GNN Explainers 2.0: A Paradigm for User-Oriented, Data-Guided Explanations

Arijit Khan
Bowling Green State University
Bowling Green, OH, USA
arijitk@bgsu.edu

***Abstract*—Graph neural networks (GNNs) have emerged as powerful deep learning models for graph-structured data, achieving state-of-the-art performance across diverse domains including social networks, knowledge graphs, bioinformatics, transportation, the World Wide Web, and finance [9]. They have been successfully applied to tasks such as node and graph classification, link prediction, entity resolution, question answering, recommendation, and fraud detection. Despite these advances, explaining the decisions of high-performing yet opaque GNNs remains a critical challenge. Over the past five years, significant progress has been made with the development of GNN explanation methods [18]–[20] (e.g., GNNExplainer [10], PGExplainer [11], SubgraphX [12], PGMExplainer [13], GraphLime [14], GCFExplainer [15], CF2 [16], GNN-LRP [17]). We collectively refer to these approaches as "GNN Explainers 1.0," which focus on identifying influential nodes, edges, subgraphs, and features to provide post hoc explanations of model outputs. While effective for narrow tasks such as node or graph classification, their one-time, output-centric nature limits broader applicability. For practical debugging, accountability, and trustworthy deployment, explanations must extend beyond final predictions to capture layer-wise provenance, enabling data scientists to trace how inputs evolve through the network and pinpoint sources of error. Moreover, non-technical stakeholders require explanations that are accessible, configurable, and queryable through familiar interfaces–structured queries, counterfactuals, or natural language–supporting interactive exploration of model behavior by both experts and non-experts.**

To address these challenges, we propose a user-centered paradigm–GNN Explainers 2.0–that moves beyond static, output-only explanations toward actionable, end-user-facing insights [1]. Grounded in data management principles, this approach leverages data-centric operations–views, slicing, skyline selection, fairness analysis, robustness checks, and adversarial constraints–to enhance comprehension, usability, and trust while enabling interactive exploration tailored to diverse stakeholders. Recent advances exemplify this paradigm. GVEX introduces a two-tier structure in which lower-tier subgraphs provide factual and counterfactual reasons for predictions, while higher-tier patterns summarize these subgraphs to support efficient search and alignment with domain knowledge [2]. Counterfactual evidence (CE) methods detect fairness violations by identifying similar nodes treated differently by pre-trained GNNs [3]. Model slicing framework SliceGX generates finer-grained, layer-wise explanations in a progressive manner–capabilities that are crucial for model diagnosis and architecture optimization [4]. Robust counterfactual witnesses (k-RCW) ensure resilience to structural disturbances, strengthening trust in post-hoc explanations [5]. Skyline-based methods, e.g., SXQ, generate diversified explanations across multiple evaluation metrics, balancing fidelity+, fidelity-, and sparsity [6]. ATEX-CF unifies adversarial attack methods with counterfactual explanation generation by leveraging their shared objective of inducing prediction flips through targeted perturbations [7]. It efficiently handles both edge additions and deletions, using adversarial insights to identify structurally minimal and impactful counterfactuals. Advances in parallel, streaming, and indexing techniques further improve scalability and efficiency [2], [3], [5]. Finally, explainability is expanding beyond classification, with frameworks such as NAEx supporting complex tasks like GNN-based network alignment [8]. Collectively, these contributions establish a foundation for interactive, trustworthy, and domain-adaptable GNN explanations as shown in Figure 1. Where GNN Explainers 1.0 primarily deliver static, instance-level or model-level rationales for predictions, GNN Explainers 2.0 emphasize interactive, multi-level, and stakeholder-aware explanations that integrate model behavior, data provenance, and user intent.

Future work may enable interactive dialogue with GNNs through ad hoc, declarative, and natural language explanatory queries; produce more accessible explanations via natural language descriptions, exemplars, and pattern- or rule-based concepts; incorporate human-in-the-loop feedback and explanatory query recommendation during model interrogation; and extend explanation methods to advanced architectures, including graph Transformers and emerging graph foundation models.

