GNN Explanations

I. Makarov & V. Pozdnyakov & D. Kiselev

BigData Academy MADE from Mail.ru Group

Graph Neural Networks and Applications

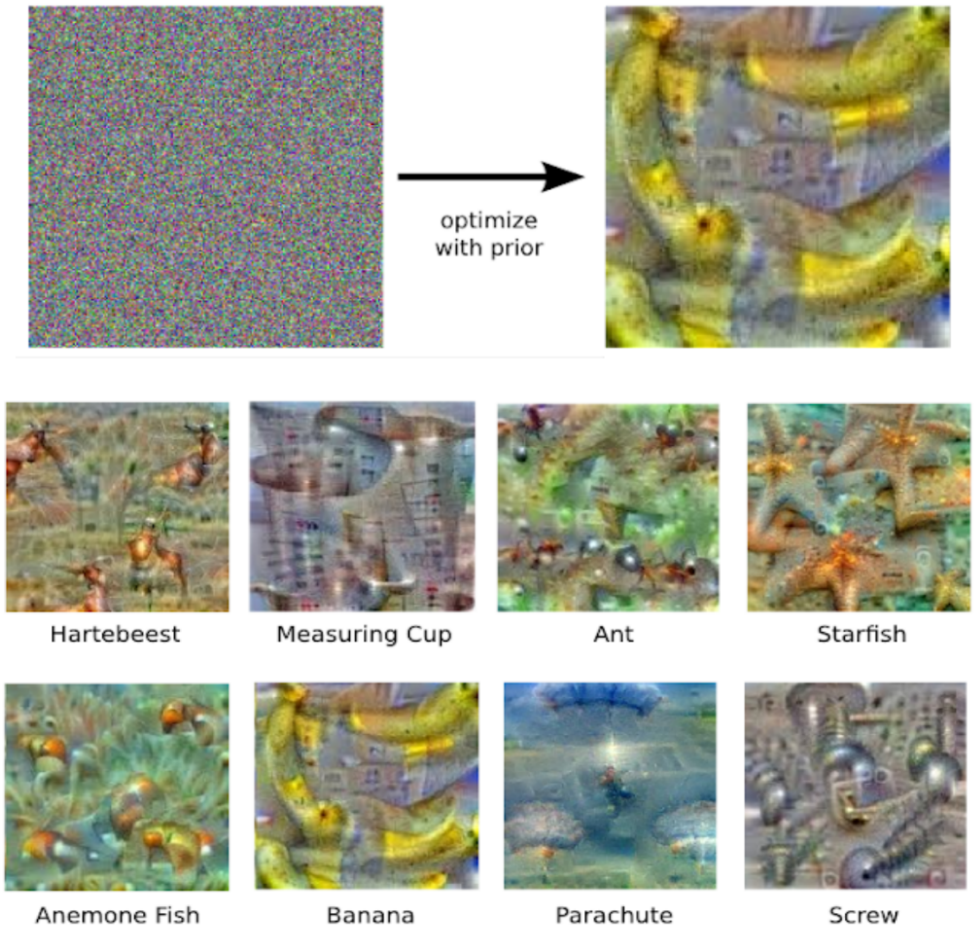
Topics

1. Explaining Neural Networks
2. Causality and Interpretability in GNNs

Deep Dream

Explaining Deep Neural Networks

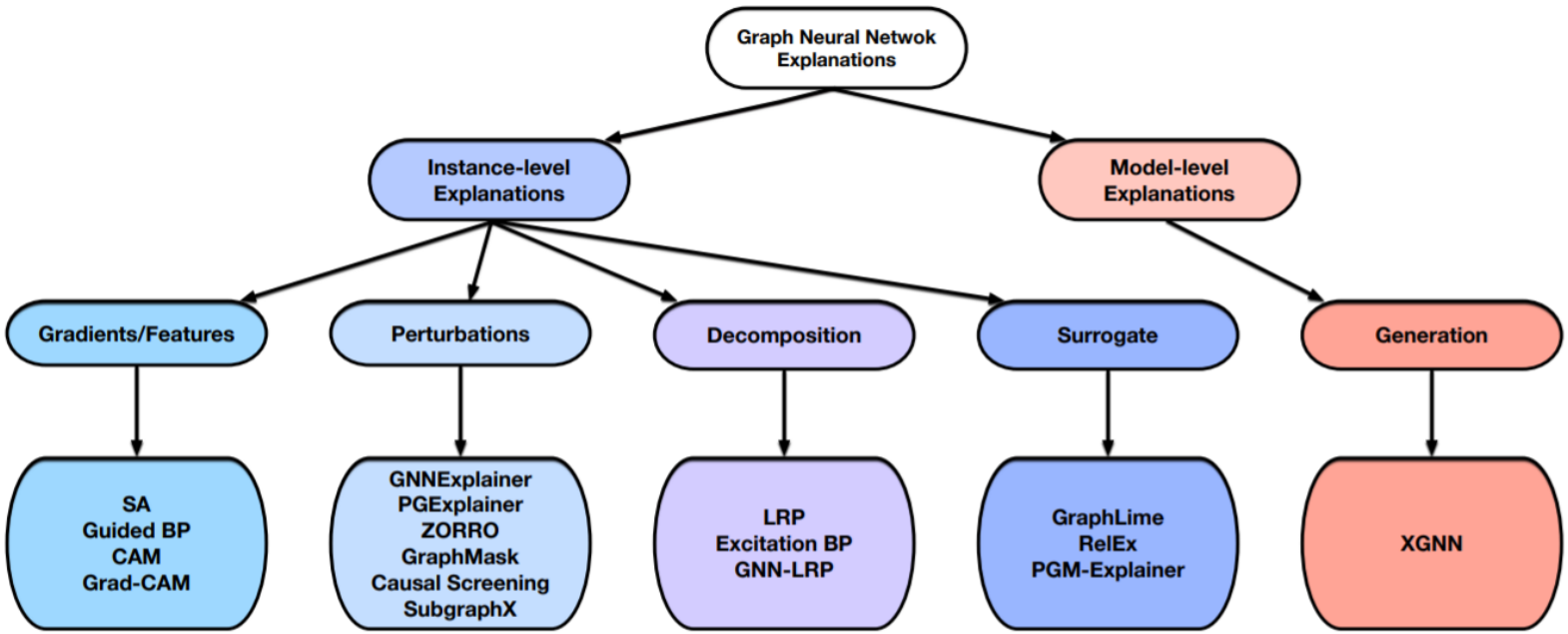
1. Train neural network
2. Optimize input for maximizing activation
3. Lucid lib for feature analysis `https://github.com/znah/lucid`



