The Road to Explainable Graph Neural Networks

Sayan Ranu
Department of Computer Science & Engineering, and Yardi School of AI (Joint Appointment), IIT Delhi
sayanranu@iitd.ac.in

GOAL: Graph Neural Networks (GNNs) are being increasingly adopted in various real-world applications, including drug and material discovery [16, 17, 27], recommendation engines [29], congestion modeling on road networks [5], and weather forecasting [11]. However, similar to other deep-learning models, GNNs are considered black boxes due to their limited capacity to provide explanations for their predictions. This lack of interpretability poses a significant barrier to their adoption in critical domains such as healthcare, finance, and law enforcement, where transparency and trustworthiness are essential for decision-making processes. In these pivotal domains, understanding the rationale behind model predictions is crucial not only for compliance with interpretability requirements but also for identifying potential vulnerabilities and gaining insights to refine the model further. *How do we make GNNs interpretable?* This is a central motivating question driving my research pursuits.

BACKGROUND: Existing GNN explainers can be grouped into *model-level* and *instance-level* explainers. Model-level explainers [8, 26, 32] aim to explain the overall behavior of the model that generalize across instances. Instance-level explainers [1, 3, 7, 12, 13, 14, 19, 21, 25, 30, 33, 34] explain a specific input graph by highlighting the subgraph components used by a GNN to make its prediction. The type of explanation offered is either *factual* [15, 24, 28, 32] or *counterfactual* [2, 4, 13, 22]. Factual explainers aim to find an important subgraph that correlates most with the underlying GNN prediction. In contrast, counterfactual reasoning attempt to identify the smallest amount of perturbation on the input graph (e.g., removal/addition of edges or nodes) that changes the GNN's prediction. Thus, while factual explainers surface factors influencing a prediction, counterfactual explainers provide insights into how small changes in the input lead to different outcomes.

CHALLENGES AND IDEAS:

Completeness: Factual explanations aim to identify the cause of a particular prediction. However, an interesting observation has emerged: when the explanations (subgraph) are removed from the input graphs, and the GNN is retrained on the residual graphs, it often manages to recover the correct predictions [9, 10]. This observation prompts a crucial inquiry: *Are the explanations complete?* The incompleteness of explainers is likely a manifestation of their *post-hoc* learning framework. Specifically, the post-hoc paradigm treats the GNN as a black box where the explainers have no visibility to model internals such as the loss surface and hyper-parameters.

Idea: Remedies may lie within the *ante-hoc* paradigm, where the GNN and the explainer undergo *joint* training [9]. Joint training enables access to a more comprehensive set of information influencing the GNN predictions, including the loss surface, topological components of the input activating neurons, and gradient flows through the layers.

Feasibility: An important facet of counterfactual reasoning lies in its ability to offer recourse options. The practical effectiveness of the recourse hinges on its alignment with specific domain constraints. In molecular datasets, for instance, a valid recourse should yield a chemically sound molecule. Current counterfactual explainers predominantly focus on pinpointing the shortest edit path that steers the graph toward the decision boundary, often overlooking the feasibility of the proposed edits [10].

Idea: The loss associated with a counterfactual explanation is typically a function of: *(1)* the size of the explanation and *(2)* the quality of the explanation. I propose the addition of a third term that models the *feasibility* of the explanation. Modeling feasibility presents non-trivial challenges, and potential solutions may be found in the area of generative modeling for graphs [6, 23, 31]. Generative models learn the underlying distribution of topological patterns in the training data and utilize that knowledge to generate new, similar graphs. This paradigm can be extended to estimate the probability of the recourse being generated, corresponding to the feasibility term in the loss function.

Stability: Humans are the intended audience for explanations. Hence, they need to be stable. Explainers are neural networks themselves optimizing parameters on a non-convex loss surface. Consequently, they are unstable to the initialization seeds [10] impacting human interpretability.

Idea: A neural network's stability is connected to its Lipschitz constant [20]. Lipschitz-constrained neural networks augment stability by bounding its Lipschitz constant [18]. Exploring Lipschitz-constrained GNN explainers represents a promising research avenue.

REFERENCES

[1] C. Abrate and F. Bonchi. Counterfactual graphs for explainable classification of brain networks. In *KDD*, page 2495–2504, 2021.

[2] C. Abrate and F. Bonchi. Counterfactual graphs for explainable classification of brain networks. In *KDD*, 2021.

[3] M. Bajaj, L. Chu, Z. Y. Xue, J. Pei, L. Wang, P. C.-H. Lam, and Y. Zhang. Robust counterfactual explanations on graph neural networks. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[4] M. Bajaj, L. Chu, Z. Y. Xue, J. Pei, L. Wang, P. C.-H. Lam, and Y. Zhang. Robust counterfactual explanations on graph neural networks. *arXiv preprint arXiv:2107.04086*, 2021.

