
2 $\mathbf{C} \leftarrow \mathbf{0}$; /* Count interactions */
3 **for** $t = 0, 1, \ldots$ **do**
4 $\quad \pi \leftarrow$ a random ordering of $\{1, 2, \ldots, |V|\}$ **for** *all subset* $S \subseteq V$ *with size* l, $|S| = l$ **do**
5 $\quad\quad i \leftarrow$ left most index of S's elements in the ordering π $T \leftarrow \{\pi_1, \ldots, \pi_{i-1}\}$; /* A set of predecessors of S in π */
6 $\quad\quad \mathbf{A}_S \leftarrow \mathbf{A}_S + \delta_S f|_E(T)$; /* $f|_E(T)$ is computed by Algorithm 2 */
7 $\quad\quad \mathbf{C}_S \leftarrow \mathbf{C}_S + 1$
8 $\quad$ **end**
9 **end**
10 $\mathbf{B} \leftarrow \mathbf{0}$ **for** *every subset* $S \subseteq V$ *up to size* k, $|S| \leq k$ **do**
11 $\quad$ **if** $|S| \leq k$ **then**
12 $\quad\quad \mathbf{B}_S \leftarrow \delta_S f|_E(\emptyset)$
13 $\quad$ **else**
14 $\quad\quad \mathbf{B}_S \leftarrow \mathbf{A}_S / \mathbf{C}_S$
15 $\quad$ **end**
16 **end**

Algorithm 2 Value of an interaction-restricted function ($f|_E(T)$).

Input : a graph $G = (V, E)$, a value function $f : 2^{|V|} \to \mathbb{R}$, a coalition T
Output : A value $v \in \mathbb{R}$

17 $\zeta(T) \leftarrow$ a set of components of subgraph (T, E_T) $v \leftarrow 0$ **for** *every component* C *in* $\zeta(T)$ **do**
18 $\quad v \leftarrow v + f(C)$
19 **end**

D.2. Motif Search

We provide a linear relaxation approach to solve the problem (3), which can be explicitly rewritten as follows

$$\max_{S_1,\ldots,S_m \subseteq V} \sum_{l=1}^{m} \left| \sum_{\substack{i,j \in S_l \\ i \leq j}} \tau \mathbf{B}_{ij}^{+} + (1-\tau)\mathbf{B}_{ij}^{-} \right| \tag{5}$$

$$\text{s.t. } S_l \cap S_h = \emptyset \qquad 1 \leq l, h \leq m \tag{6}$$

$$|\cup_{l=1}^{m} S_l| \leq M \tag{7}$$

$$(S_l, E_{S_l}) \text{ is connected} \qquad 1 \leq l \leq m \tag{8}$$

The above optimization problem is a variant of the quadratic multiple knapsack problem (Hiley & Julstrom, 2006) with absolute values. One strategy to solve (3) is by linear relaxations and then using off-the-shelf MILP solvers such as MOSEK (ApS, 2019) or GUROBI (Gurobi Optimization, LLC, 2023). The absolute operator in the objective can be cast to linear constraints using the big-M method, a widely used technique in integer programming to linearize constraints with absolute values. To capture the connectivity constraints (8), we employ the *linear connectivity constraints* as in (Mak-Hau & Yearwood, 2019; Althaus et al., 2014). The linear connectivity constraints are initially relaxed and added in a *lazy* fashion whenever the incumbent optimal solution violates them. Problem (3) also exhibits symmetry in the solution (shuffling the set indices returns the same solution), and we break this symmetry using aggressive symmetry-breaking constraints of the MILP solvers.

E. Experiments

E.1. Experimental Details

E.1.1. Datasets

We use the popular datasets in the literature of GNN explanability (Sato, 2020; Yuan et al., 2022; Agarwal et al., 2023).

- *BA-2Motifs (Luo et al., 2020)*: The dataset is a binary classification task where each graph incorporates a Barabasi-Albert base structure linked with either a house or five-cycle motif. The label and ground-truth explanation of a graph is determined by the motif the graph contains.
- *SPMotif (Wu et al., 2022)*: The dataset contains graphs combining a base structure (Tree, Ladder, or Wheel) and a motif (Cycle, House, Crane). A spurious correlation between the base and the motif is manually injected into each graph. A graph's label and ground truth explanation is determined based on the motif it contains.
- *BA-HouseGrid*: Similar to BA-2Motifs but with two distinct motifs, house and 3×3 grid. The house and grid motif are chosen because they do not have overlapping structures such as those found in the house and five-cycle.
- *BA-HouseAndGrid*: Each graph is a Barabási structure that may linked with either house or grid motifs. Graphs containing both motif types are labeled as 1, otherwise 0.
- *BA-HouseOrGrid*: Similar to BA-HouseAndGrid, however, graphs with either house or grid motifs are labeled as 1, otherwise 0.
- *Mutagenic (Kazius et al., 2005)*: The dataset is a molecular property prediction task, which is to identify if a molecule is mutagenic or not. The functional groups -NO2 and -NH2 are considered as ground-truth explanations that lead to mutagenicity (Luo et al., 2020).
- *Benzene (Sanchez-Lengeling et al., 2020)*: The dataset contains 12000 molecular graphs extracted from ZINC15 (Sterling & Irwin, 2015). The task is to determine the presence of benzene rings in a molecule. Ground-truth explanations are the carbon atoms in the benzene rings.
- *MNIST75SP (Monti et al., 2017)*: An image classification dataset where each image in MNIST is converted to a superpixel graph. Each node represents a superpixel where node features are the superpixel's central coordinate and

Table 4. Statistics and properties of datasets. The datasets above the dashed blue line are synthetic, and below the dashed blue line are real-world ones.