Explaining GNNs

Explainability in Graph Neural Networks: A Taxonomic Survey

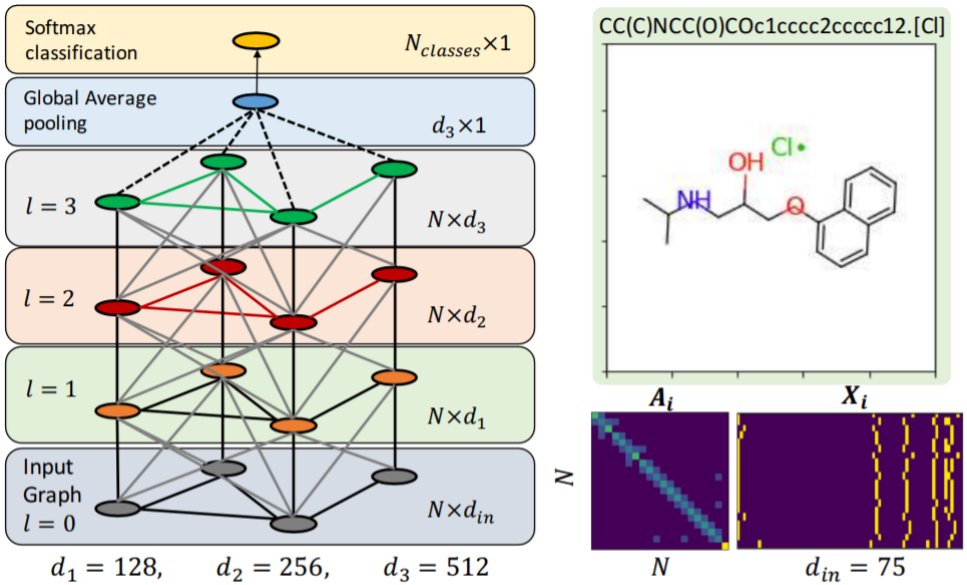
1. Gradient-based optimize input for activation
2. Perturbation-based sample neighborhood and feature subsets
3. Decomposition-based methods model separate samples impacts
4. Surrogate models deal with generating new datasets and fitting them to local input graph and its prediction



from Ji et al., 2020

CAM Explainability Methods for Graph Convolutional Neural Networks

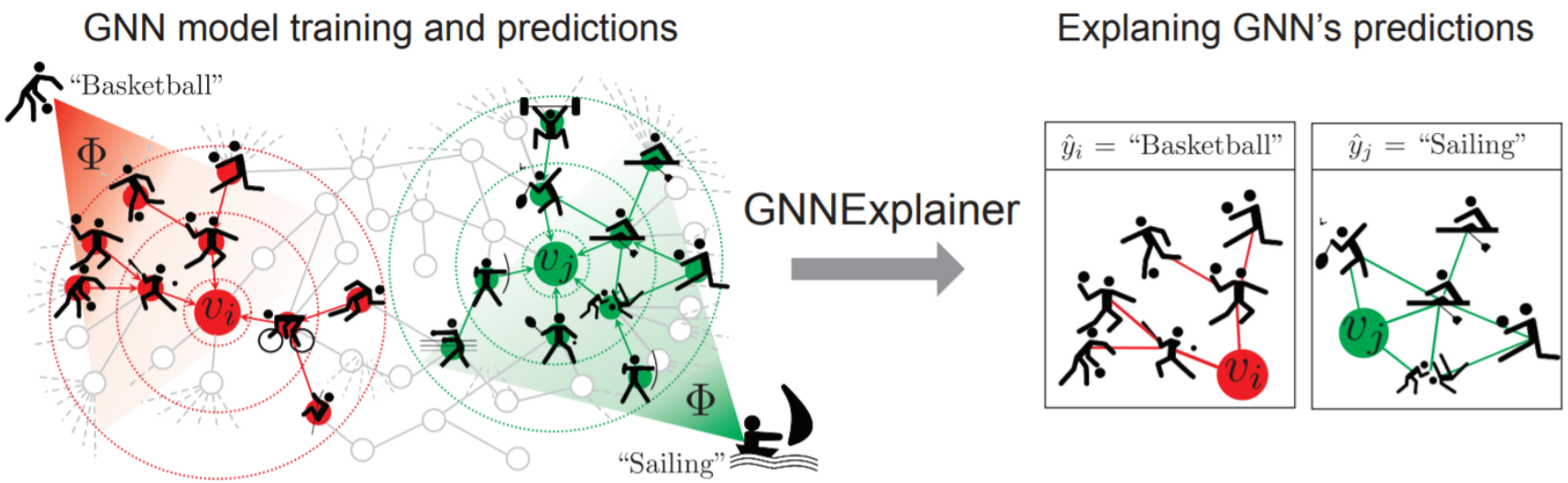
1. Class Activation Mapping (CAM) improves saliency maps for CNNs/GCNNs, by identifying semantic class-specific features at the last convolutional layer as opposed to the input space.
2. The downside of CAM is the convolutional layer followed by a global average pooling (GAP) before the softmax classifier.



from Hoffman et al., 2019, see also `https://mrsalehi.medium.com/`

GNNExplainer: Generating Explanations for Graph Neural Networks

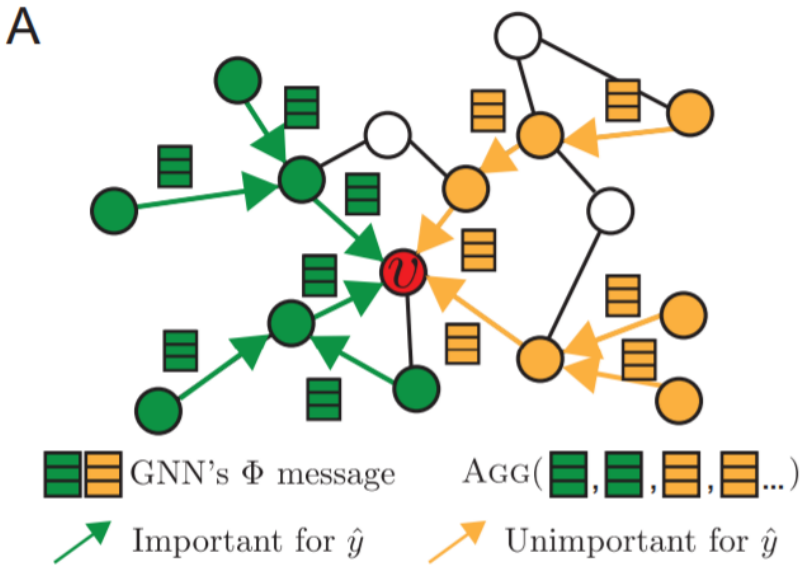
1. Extract meaningful subgraphs for prediction task
2. Discard neighbors that have no impact on node classification
3. See also applications to NLP: Titov et al., 2020 (GraphMASK), and CV: Chua et al., 2020 (Causal Screening).

