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

[118] Srijan Kumar, Francesca Spezzano, VS Subrahmanian, and Christos Faloutsos. Edge weight prediction in weighted signed networks. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 221–230. IEEE, 2016.

[119] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, 2013.

[120] Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association For Computational Linguistics (volume 2: Short papers)*, pages 49–54, 2014.

[121] Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3:127–163, 2000.

[122] C Lee Giles, Kurt D Bollacker, and Steve Lawrence. Citeseer: An automatic citation indexing system. In *Proceedings of the third ACM Conference on Digital libraries*, pages 89–98, 1998.

[123] Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1):396–413, 2020.

A Datasets

We summarize synthetic and real-world datasets that are commonly used for evaluating the effectiveness of GNN explanation methods. The details of these datasets are presented in Table 5.

Synthetic Datasets. BA-Shapes [64] dataset consists of a base Barabási-Albert (BA) graph with 300 nodes and 80 "house"-structured network motifs, each consisting of five nodes. These motifs are randomly attached to nodes in the base graph. The nodes in the dataset are classified into four classes based on their structural roles: top, middle, and bottom nodes of the house motifs, and nodes that do not belong to any house.

BA-Community [64] dataset is created by combining two BA-Shapes graphs. The nodes in this dataset have feature vectors that follow a normal distribution. Each node is assigned to one of eight classes based on their structural roles within the house motifs and their community memberships.

BA-2motifs [66] dataset uses BA graphs as the basis. It includes graphs with two types of motifs: "house" motifs and five-node cycle motifs. The dataset is divided into two classes based on the type of motif attached to the graphs. Similarly, BA-3motif [78] dataset utilizes BA graphs as the basis. Each base graph is augmented with one of three motifs: house, cycle, or grid.

Tree-Cycles [64] dataset begins with a basic graph that is an 8-level balanced binary tree. Additionally, the dataset includes 80 six-node cycle motifs, which are randomly attached to nodes in the base graph.

Tree-Grids [64] dataset shares similarities with the Tree-Cycles dataset. However, instead of attaching cycle motifs to the base tree graph, 3-by-3 grid motifs are attached in the Tree-Grids dataset.

Other synthetic datasets include Is_Acyclic [116], BA-3motif [78], Infection [45], and Tree-BA [50].

Real-world Datasets. Depending on the specific domain, real-world datasets can be divided into the following categories:

- **Chemistry**: MUTAG [117] consists of 4,337 molecule graphs that have been labeled based on their mutagenic effect on a specific type of bacteria. Each graph in the dataset represents a molecule, and its label indicates whether it is mutagenic or non-mutagenic. ClinTox, Tox21, BBBP, and BACE [6] are publicly available and widely used in drug discovery and chemical informatics. These datasets consist of chemical molecule graphs that are labeled based on their specific chemical properties. NCI1 [32] dataset consists of 4,110 chemical compounds that are labeled as positive or negative in relation to cell lung cancer.

Table 5. Datasets of GNN explanation tasks.

Class	Dataset	Domain	Paper
Synthetic Datasets	BA-Shapes	Artificially generated	[64], [66], [70], [68], [77], [48], [91], [90], [98], [89], [67], [61]
	BA-Community		[64], [66], [67], [77], [14], [89]
	BA-2motifs		[66], [68], [77], [85]
	BA-3motif		[78]
	Tree-Cycles		[64], [66], [67], [70], [77], [48], [91], [98], [89]
	Tree-Grids		[64], [66], [67], [70], [77], [50], [89], [61]
	Is_Acyclic		[97]
	Infection		[61]
	Tree-BA		[50]
Real-world Datasets	MUTAG	Chemistry	[64], [66], [97], [68], [77], [78], [48], [91], [51], [98], [50], [89], [85], [61]
	ClinTox		[90]
	Tox21		[90]
	BBBP		[68], [90], [85]
	BACE		[90], [85]
	NCI1		[48], [91], [98]
	REDDIT-BINARY	Social network	[64]
	REDDIT-MULTI-5K		[51]
	Reddit		[76]
	Bitcoin-Alpha	Bitcoin exchange network	[67]
	Bitcoin-OTC		[67]
	Graph-SST2	Text classification	[68], [90], [89], [85]
	Twitter		[85]
	Wikipedia		[72]
	Cora	Citation network of machine learning	[14], [87]
	CiteSeer		[14], [48]
	PubMed	Citation network of biomedicine	[14], [87]
	VG-5	Computer vision	[78]
	MNIST		[67], [78]
	OGBN-MAG	Citation network	[75]

- **Social Networks**: Reddit [11] is a graph dataset constructed from posts made in September 2014, where nodes represent posts and edges connect posts if the same user comments on both, with the goal of predicting the community ("subreddit") each post belongs to. REDDIT-BINARY [29] is a dataset consisting of 2,000 graphs, where each graph represents an online discussion thread from Reddit. Nodes in the graphs represent participating users, and edges indicate replies between users' comments. REDDIT-MULTI-5K [29] is a dataset that includes 4,999 social networks labeled with five different classes, representing the topics of question/answer communities.
- **Bitcoin Exchange Network**: Bitcoin-Alpha and Bitcoin-OTC datasets [118] are networks comprising 3,783 and 5881 accounts involved in Bitcoin trading on platforms. Each account is rated by other members on a scale of -10 (total distrust) to +10 (total trust).
- **Text Classification**: Wikipedia is a temporal graph datasets composed of approximately 9,300 active users and the most frequently edited pages, containing around 160,000 temporal edges. Each temporal edge is accompanied by a 172-dimensional user editing feature vector, capturing user editing activities over time. Graph-SST2 [119] and Twitter [120] are sentiment graph datasets for graph classification. Nodes denote words and edges represent the relationships between words.