Dataset	no. graphs	no. classes	no. nodes	no. edges
BA-2Motifs	1000	2	25.00	25.48
BA-HouseGrid	10000	2	26.97	28.95
SPMotif	18000	3	45.98	66.72
BA-HouseAndGrid	10000	2	29.42	41.01
BA-HouseOrGrid	10000	2	24.73	34.31
Mutagenic	2951	2	30.13	30.45
Benzene	12000	2	20.58	21.83
MNIST75SP	70000	10	70.57	295.25
GraphSST2	70042	2	10.20	9.20
Twitter	6940	3	21.10	20.10

Table 5. Accuracy of GNNs models on evaluated datasets. The datasets above the dashed blue line are synthetic, and below the dashed blue line are real-world ones.

Dataset	GCN	GIN	GAT
BA-2Motifs	97.00	100.00	41.00
BA-HouseGrid	100.00	99.90	52.30
SPMotif	86.00	96.17	50.11
BA-HouseAndGrid	99.30	99.90	51.10
BA-HouseOrGrid	100.00	100.00	51.10
Mutagenic	89.86	91.55	92.57
Benzene	85.42	91.92	85.92
MNIST75SP	82.44	89.43	94.64
GraphSST2	89.13	87.86	89.24
Twitter	69.22	63.87	67.34

brightness intensity. The spatial proximity between the superpixels determines the edges. Graph-truth explanations are the top 15 superpixels with the highest intensity.

- *GraphSST2 and Twitter (Yuan et al., 2022)*: The datasets are sentiment classification tasks. GraphSST2 has two classes, and Twitter has three classes. Each node corresponds to one word in the text, edges are constructed by the Biaffine parser (Gardner et al., 2018), and node features are pre-trained BERT embedding of words. No ground-truth explanations are available for these datasets.

Table 4 provides statistics for chosen datasets.

E.1.2. Models

In this paper, we use three GNN models commonly used in explainability literature: GCN (Kipf & Welling, 2016), GIN (Xu et al., 2018), and GAT (Veličković et al., 2017). The accuracy of these models is provided in Table 5. As SubgraphX released the checkpoint for GCN for BA-2Motifs, we used their checkpoint. As GAT performs poorly on synthetic datasets, we only explain for GAT on real-world datasets. Note that GAT's poor performance on synthetic datasets is also reported in (Amara et al., 2023, Table 3) and (Zheng et al., 2024, Table 2).

E.1.3. Metrics

For a GNN model f and input graph $G = (V, E)$, the output of explainers is a subgraph (S, E_S). Let (S^{gt}, E^{gt}) be the ground truth explanation. Following previous practice in (Ying et al., 2019; Luo et al., 2020), we use the below metrics for datasets with ground truth explanations:

- *Area Under the ROC Curve (AUC)*: To measure AUC metrics, we treat the explanation task as the binary classification task on the edges of the input graph. We then compute the explanation accuracy by comparing binary edge masks generated by explainers E_S against the ground truth edge masks corresponding to E^{gt}.
- *F1 score*: The F1 score reports the overlap of the nodes highlighted by the explainers S compared to the ground truth explanation S^{gt}. We use the F1 score for datasets where graphs contain only a single motif.
- *Adjusted Mutual Information (AMI)*: AMI score is a common metric for evaluating different clustering algorithms. We use the AMI score to measure explainers' ability to identify different explanatory substructures. Specifically, for an

Table 6. Explanation accuracy for the GIN model on datasets with a single motif. Note that GradCAM is a gradient-based method; GNNExplainer, PGExplainer, Refine, and MatchExplainer are perturbation-based; The last four methods are cooperative game-based.

Method	BA-2Motifs		BA-HouseGrid		SPMotif		MNIST75SP	
	F1 ↑	*AUC*↑	*F1* ↑	*AUC*↑	*F1* ↑	*AUC*↑	*F1* ↑	*AUC*↑
GradCAM	**0.934**	**0.988**	0.876	**0.978**	**0.776**	**0.887**	0.230	0.486
GNNExplainer	0.280	0.486	0.360	0.508	0.176	0.491	0.257	0.525
PGExplainer	0.314	0.363	0.449	0.491	0.429	0.415	0.202	0.484
Refine	0.116	0.427	0.231	0.506	0.153	0.475	0.194	0.543
MatchExplainer	0.672	0.765	0.482	0.635	0.189	0.494	0.196	0.499
SubgraphX	0.863	0.901	0.862	0.839	0.484	0.673	0.205	0.501
GStarX	0.185	0.484	0.268	0.544	0.170	0.486	0.307	0.532
SAME	0.909	0.914	0.846	0.827	0.347	0.607	0.275	0.525
MAGE (ours)	**0.940**	0.942	**0.954**	0.942	0.630	0.729	**0.586**	**0.675**

Table 7. Explanation accuracy for the GCN model on datasets with multiple motifs.

Method	BA-HouseAndGrid		BA-HouseOrGrid		Mutagenic		Benzene	
	AMI↑	*AUC*↑	*AMI*↑	*AUC*↑	*AMI*↑	*AUC*↑	*AMI*↑	*AUC*↑
GradCAM	0.570	0.776	0.980	**0.998**	0.423	0.887	**0.669**	**0.928**
GNNExplainer	0.242	0.491	0.184	0.485	0.151	0.578	0.173	0.492
PGExplainer	0.483	0.699	0.428	0.541	0.689	0.949	0.216	0.056
Refine	0.241	0.457	0.114	0.452	0.463	0.876	0.199	0.618
MatchExplainer	0.612	0.842	0.606	0.818	0.231	0.601	0.165	0.508
SubgraphX	0.532	0.714	0.533	0.779	0.672	0.840	0.326	0.643
GStarX	0.201	0.497	0.130	0.481	0.003	0.472	0.049	0.515
SAME	0.528	0.714	0.631	0.813	0.787	0.895	0.109	0.536
MAGE (ours)	**0.675**	**0.879**	**0.999**	**0.999**	**0.987**	**0.993**	0.584	0.784