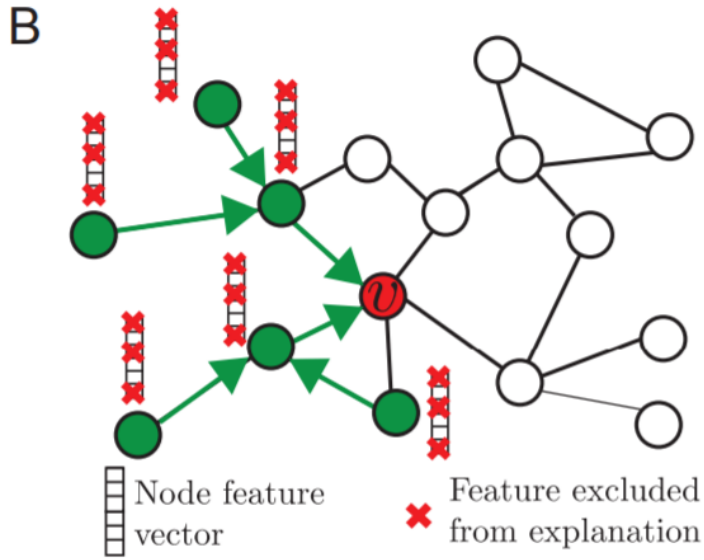


from Leskovec et al., 2019

GNNExplainer: Generating Explanations for Graph Neural Networks

1. GNN computation graph for inference of v
2. GNN needs to aggregate important as well as unimportant messages
3. GNNEXPLAINER identifies a small set of important features and pathways (green) that are crucial for prediction.
4. GNNEXPLAINER identifies what feature dimensions are important for prediction by learning a node feature mask.

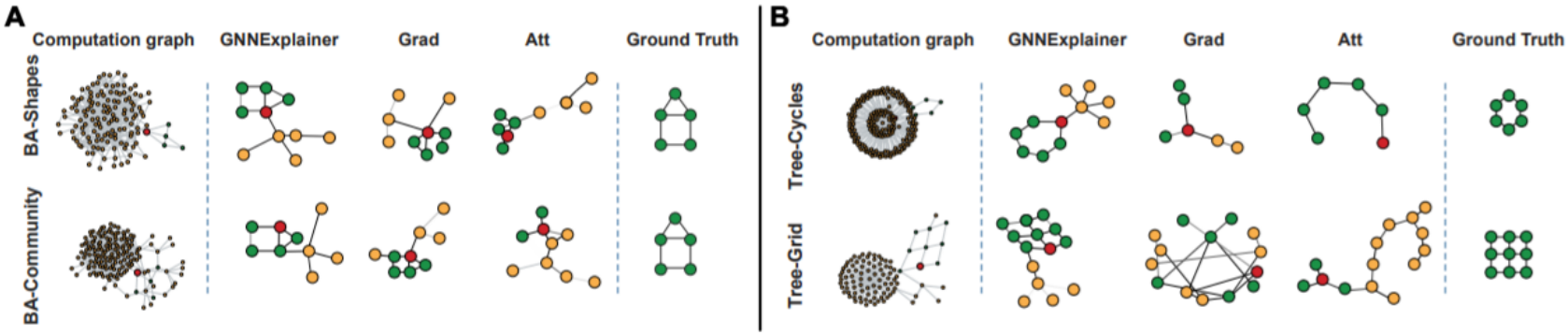




from Leskovec et al., 2019

GNNExplainer: Generating Explanations for Graph Neural Networks

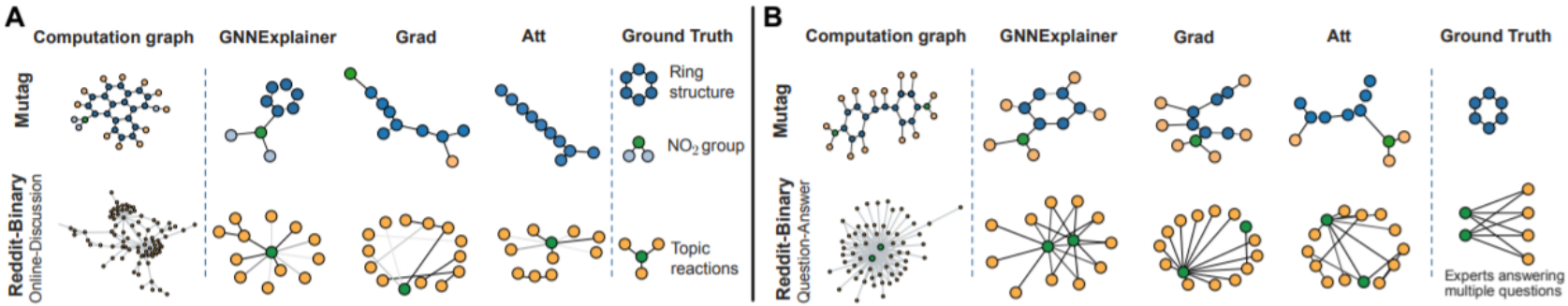
1. Explaining node-level classification
2. Red node is target node for classification
3. Grad optimizes features and adjacency matrix, Att learns GAT for edge attention



from Leskovec et al., 2019

GNNExplainer: Generating Explanations for Graph Neural Networks

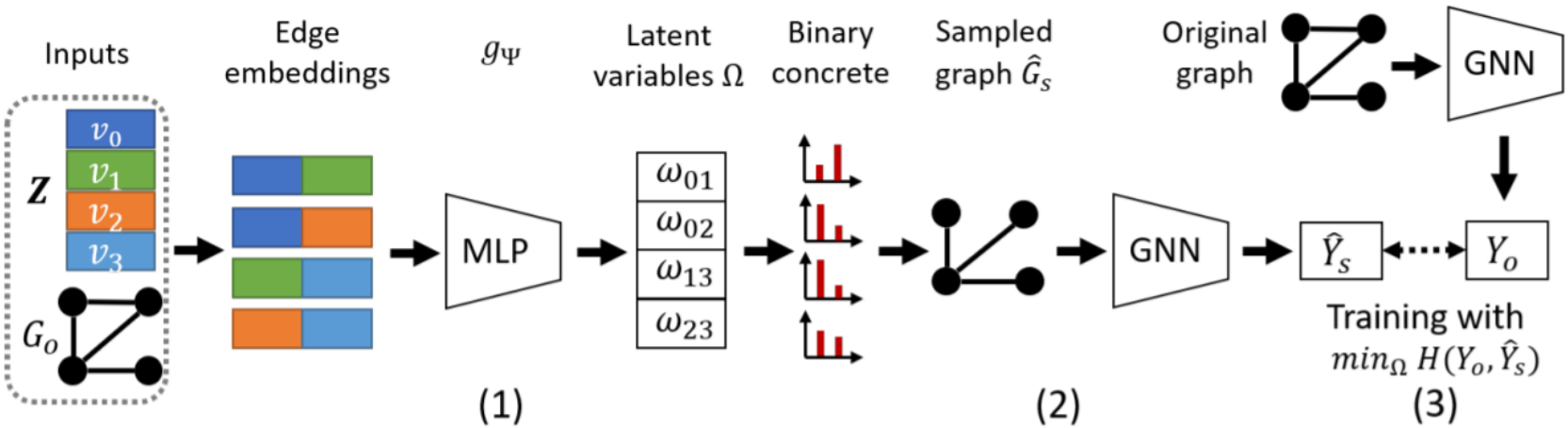
1. Explaining graph-level classification
2. Sub-graph pattern mining