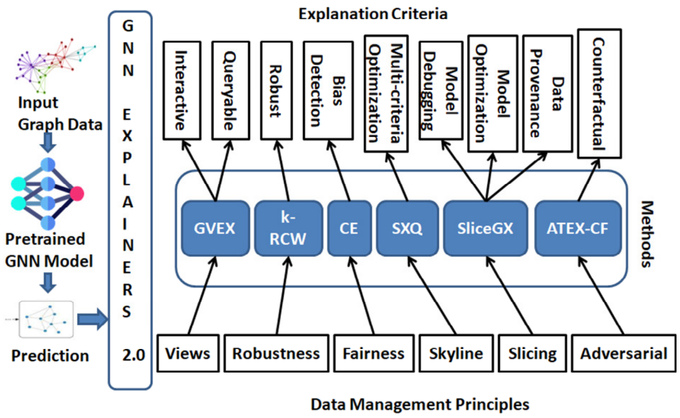

Fig. 1: The GNN Explainers 2.0 paradigm applies data-centric operations–views, slicing, skyline selection, fairness analysis, robustness checks, and adversarial constraints–to improve interpretability, usability, and trust while supporting interactive exploration based on diverse stakeholders' explanation criteria.

***Index Terms*—Graph Neural Networks, Explainable AI, User-friendly Explanations**

I. Acknowledgement

This talk is based on a sequence of recent works [1]–[8].

II. AI-Generated Content Acknowledgement

This paper has no AI-generated content.

References

[1] A. Khan, X. Ke, Y. Wu, and F. Bonchi, "GNN Explainers 2.0: User-centric and Data Driven Insights," WSDM (2026).

[2] T. Chen, D. Qiu, Y. Wu, A. Khan, X. Ke, and Y. Gao, "View-based Explanations for Graph Neural Networks," Proc. ACM Manag. Data, 2, 1 (2024), 40:1–40:27.

[3] D. Qiu, J. Chen, A. Khan, Y. Zhao, and F. Bonchi, "Finding Counterfactual Evidences for Node Classification," KDD (2025), 2362–2373.

[4] C. Yu, T. Zhu, T. Chen, Y. Wu, A. Khan, and X. Ke, "SliceGX: Layer-wise GNN Explanation with Model-slicing," The Web Conference (2026).

[5] D. Qiu, M. Wang, A. Khan, and Y. Wu, "Generating Robust Counterfactual Witnesses for Graph Neural Networks," ICDE (2024), 3351–3363.

[6] D. Qiu, H. Che, A. Khan, and Y. Wu, "Interpreting Graph Inference with Skyline Explanations," ICDE (2026).

[7] Y. Zhang, S. B. Yang, A. Khan, and C. G. Akcora, "ATEX-CF: Attack-Informed Counterfactual Explanations for Graph Neural Networks," ICLR (2026).

[8] S. Saxena, A. Khan, and J. Chandra, "NAEx: A Plug-and-Play Framework for Explaining Network Alignment," CoRR 2508.04731 (2025).

[9] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A Comprehensive Survey on Graph Neural Networks," IEEE Trans. Neural Networks Learn. Syst. 32(1) (2021), 4-24.

[10] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNExplainer: Generating Explanations for Graph Neural Networks," NeurIPS (2019), 9240-9251

[11] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang, "Parameterized Explainer for Graph Neural Network," NeurIPS (2020).

[12] H. Yuan, H. Yu, J. Wang, K. Li, and S. Ji, "On Explainability of Graph Neural Networks via Subgraph Explorations," ICML (2021) 12241-12252.

[13] M. N. Vu and M. T. Thai, "PGM-Explainer: Probabilistic Graphical Model Explanations for Graph Neural Networks," NeurIPS (2020).

[14] Q. Huang, M. Yamada, Y. Tian, D. Singh, and Y. Chang, "GraphLIME: Local Interpretable Model Explanations for Graph Neural Networks," IEEE Trans. Knowl. Data Eng. 35(7) (2023) 6968-6972.

[15] Z. Huang, M. Kosan, S. Medya, S. Ranu, and A. K. Singh, "Global Counterfactual Explainer for Graph Neural Networks," WSDM (2023) 141-149.

[16] 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, " WWW (2022) 1018-1027.

[17] T. Schnake, O. Eberle, J. Lederer, S. Nakajima, K. T. Schütt, K.-R. Müller, and G. Montavon, "Higher-Order Explanations of Graph Neural Networks via Relevant Walks," IEEE Trans. Pattern Anal. Mach. Intell. 44(11) (2022) 7581-7596.

[18] A. Khan and E. B. Mobaraki, "Interpretability Methods for Graph Neural Networks," DSAA (2023) 1-4.

[19] H. Yuan, H. Yu, S. Gui, and S. Ji, "Explainability in Graph Neural Networks: A Taxonomic Survey," IEEE Trans. Pattern Anal. Mach. Intell. 45(5) (2023) 5782-5799.

[20] J. Kakkad, J. Jannu, K. Sharma, C. Aggarwal, and S. Medya, "A Survey on Explainability of Graph Neural Networks," IEEE Data Eng. Bull. 47(2) (2023) 35-63.