explanation $S = (S_1, \ldots, S_m)$ and ground truth explanation $S^{gt} = (S_1^{gt}, \ldots S_l^{gt})$, the mutual information between two partitions S and S^{gt} is calculated by

$$\mathrm{MI}(S, S^{gt}) = \sum_{i=1}^{m} \sum_{j=1}^{l} p(i,j) \log \frac{p(i,j)}{p(i)p'(j)},$$

where $p(i) = \frac{|S_i|}{|V|}$ is the probability that a node picked at random from explanation S falls into a motif S_i. Similarity, we have $p'(j) = \frac{|S_j^{gt}|}{|V|}$ and $p(i,j) = \frac{|S_i \cap S_j^{gt}|}{|V|}$. The vanilla MI tends to favor partitions with a higher number of clusters, regardless of the actual amount of 'mutual information' between the label assignments, *adjusted* MI (AMI) is used to alleviate this bias (Vinh et al., 2009). Even though most of the current explainers do not consider the multiple motif setting, they might highlight multiple disconnected components. For these baselines, we consider each connected component in the highlighted subgraph as one identified motif.

For datasets without ground truth explanations, we use the following metrics:

- *Fidelity* (Fid) *(Pope et al., 2019; Yuan et al., 2021)*: Fidelity measures the faithfulness of an explanation by reporting the changes in the model output when removing or keeping only selected nodes S

$$\begin{aligned} \mathrm{Fid}^{+}(S) &= f(V) - f(V \setminus S), \\ \mathrm{Fid}^{-}(S) &= f(V) - f(S), \\ \mathrm{Fid}(S) &= \mathrm{Fid}^{+}(S) - \mathrm{Fid}^{-}(S). \end{aligned}$$

  The higher fidelity means that S is more important to the model prediction.

- *Robust Fidelity* (Fid_α): Since it is known that Fid is sensitive to the OOD samples, we report Fid_α, a new metric proposed in (Zheng et al., 2023) to alleviate the OOD problem of Fid

$$\begin{aligned} \mathrm{Fid}_\alpha^{+}(S) &= f(V) - \mathbb{E} f(V \setminus \Omega^{\alpha}(S)), \\ \mathrm{Fid}_\alpha^{-}(S) &= f(V) - \mathbb{E} f(S \cup \Omega^{1-\alpha}(V \setminus S)), \\ \mathrm{Fid}_\alpha(S) &= \mathrm{Fid}_\alpha^{+}(S) - \mathrm{Fid}_\alpha^{-}(S), \end{aligned}$$

Table 8. Results on sentiment classification tasks. Note that GradCAM is a gradient-based method, while the other methods are cooperative game-based.

(a) GIN

Method	GraphSST2		Twitter	
	$\text{Fid}_\alpha \uparrow$	$\text{Fid} \uparrow$	$\text{Fid}_\alpha \uparrow$	$\text{Fid} \uparrow$
GradCAM	0.140	0.259	0.251	0.412
SubgraphX	0.147	**0.307**	0.239	0.398
GStarX	0.132	0.264	0.243	0.403
SAME	0.149	0.281	0.301	0.475
SAGE (ours)	**0.172**	**0.309**	**0.374**	**0.590**

(b) GAT

Method	GraphSST2		Twitter	
	$\text{Fid}_\alpha \uparrow$	$\text{Fid} \uparrow$	$\text{Fid}_\alpha \uparrow$	$\text{Fid} \uparrow$
GradCAM	–	–	–	–
SubgraphX	0.112	0.192	0.086	0.129
GStarX	**0.141**	0.250	0.093	0.147
SAME	0.123	0.203	**0.106**	0.158
SAGE (ours)	**0.148**	**0.266**	**0.112**	**0.176**

where $\Omega^\alpha(T), 0 \leq \alpha \leq 1$ is a random subset of T where a node in T is included with probability α and erased with probability $1 - \alpha$. The key idea of Fid_α is that if a subset S is important to the model prediction, removing or keeping a superset of S should also change the model output significantly. In case $\alpha = 1$, Fid_α coincides with the original Fid metric. In this paper, we set $\alpha = 0.8$.

E.1.4. Experimental Setup and Implementation

We mainly follow the experimental settings as in (Yuan et al., 2021; 2022), where we leverage their codebase[3] for GNN models and baseline explainers. We use the default hyperparameters for the baselines as in (Yuan et al., 2022).

Following (Zheng et al., 2023), we only explain for well-trained models with reasonable performance and graph instances that the GNN models correctly predict. We split the dataset into training, validation, and test subsets with respective ratios of 0.8, 0.1, and 0.1. We train GNN models to a reasonable performance and then run the explainers for graph instances in the test datasets. We report the average metrics over instances in the test dataset. For synthetic and molecular datasets, we explain all instances in the test set with ground truth explanations. For MNIST75SP, GraphSST2, and Twitter, we randomly select 200 instances for evaluation.

E.2. Remaining Results

E.2.1. Quantitative Results

We provide the remaining results for other combinations of GNN models and datasets in Table 6, 7, and 8. In terms of explanation accuracy (Table 6 and Table 7), MAGE outperforms other cooperative game-based explainers in most settings while providing competitive performance with GradCAM, a while-box explainer.

E.2.2. Ablation Study

Ablation for the group attribution. This experiment validates that higher-order interactions can better approximate the group attribution than node-wise values such as Shapley and Myerson values.