from Leskovec et al., 2019

Parameterized Explainer for Graph Neural Network

1. PGExplainer takes node representations and the original graph to compute the latent variables in edge distributions.
2. In case that an explanatory subgraph is wanted, model selects top-ranked edges according to latent variables.
3. A random graph is sampled from edge distributions and then feed to the trained GNN model to get the prediction and optimize cross-entropy between original graph and sampled graph

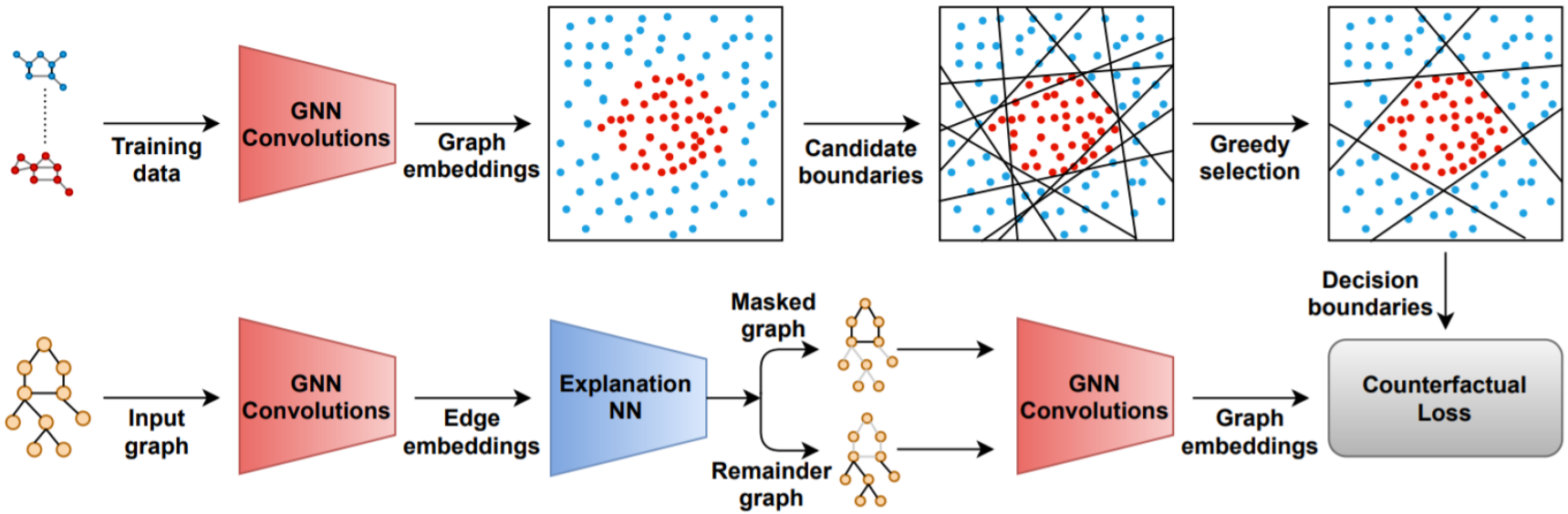


from Zhang et al., 2020

Robust Counterfactual Explanations on Graph Neural Networks

Given a trained GNN model for an input graph G, the goal of RCExplainer is to explain prediction by identifying a small subset of edges S, such that

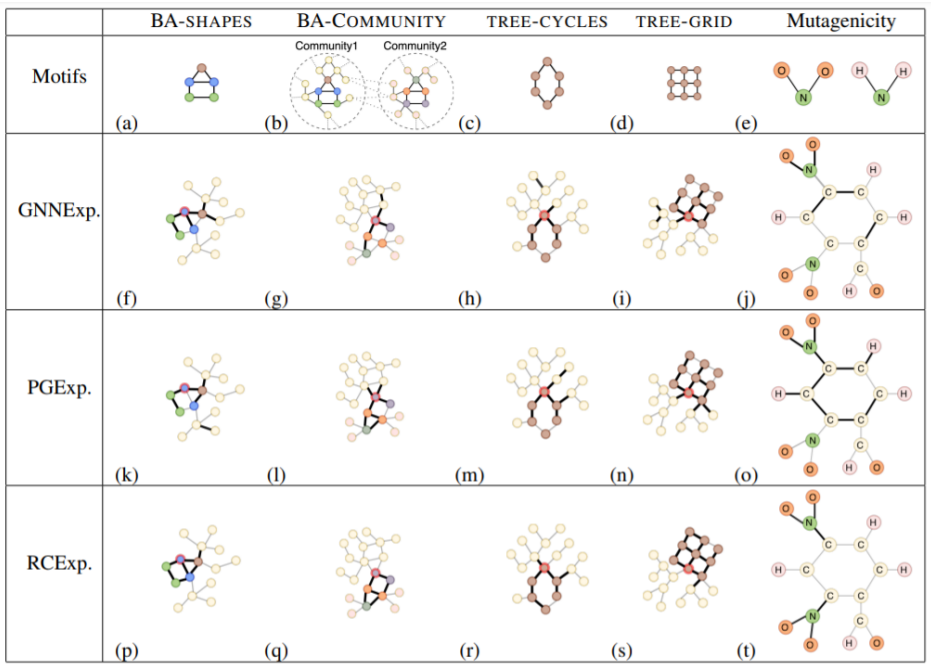
1. removing the set of edges in S from G changes the prediction on the remainder of G significantly;
2. S is stable with respect to slight changes on the edges of G and the feature representations of the nodes of G.