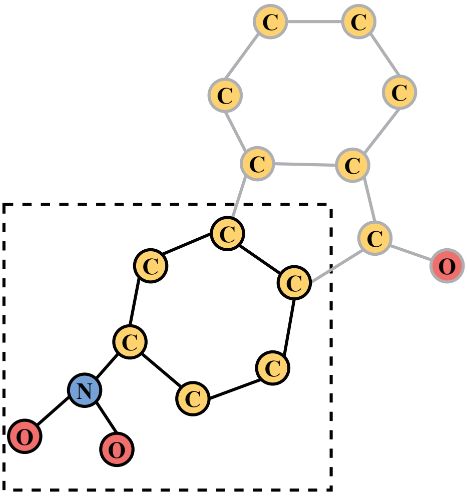

Fig. 11. The explanation in the form of nodes subgraph.

- **Citation Network of Machine Learning**: Cora [121] and CiteSeer [122] consist of machine learning papers. Nodes in these datasets represent individual papers, while edges represent the citations between papers. Cora dataset is labeled with seven class labels, whereas CiteSeer dataset has six class labels.
- **Citation Network of Biomedicine**: PubMed [30] is a citation network in the field of biomedicine, comprising papers from the biomedical domain. In this dataset, papers are represented as nodes, and edges indicate that one paper has been cited by another paper.
- **Computer Vision**: VG-5 [78] is constructed using 4,443 $(images, scene\ graphs)$ pairs from the Visual Genome dataset. Wherein, the graphs are labeled with five classes: stadium, street, farm, surfing, and forest. Each graph contains regions of the objects as nodes, while edges indicate the relationships between object nodes. MNIST [34] dataset consists of 70,000 images that are converted into graphs based on superpixel adjacency. Each graph in the MNIST dataset is labeled with one of ten digit classes.
- **Citation Network**: OGBN-MAG is a heterogeneous temporal graph dataset extracted from the Microsoft Academic Graph (MAG) [123], consisting of yearly sliced academic events, with nodes and edges representing authors, papers, fields, institutions, and their relationships.

B Explanation Sample

Fig. 11 presents a subgraph explanation, where the subgraph within the dashed box represents the generated explanation. Fig. 12 shows an explanation in the form of nodes, node features, and edges. The black circles indicate the explanation nodes, the red edges represent the explanation edges, and the gray box marks the positions of the retained node feature vectors. Fig. 13 illustrates the explanation in the form of walks. For simplicity, we present the explanation as 1-edge walks, with the black edges indicating these walks. Fig. 14 demonstrates the explanation in the form of flows, where the red lines represent the flows in the context of a two-layer GNN.

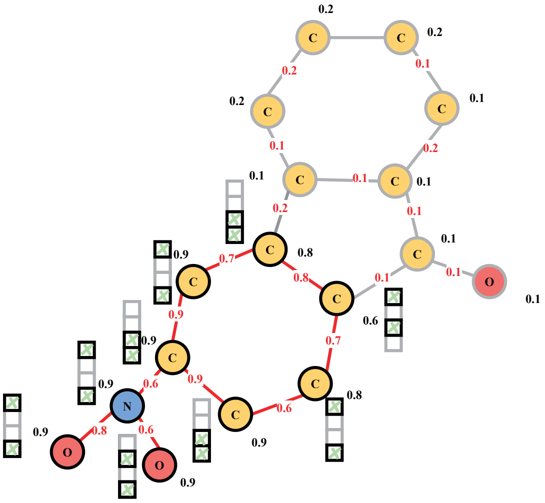

Fig. 12. The explanation in the form of nodes, edges, and node features.

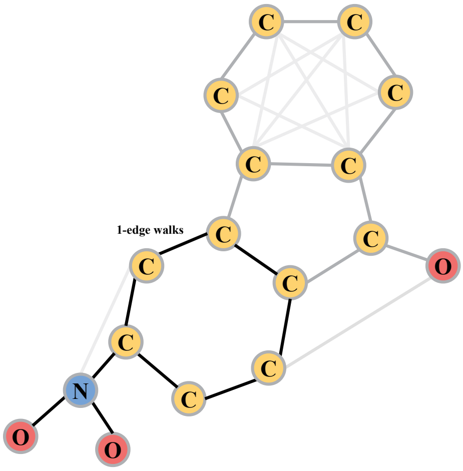

Fig. 13. The explanation in the form of walks.

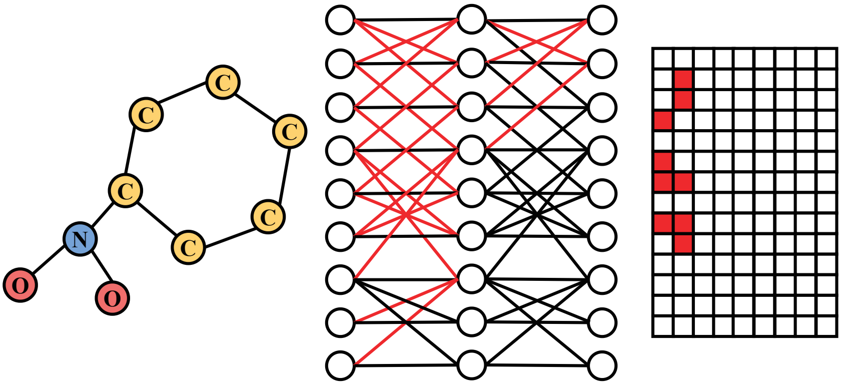

Fig. 14. The explanation in the form of flows.