For a given group S, the Shapley (ϕ) or Myerson (ψ) values estimate the contribution of a group S to the model prediction by the sum of node-wise importance

$$\text{GrAttr}(\psi, S) = \sum_{i \in S} \psi_i.$$

The change in the model prediction in the absence of the group S is then estimated by

$$\begin{aligned} \text{GrAttr}(\psi, \bar{S}) &= \text{GrAttr}(\psi, V) - \text{GrAttr}(\psi, V \setminus S) \\ &= \text{GrAttr}(\psi, S). \end{aligned}$$

Here, we can see that $\text{GrAttr}(\psi, \bar{S}) = \text{GrAttr}(\psi, S)$ when we use node-wise importance to approximate the group attribution.

Meanwhile, we can estimate the group attribution of S using the second-order Shapley-Taylor (Φ) or Myerson-Taylor (Ψ)

[3] https://github.com/divelab/DIG/tree/dig-stable/benchmarks/xgraph

index as follows

$$\mathrm{GrAttr}(\Psi, S) = \sum_{i,j \in S} \Psi_{ij}$$
$$\mathrm{GrAttr}(\Psi, \bar{S}) = \mathrm{GrAttr}(\Psi, V) - \mathrm{GrAttr}(\Psi, S).$$

To show that second-order interaction indices provide a better approximation of the group attribution, we conduct an experiment on image classification tasks where the Shapley values were more frequently used. Specifically, we use ResNet50 (He et al., 2016) and ViT-B/16 (Dosovitskiy et al., 2020) models pre-trained on Imagenet. We then compute four attribution methods, including the Shapley values, Myerson values, 2nd-order Shapley-Taylor index, and 2nd-order Myerson-Taylor index for 50 representative images provided by SHAP (Lundberg & Lee, 2017). Each image is segmented into 49 (7×7) patches to compute the Shapley values and Shapley-Taylor indices. We build a grid graph on 49 patches as the interaction-restricted function for the Myerson value and Myerson-Taylor index. Two patches are connected if they are spatially adjacent.

We measure the faithfulness of the group attributions of four methods using infidelity metric (Yeh et al., 2019), which evaluates the difference between the estimated group attribution and the model's prediction in the presence (Infid^+) and absence of the group (Infid^-)

$$\mathrm{Infid}^+(\mathcal{I}^k) = \mathbb{E}_S \left| \mathrm{GrAttr}(\mathcal{I}^k, S) - (f(S) - f(\emptyset)) \right|,$$
$$\mathrm{Infid}^-(\mathcal{I}^k) = \mathbb{E}_S \left| \mathrm{GrAttr}(\mathcal{I}^k, \bar{S}) - [f(V) - f(V \setminus S)] \right|.$$

The lower infidelity indicates that the group attribution method more accurately captures potential shifts in the model's predictions in the presence or absence of a group.

Table 9. Ablation study for computing group attribution on Imagenet. The last column is the number of model queries needed to compute the attributions.

Method	ResNet50		ViT-B/16		
	$\mathrm{Infid}_+(\downarrow)$	$\mathrm{Infid}_-(\downarrow)$	$\mathrm{Infid}_+(\downarrow)$	$\mathrm{Infid}_-(\downarrow)$	#queries
Shapley (ϕ)	0.083	0.123	0.058	0.065	18K
Myerson (ψ)	0.083	0.123	0.057	**0.064**	13K
Shapley-Taylor (Φ^2)	0.060	0.110	0.062	0.105	57K
Myerson-Taylor (Ψ^2)	**0.056**	**0.108**	**0.032**	0.075	19K

Table 9 shows that second-order indices (Shapley-Taylor and Myerson-Taylor) provide a better approximation for group attributions than using (first-order) marginal contributions (Shapley and Myerson). Notably, Myerson-based attributions improve the estimation of group attributions with fewer queries on the black-box model.

Ablating the interaction indices by edge-based explainers. To investigate the effectiveness of the interaction indices compared to simple edge-based explainers that assign an importance score for every edge in the graph, we ablate the interaction indices with edge-based explainers in our MAGE framework. Here, the edge-based explainer could act as the interaction matrix $\mathbf{B}$ where the interactions between two nodes without a direct connection will be zero. We apply our motif search component directly upon this interaction matrix to find explanatory motifs.

We conduct experiments with two edge-based explainers: GNNExplainer and EdgeShaper, and their combinations with our motif search module. GNNExplainer (Ying et al., 2019) is not a game-based method, so it may not have theoretical properties like Shapley values. Therefore, we also apply EdgeShaper (Mastropietro et al., 2022), which applies Shapley values on edges (consider an edge as a player) to compute importance scores for edges. EdgeShaper is an edge-wise importance method based on Shapley value, thus has Shapley's properties. However, it does not satisfy the interaction distribution (ID) axiom, which supports the Shapley-Taylor and Myerson-Taylor indices.

The result shown in table 10 shows the enhanced performance of edge-based explanations when augmented with the motif search module in most of the settings. However, this is not always the case since we can also observe a drop from EdgeShaper + Motif Search compared to EdgeShaper. This drop may be because EdgeShaper does not align with the definition of the group attribution in the motif search module since EdgeShaper (and other edge-based explainers) assume that two nodes without a direct edge do interact with each other and will have a zero interaction score. However, Myerson-Taylor

Table 10. Performance of MAGE when ablating the interaction indices by edge-based explainers.