from Zhang et al., 2021

Robust Counterfactual Explanations on Graph Neural Networks

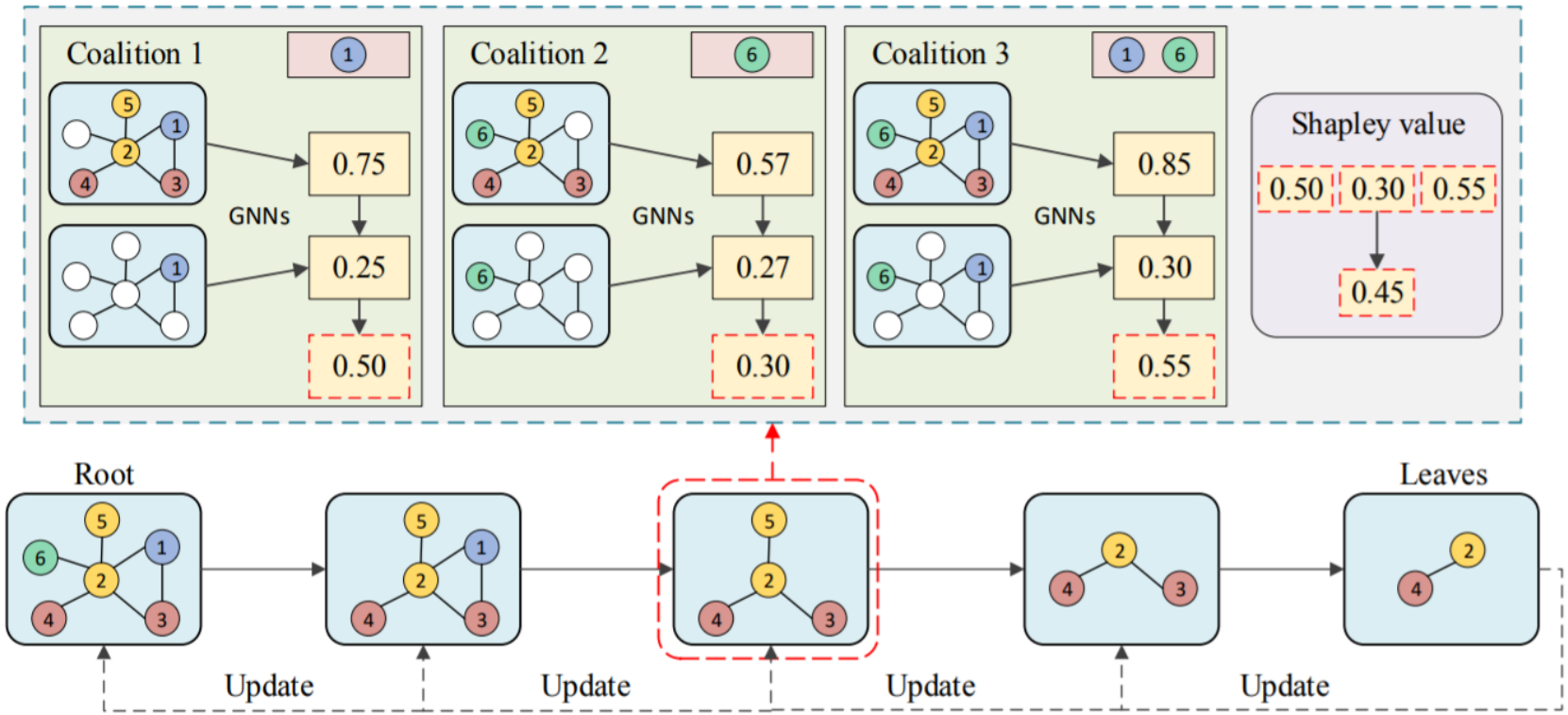
1. Graph motifs are followed by explanations.
2. First four columns correspond to node classification of red node and the node colors denote different labels.
3. Last column corresponds to graph classification for a mutagenic sample and explanations are highlighted in black.



from Zhang et al., 2021, see also Anand et al., 2021 (ZORRO)

On Explainability of Graph Neural Networks via Subgraph Explorations

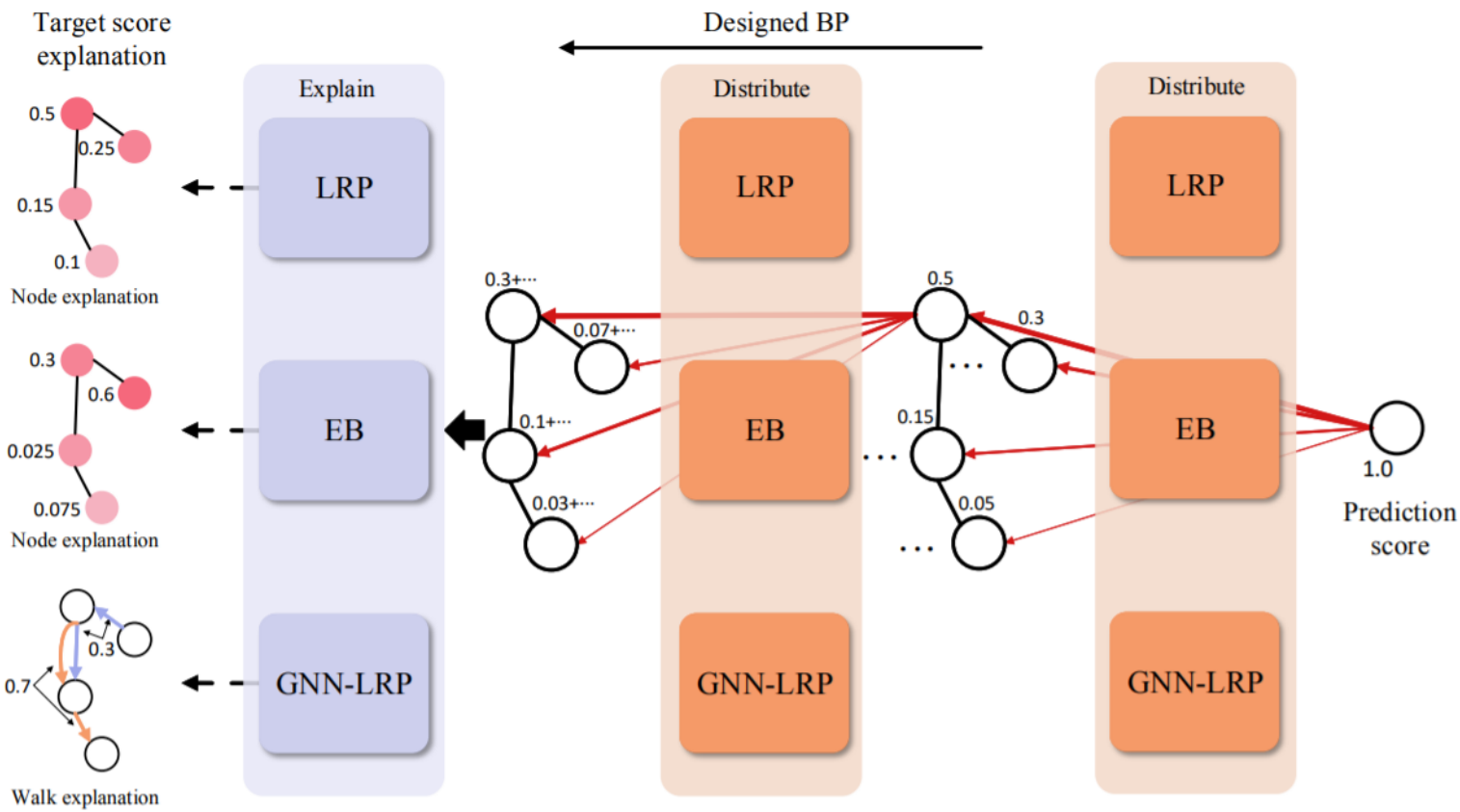
1. SubgraphX uses MCTS for subgraph search
2. Result is evaluated by computing the Shapley value via Monte-Carlo sampling.