[5] A. Derrow-Pinion, J. She, D. Wong, O. Lange, T. Hester, L. Perez, M. Nunkesser, S. Lee, X. Guo, B. Wiltshire, P. W. Battaglia, V. Gupta, A. Li, Z. Xu, A. Sanchez-Gonzalez, Y. Li, and P. Velickovic. Eta prediction with graph neural networks in google maps. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*. ACM, Oct. 2021.

[6] N. Goyal, H. V. Jain, and S. Ranu. Graphgen: a scalable approach to domain-agnostic labeled graph generation. In *Proceedings of The Web Conference 2020*, pages 1253–1263, 2020.

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

[8] Z. Huang, M. Kosan, S. Medya, S. Ranu, and A. Singh. Global counterfactual explainer for graph neural networks. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 141–149, 2023.

[9] M. Kosan, A. Silva, and A. Singh. Robust ante-hoc graph explainer using bilevel optimization, 2023.

[10] M. Kosan, S. Verma, B. Armgaan, K. Pahwa, A. Singh, S. Medya, and S. Ranu. Gnnx-bench: Unravelling the utility of perturbation-based gnn explainers through in-depth benchmarking, 2023.

[11] R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, A. Merose, S. Hoyer, G. Holland, O. Vinyals, J. Stott, A. Pritzel, S. Mohamed, and P. Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 0(0):eadi2336, 2023.

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

[13] A. Lucic, M. A. Ter Hoeve, G. Tolomei, M. De Rijke, and F. Silvestri. Cf-gnnexplainer: Counterfactual explanations for graph neural networks. In *AISTATS*, pages 4499–4511, 2022.

[14] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang. Parameterized explainer for graph neural network. *Advances in neural information processing systems*, 33:19620–19631, 2020.

[15] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang. Parameterized explainer for graph neural network. *arXiv preprint arXiv:2011.04573*, 2020.

[16] A. Merchant, S. Batzner, S. Schoenholz, M. Aykol, G. Cheon, and E. Cubuk. Scaling deep learning for materials discovery. *Nature*, 624:1–6, 11 2023.

[17] L. Rampášek, M. Galkin, V. P. Dwivedi, A. T. Luu, G. Wolf, and D. Beaini. Recipe for a General, Powerful, Scalable Graph Transformer. *Advances in Neural Information Processing Systems*, 35, 2022.

[18] K. Scaman and A. Virmaux. Lipschitz regularity of deep neural networks: Analysis and efficient estimation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 3839–3848, 2018.

[19] C. Shan, Y. Shen, Y. Zhang, X. Li, and D. Li. Reinforcement learning enhanced explainer for graph neural networks. In *NeurIPS 2021*, December 2021.

[20] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[21] J. Tan, S. Geng, Z. Fu, Y. Ge, S. Xu, Y. Li, and Y. Zhang. Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 1018–1027, 2022.

[22] J. Tan, S. Geng, Z. Fu, Y. Ge, S. Xu, Y. Li, and Y. Zhang. Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning. In *WebConf*, 2022.

[23] C. Vignac, I. Krawczuk, A. Siraudin, B. Wang, V. Cevher, and P. Frossard. Digress: Discrete denoising diffusion for graph generation. In *The Eleventh International Conference on Learning Representations*, 2023.

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

[25] G. P. Wellawatte, A. Seshadri, and A. D. White. Model agnostic generation of counterfactual explanations for molecules. *Chemical science*, 13(13):3697–3705, 2022.

[26] H. Xuanyuan, P. Barbiero, D. Georgiev, L. C.

Magister, and P. Lió. Global concept-based interpretability for graph neural networks via neuron analysis. 2023.

[27] C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen, and T.-Y. Liu. Do transformers really perform badly for graph representation? In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[28] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32:9240, 2019.

[29] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *KDD*, page 974–983, 2018.

[30] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32, 2019.

[31] J. You, R. Ying, X. Ren, W. L. Hamilton, and J. Leskovec. Graphrnn: Generating realistic graphs with deep auto-regressive models. In *ICML, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5694–5703. PMLR, 2018.

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

[33] H. Yuan, H. Yu, S. Gui, and S. Ji. Explainability in graph neural networks: A taxonomic survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[34] H. Yuan, H. Yu, J. Wang, K. Li, and S. Ji. On explainability of graph neural networks via subgraph explorations. In *ICML*, pages 12241–12252. PMLR, 2021.