Method	BA-2Motifs		BA-HouseGrid		Mutagenic		Benzene		GraphSST2	
	F1	AUC	F1	AUC	AMI	AUC	AMI	AUC	Fid_α	Fid
GNNExplainer	0.22	0.44	0.30	0.55	0.23	0.68	0.18	0.49	-0.02	0.07
GNNExplainer + Motif Search	0.43	0.64	0.60	0.75	0.06	0.46	0.27	0.51	0.08	0.15
EdgeShaper	0.48	0.72	0.49	0.65	0.70	0.86	0.37	0.70	-0.11	0.03
EdgeShaper + Motif Search	0.55	0.71	0.67	0.77	0.59	0.79	0.40	0.71	-0.13	0.01
Myerson-Taylor + Motif search (**MAGE**)	**0.86**	**0.89**	**0.83**	**0.85**	**1.00**	**1.00**	**0.92**	**0.96**	**0.20**	**0.34**

may attribute a positive (or negative) interaction score to connected nodes from multiple hops away. Thanks to that, our framework (Myerson-Taylor + Motif Search) outperforms other ablated methods by a considerable margin.

Ablation on the higher-order interaction indices. In the main paper, we use the second-order interaction index mainly because of its efficacy and efficiency. Using a higher-order interaction index ($k > 2$) will require more computation cost in approximating the interaction index and solving the motif search problem. For example, if we use $k = 3$, the motif search problem would become a cubic program, which would be more difficult to solve. For demonstration purposes, we conduct an ablation study using a third-order interaction index ($k = 3$) and apply linear relaxation to solve the motif search. The result is reported for BA-2Motifs in Table 11.

Table 11. Comparison of higher-order Shapley-Taylor and Myerson-Taylor.

	F1	AUC	$\#Q$	Running Time (s)
MAGE (Shapley-Taylor $k = 2$)	0.709	0.787	51K	11.4
MAGE (Myerson-Taylor $k = 2$)	0.86	0.89	7K	5.1
MAGE (Shapley-Taylor $k = 3$)	0.92	0.93	343K	104.2
MAGE (Myerson-Taylor $k = 3$)	0.96	0.97	25K	56.8

We observe that third-order interaction indices increase the performance on the BA-2Motifs dataset compared to second-order interaction indices. However, as a trade-off, the number of model queries and the running time increase significantly.

E.2.3. Complexity Analysis

Running time analysis. Table 12 shows the running time for our method and competing methods on evaluated datasets. Note that due to the additional training stages or the necessity for access to the model's gradient or training data, GradCAM, PGExplainer, Refine, and MatchExplainer are excluded from the direct comparison. Therefore, our comparisons are focused on GNN explainers and other cooperative game-based methods. Significantly, MAGE exhibits the most competitive running time among cooperative game-based explainers while showing superior performance.

Table 12. The average running time of competing methods on evaluated datasets (second/sample).

Method	BA-2Motifs	BA-HouseGrid	SPMotif	MNIST75SP	BA-HouseAndGrid	BA-HouseOrGrid	Mutagenic	Benzene
GNNExplainer	2.27	2.30	3.50	15.83	2.62	3.41	3.78	4.91
SubgraphX	66.60	54.40	74.14	823.55	70.18	81.41	63.63	4.40
GStarX	19.63	15.53	21.51	35.75	29.47	17.06	16.43	20.02
SAME	11.61	44.42	19.03	142.11	90.73	41.27	6.12	7.14
MAGE (ours)	5.16	6.18	14.61	44.28	17.74	9.38	4.89	3.63

Table 13. Model query efficiency across game-based methods.

Method	Single Motif				Sentiment Analysis		Multiple Motifs			
	BA-2Motifs $\#Q\downarrow$	BA-HouseGrid $\#Q\downarrow$	SPMotif $\#Q\downarrow$	MNIST75SP $\#Q\downarrow$	GraphSST2 $\#Q\downarrow$	Twitter $\#Q\downarrow$	BA-HouseAndGrid $\#Q\downarrow$	BA-HouseOrGrid $\#Q\downarrow$	Mutagenic $\#Q\downarrow$	Benzene $\#Q\downarrow$
SubgraphX	333K	340K	375K	2865K	255K	313K	430K	350K	258K	19K
SAME	45K	232K	108K	881K	24K	31K	466K	217K	27K	28K
GStarX	26K	28K	37K	**72K**	18K	21K	34K	30K	28K	23K
MAGE	**7K**	**9K**	**18K**	128K	**2K**	**3K**	**17K**	**15K**	**3K**	**1K**

Analysis of the number of model queries. We report the average number of model queries needed to explain an instance by game-based explainers. Table 13 shows that MAGE requires fewer queries than other game-based methods, especially

when compared with MCTS methods such as SubgraphX and SAME. MAGE demonstrates its superiority by reducing queries by averaged factors of 51.40 (SubgraphX), 14.63 (SAME), and 6.17 (GStarX).

E.2.4. Qualitative Results

For image classification (Figure 5), only MAGE can provide a meaningful explanation that aligns with the brightest superpixels in the input image. Meanwhile, graph explainers fail to provide meaningful explanations for class '8', despite having high fidelity scores. This observation aligns with arguments in (Zheng et al., 2023), suggesting that the fidelity metric is sensitive to out-of-distribution explanations. Note that GradCAM (Pope et al., 2019) is adapted to explain GNN models on the graph inputs constructed from superpixels in the original images, which is different from typical GradCAM (Selvaraju et al., 2017) running on images. We also provide more examples in the GraphSST2 dataset (Figure 6). MAGE can highlight both structures that support and contradict the sentiment predicted by the model prediction. We provide more examples for Mutagentic, Benzene, and synthetic datasets in Figure 7-12.