from Ji et al., 2021

Decomposition Models

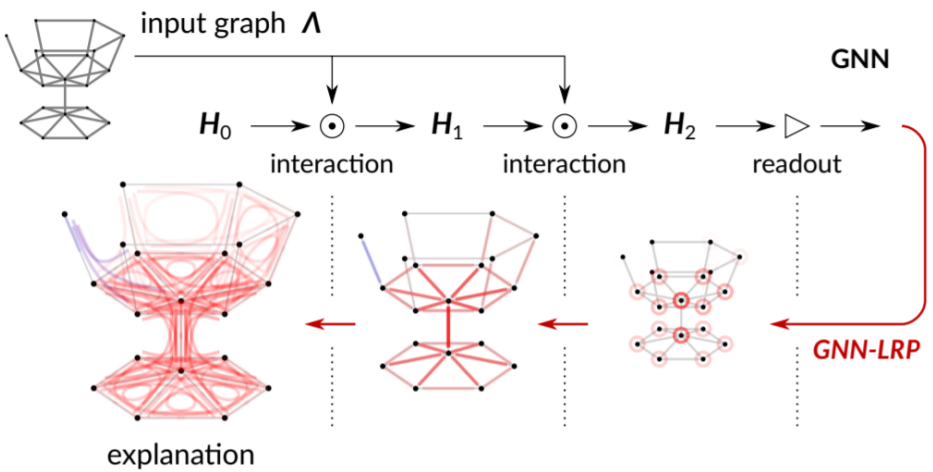
1. These methods distribute the prediction score to input space to indicate the input importance.
2. The target score is decomposed layer by layer via backpropagation.