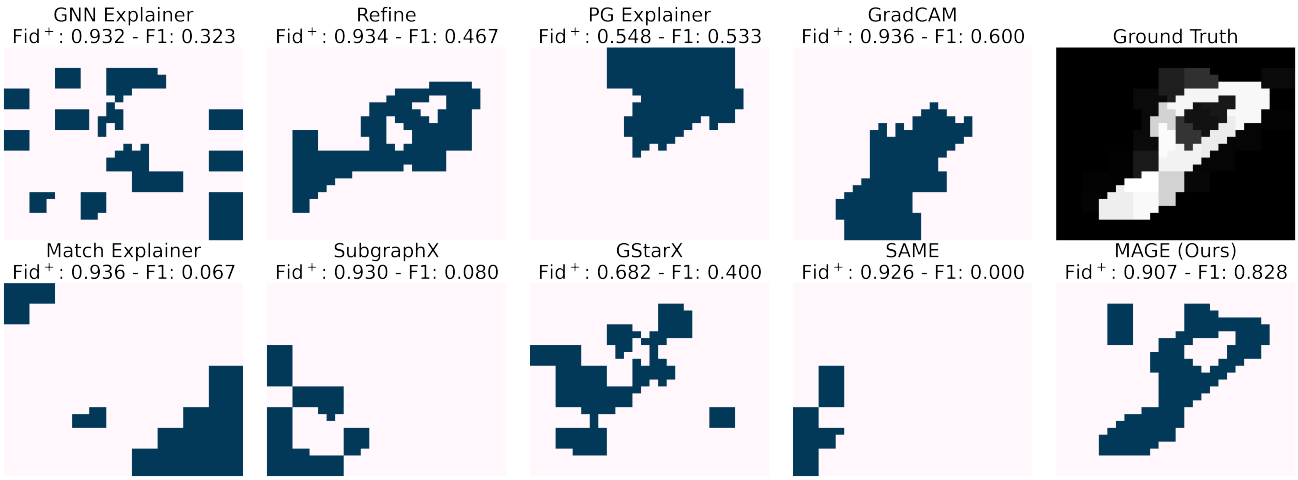
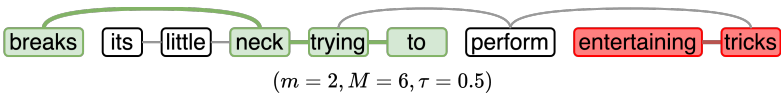
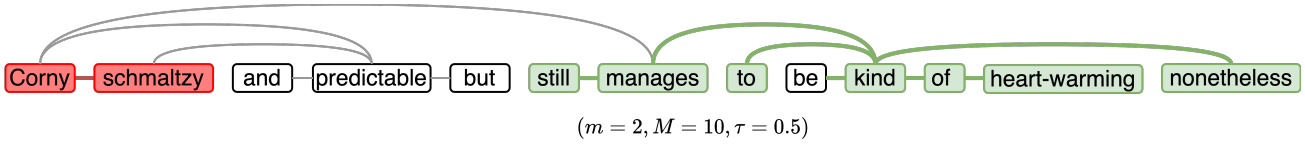


Figure 5. This example visualizes the explanation for the GCN model of MAGE against competing baselines on MNIST75SP. Despite achieving high fidelity (Fid^+) scores, the explanations of baselines are not meaningful. Meanwhile, only MAGE can generate an explanation that aligns with pixels that describe number '8'



(a) Model prediction: Negative sentiment. MAGE highlights the main negative verb phrase 'breaks neck', which contributes to the overall negative sentiment, and the phrase 'entertaining tricks', which shows a slightly positive sentiment.



(b) Model prediction: positive sentiment. MAGE highlights the main adjective, 'heart-warming,' which contributes to the overall positive sentiment, and two minor adjectives, 'corny' and 'schmaltzy,' which display some negative sentiment.

Figure 6. MAGE can highlight text subgraphs with contradicting sentiments in GraphSST2 Dataset.

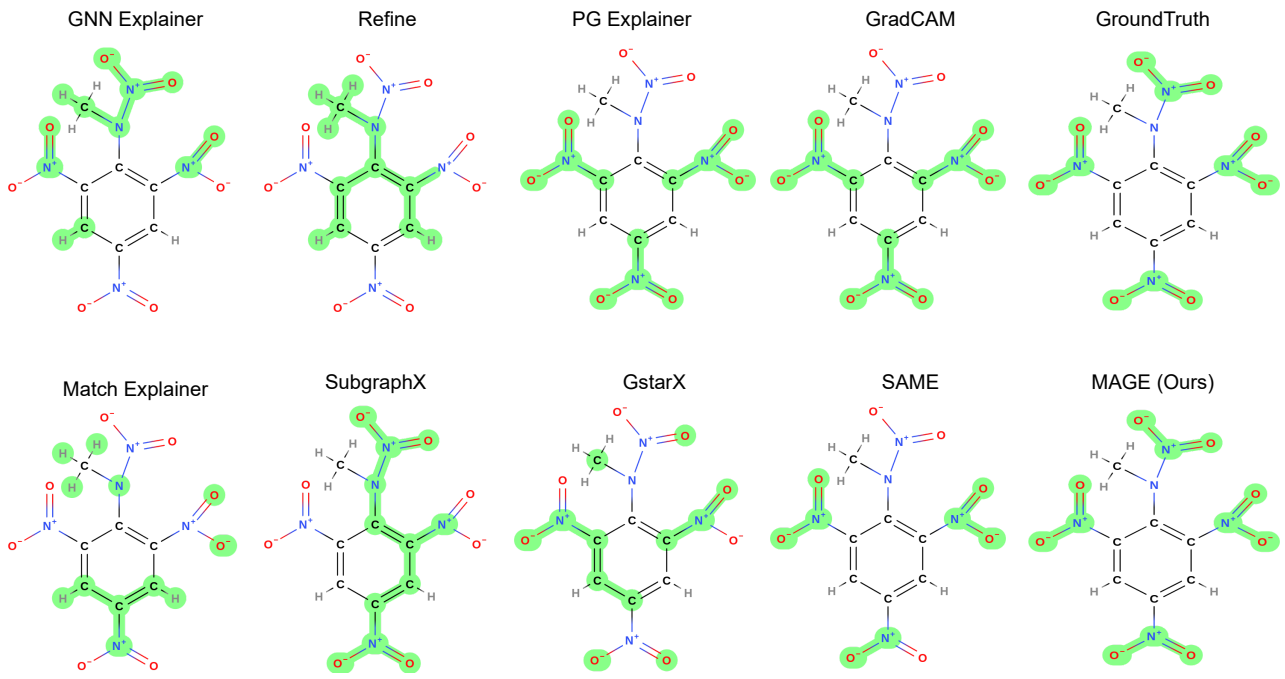


Figure 7. Explanations of competing methods on a molecular graph from Mutagenic dataset.

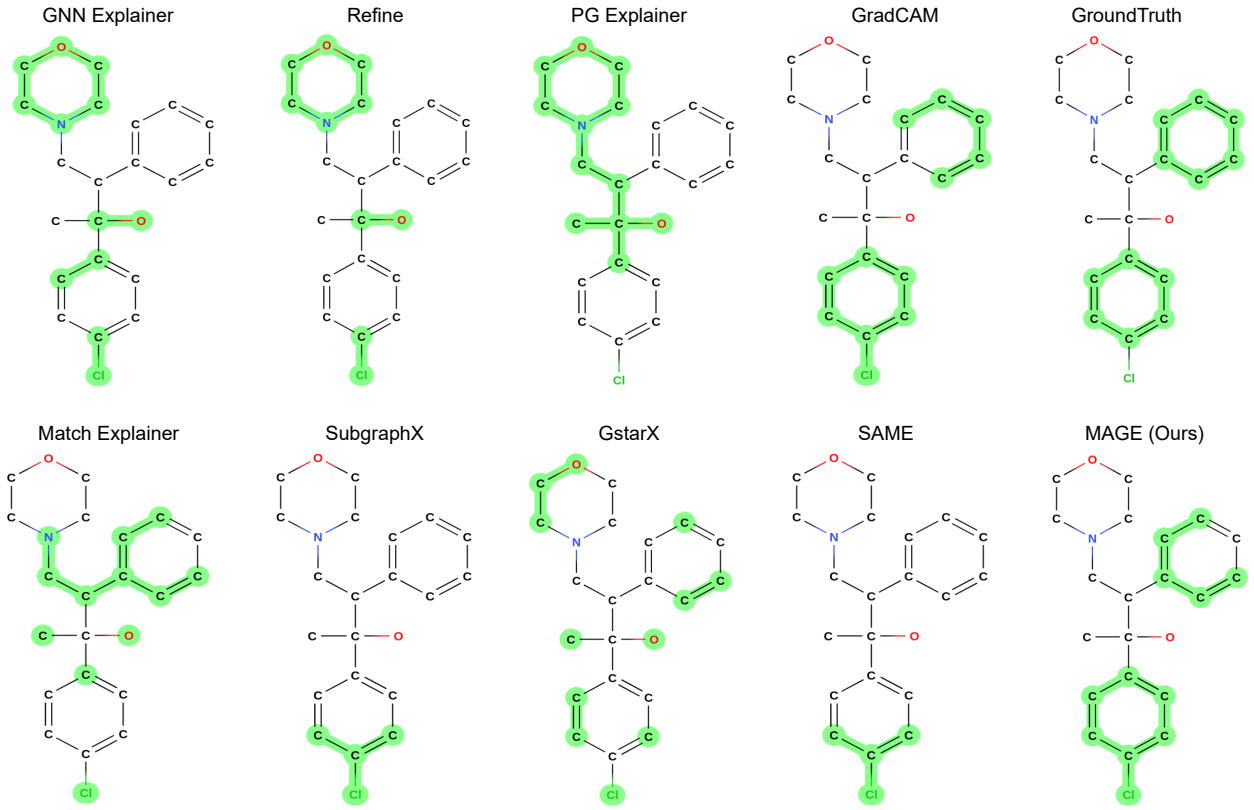


Figure 8. Explanations of competing methods on a molecular graph from Benzene dataset.

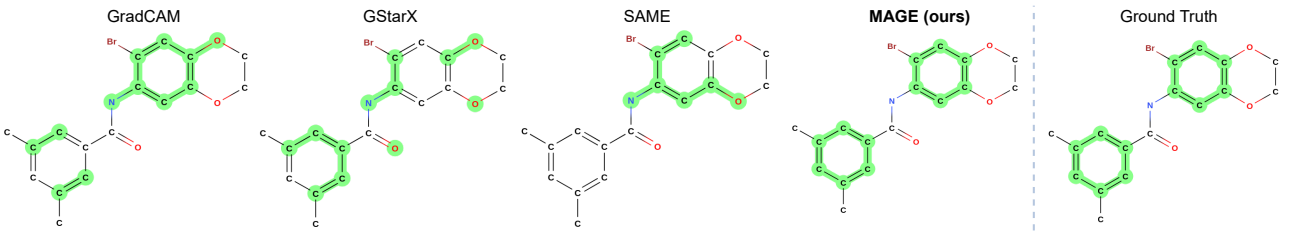


Figure 9. Molecule C17H16BrNO3 input is predicted in class 'have benzene ring' by GNN. Our MAGE multi-motif explanations correctly identify the two benzene rings; while competing methods such as GradCAM, GStarX, and SAME fail.