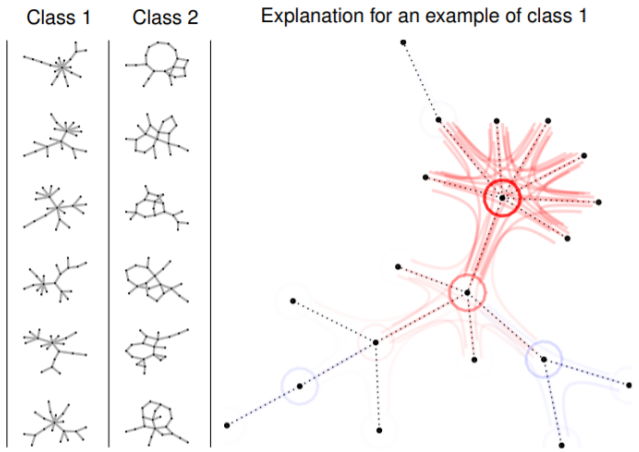


from Ji et al., 2020

Higher-Order Explanations of Graph Neural Networks via Relevant Walks

1. Score relevance of walks
2. Iteratively train edge importance and layer-wise transitions

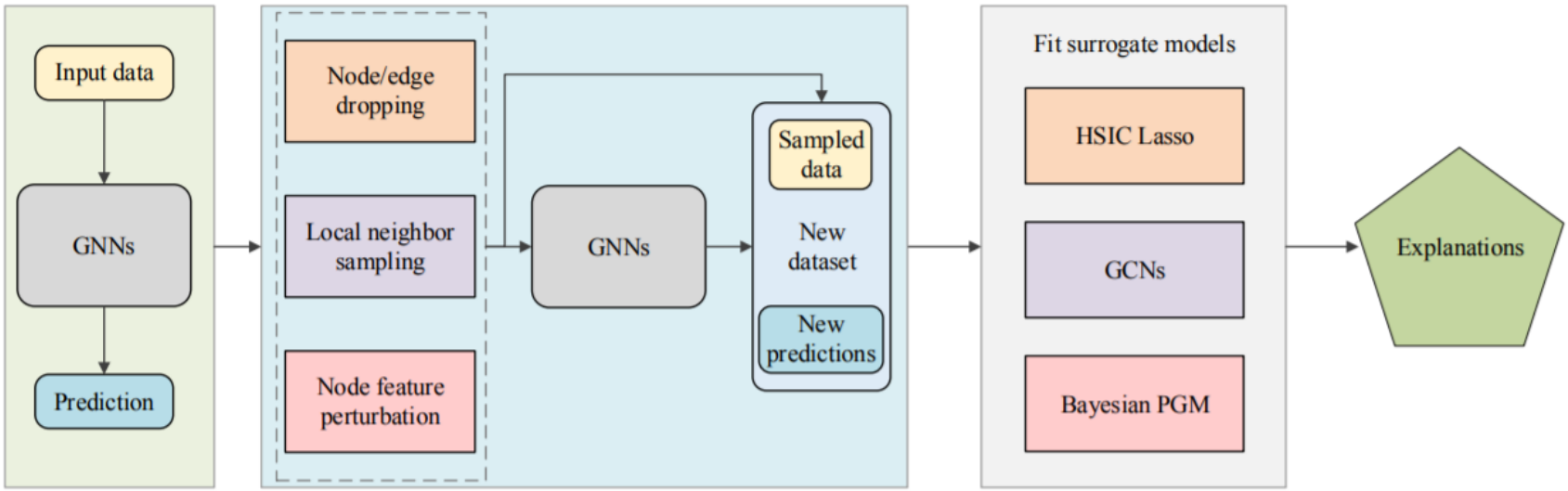




from Montavon et al., 2020

Surrogate Models

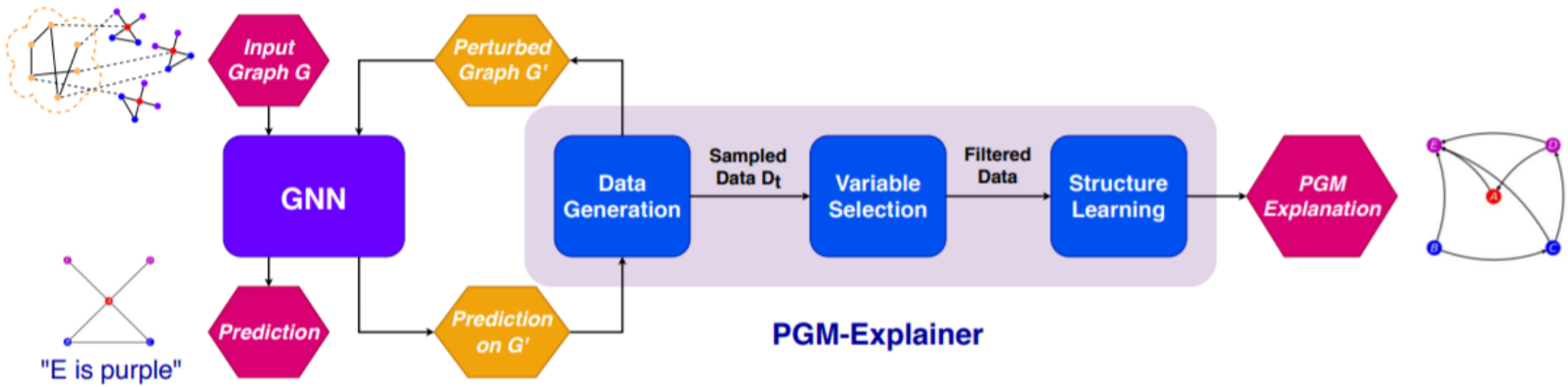
1. First sample a local dataset to represent the relationships around the target data.
2. Different interpretable surrogate methods are applied to fit the local dataset.
3. Explanations from the surrogate model can be regarded as the explanations of the original prediction.



from Ji et al., 2020

PGM-Explainer: Probabilistic graphical model explanations for graph neural networks.

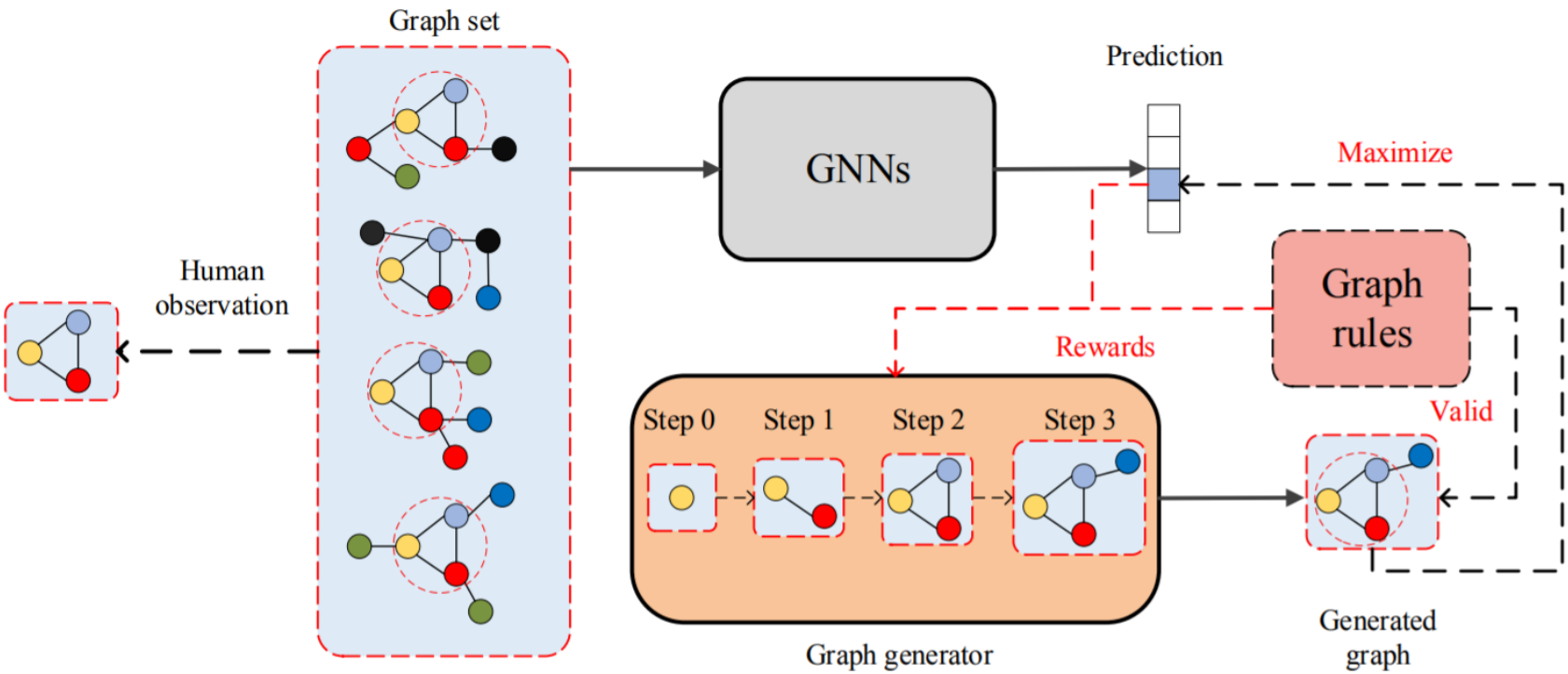
1. PGM-Explainer generates perturbed graphs and records GNN's predictions on those graphs in the data generation step.
2. The variable selection step eliminates unimportant explained features in this data and forwards the filtered data.
3. Finally, the PGM is generated in the structure learning step.