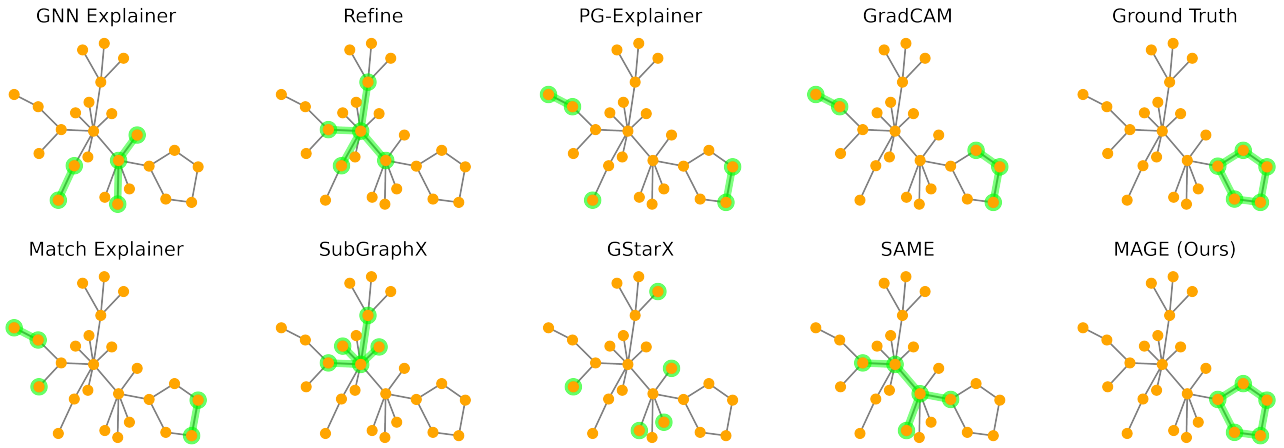


Figure 10. Explanations of competing methods on a synthetic graph from BA-2Motifs dataset.

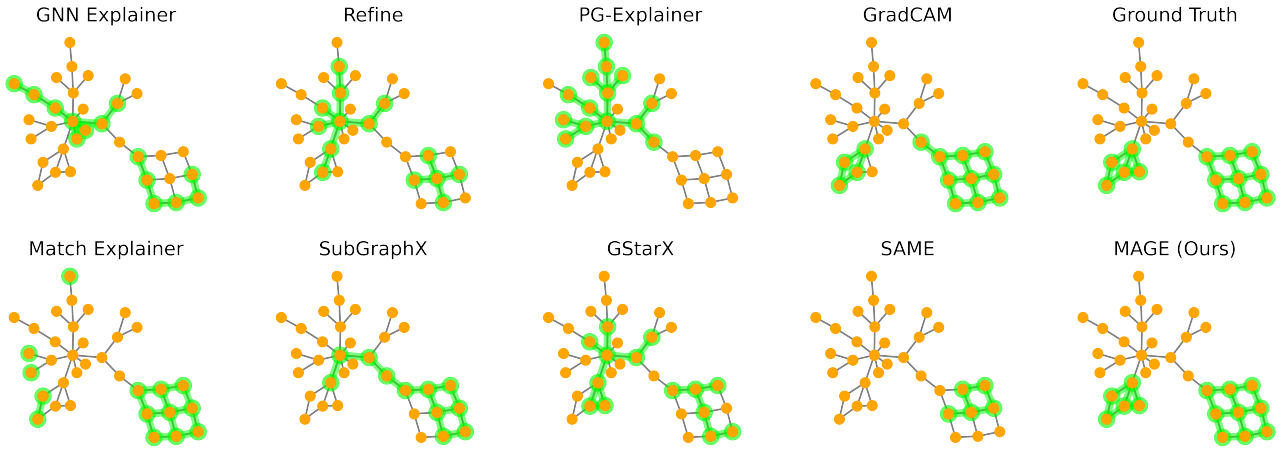


Figure 11. Explanations of competing methods on a synthetic graph from BA-HouseOrGrid dataset.

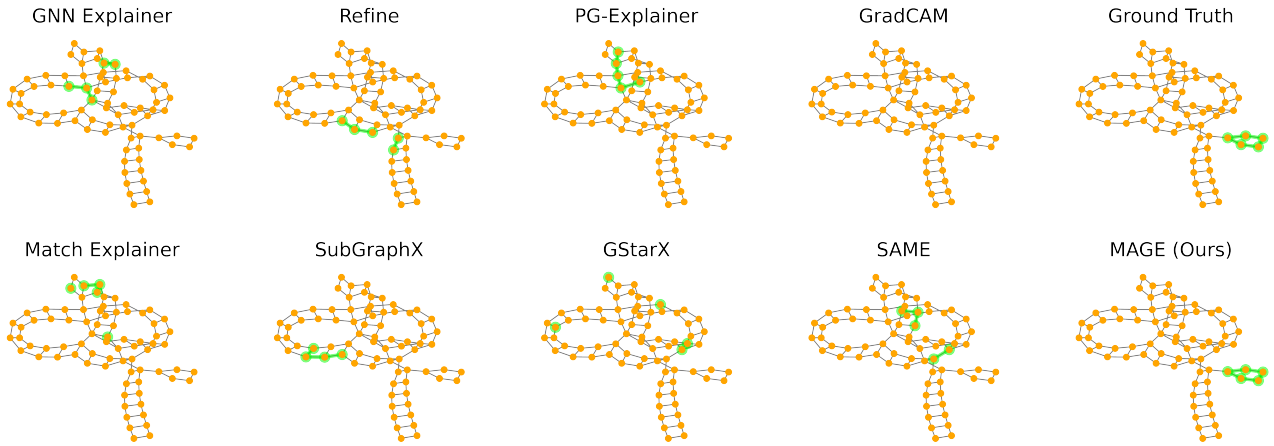


Figure 12. Explanations of competing methods on a synthetic graph from SPMotif dataset.