from Thai et al., 2020

XGNN: Towards Model-Level Explanations of Graph Neural Networks

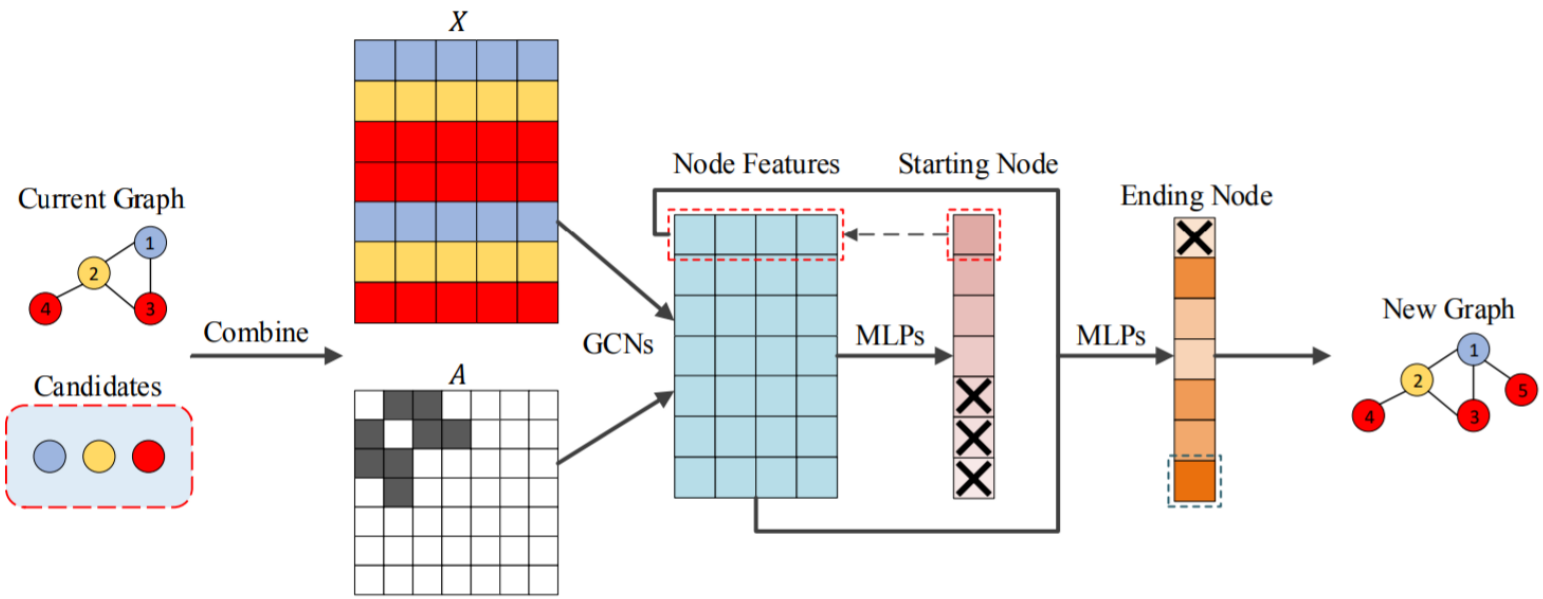
1. GNNs represent a trained graph classification model that we try to explain.
2. Interpreting GNNs via Graph Generation with mutual patterns



from Ji et al., 2020

XGNN: Towards Model-Level Explanations of Graph Neural Networks

1. Combine given graph and candidate.
2. Employ several GCN layers to aggregate and learn node features.
3. First MLPs predict a probability distribution for starting node.
4. Second MLPs predict the ending node conditioned on the starting node. Black crosses indicates masking out nodes.



from Ji et al., 2020

References

- Ying, R., Bourgeois, D., You, J., Zitnik, M. and Leskovec, J., 2019. Gnnexplainer: Generating explanations for graph neural networks. Advances in neural information processing systems, 32, p.9240.
- Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., Chen, H. and Zhang, X., 2020. Parameterized explainer for graph neural network. arXiv preprint arXiv:2011.04573.
- Bajaj, M., Chu, L., Xue, Z.Y., Pei, J., Wang, L., Lam, P.C.H. and Zhang, Y., 2021. Robust Counterfactual Explanations on Graph Neural Networks. Advances in Neural Information Processing Systems, 34.
- Rao, J., Zheng, S. and Yang, Y., 2021. Quantitative Evaluation of Explainable Graph Neural Networks for Molecular Property Prediction. arXiv preprint arXiv:2107.04119.
- Lucic, A., ter Hoeve, M., Tolomei, G., de Rijke, M. and Silvestri, F., 2021. CF-GNNExplainer: Counterfactual Explanations for Graph Neural Networks. arXiv preprint arXiv:2102.03322.

References

- Schnake, T., Eberle, O., Lederer, J., Nakajima, S., Schütt, K.T., Müller, K.R. and Montavon, G., 2020. Higher-order explanations of graph neural networks via relevant walks. arXiv preprint arXiv:2006.03589.
- Yuan, H., Tang, J., Hu, X. and Ji, S., 2020, August. 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).
- Yuan, H., Yu, H., Wang, J., Li, K. and Ji, S., 2021. On explainability of graph neural networks via subgraph explorations. arXiv preprint arXiv:2102.05152.
- Lin, W., Lan, H. and Li, B., 2021. Generative Causal Explanations for Graph Neural Networks. arXiv preprint arXiv:2104.06643.
- https://github.com/AstraZeneca/awesome-explainable-graph-reasoning

References

- Yuan, H., Yu, H., Gui, S. and Ji, S., 2020. Explainability in graph neural networks: A taxonomic survey. arXiv preprint arXiv:2012.15445.
- Agarwal, C., Zitnik, M. and Lakkaraju, H., 2021. Towards a Rigorous Theoretical Analysis and Evaluation of GNN Explanations. arXiv preprint arXiv:2106.09078.
- Liu, M., Luo, Y., Wang, L., Xie, Y., Yuan, H., Gui, S., Yu, H., Xu, Z., Zhang, J., Liu, Y. and Yan, K., 2021. DIG: A Turnkey Library for Diving into Graph Deep Learning Research. arXiv preprint arXiv:2103.12608.
- Huang, Q., Yamada, M., Tian, Y., Singh, D., Yin, D. and Chang, Y., 2020. Graphlime: Local interpretable model explanations for graph neural networks. arXiv preprint arXiv:2001.06216.
- Magister, L.C., Kazhdan, D., Singh, V. and Liò, P., 2021. GCExplainer: Human-in-the-Loop Concept-based Explanations for Graph Neural Networks. arXiv preprint arXiv:2107.11889.

References

- Pope, P.E., Kolouri, S., Rostami, M., Martin, C.E. and Hoffmann, H., 2019. Explainability methods for graph convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10772-10781).
- Funke, T., Khosla, M. and Anand, A., 2021. Zorro: Valid, Sparse, and Stable Explanations in Graph Neural Networks. arXiv preprint arXiv:2105.08621.
- Schlichtkrull, M.S., De Cao, N. and Titov, I., 2020. Interpreting graph neural networks for nlp with differentiable edge masking. arXiv preprint arXiv:2010.00577.
- Wang, X., Wu, Y., Zhang, A., He, X. and Chua, T.S., 2020. Causal Screening to Interpret Graph Neural Networks. https://openreview.net/forum?id=nzKv5vxZfge

References

- Zhang, Y., Defazio, D. and Ramesh, A., 2021, July. Relex: A model-agnostic relational model explainer. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 1042-1049).
- Vu, M.N. and Thai, M.T., 2020. Pgm-explainer: Probabilistic graphical model explanations for graph neural networks. arXiv preprint arXiv:2010.05788.
- Li, X., 2020, August. Explain graph neural networks to understand weighted graph features in node classification. In International Cross-Domain Conference for Machine Learning and Knowledge Extraction (pp. 57-76). Springer, Cham.
- Duval, A. and Malliaros, F.D., 2021. GraphSVX: Shapley Value Explanations for Graph Neural Networks. arXiv preprint arXiv:2104.10482.
- Hu, J., Li, T. and Dong, S., 2020, July. GCN-LRP explanation: exploring latent attention of graph convolutional networks. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.