GNN Explainers 2.0: User-centric and Data-driven Insights

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

Xiangyu Ke
Zhejiang University
Zhejiang, China
xiangyu.ke@zju.edu.cn

Yinghui Wu
Case Western Reserve University
Cleveland, Ohio, USA
yxw1650@case.edu

Francesco Bonchi
CENTAI Institute
Turin, Italy
bonchi@centai.eu

Abstract

Graph neural networks (GNNs) are deep learning models designed for graph-structured data that have achieved strong results across domains–social networks, knowledge graphs, bioinformatics, transportation, World Wide Web, and finance–on tasks such as node and graph classification, link prediction, entity resolution, question answering, recommendation, and fraud detection. Explaining the decisions of high-performing, yet "black-box" GNNs remains both challenging and essential. The initial five years have produced tremendous progress with many GNN explainers (e.g., GNNExplainer, PGExplainer, SubgraphX, PGMExplainer, GraphLime, GCFExplainer, CF2, GNN-LRP) that identify the influential nodes, edges, subgraphs, and features aiming to explain the output of GNNs.

We refer to those works as GNN Explainers 1.0, since they provide one-time, final-output explanations and are focused on narrow tasks like node or graph classification, which limits their usefulness for broader, user-centered needs. Practical debugging and accountability require robust, multi-faceted, and GNN's layer-wise provenance so that data scientists can trace how inputs transform through layers and locate where errors occur. Non-technical stakeholders need explanations that are accessible, configurable, and queryable through familiar interfaces–structured queries, ad-hoc instructions, counterfactual evidence, or natural language–so both experts and non-experts can interactively explore model behavior.

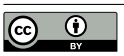
This tutorial surveys latest advances in user-centered GNN explanations that shift focus from merely explaining model outputs to producing actionable, end-user-facing explanations. We show how data mining principles can improve comprehension, usability, and trust, and outline practical strategies for creating configurable, interpretable explanations tailored to diverse stakeholders. We refer to this paradigm as GNN Explainers 2.0. We demonstrate key works under this paradigm, summarize open challenges, and highlight opportunities for the web and data mining community.

CCS Concepts

• **Computing methodologies** → **Neural networks**; • **Information systems** → **Graph-based database models**.

Keywords

Graph Neural Networks, Explainable AI, User-friendly Explanations



WSDM '26, Boise, ID, USA

ACM ISBN 979-8-4007-2292-9/2026/02
https://doi.org/10.1145/3773966.3777915

ACM Reference Format:
Arijit Khan, Xiangyu Ke, Yinghui Wu, and Francesco Bonchi. 2026. GNN Explainers 2.0: User-centric and Data-driven Insights. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining (WSDM '26), February 22–26, 2026, Boise, ID, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3773966.3777915

1 Motivation

Deep learning's rapid proliferation across domains has driven equal growth in explainable AI to address the field's black-box nature. Although deep models often deliver state-of-the-art results, their complexity raises concerns about fairness, transparency, accountability, privacy, security, and ethics. Safe, trustworthy deployment therefore requires not only SOTA performance but also clear, human-intelligible explanations for data scientists and non-machine-learning stakeholders such as biologists, chemists, social scientists, journalists, and policymakers [3, 48]. Explainable AI has surged with the rise of powerful AI/ML models and growing reliance on AI across industries and data science, driving interest in fairness, accountability, and transparency (FAT) in algorithmic decision-making systems[1]. Major data mining and web venues now feature XAI events such as Responsible AI Day (KDD 2025), Responsible AI Engineering (WWW 2025), Time Series XAI tutorial & Trust and Responsibility in Recommendation Systems workshop (WSDM 2025), Human-centric AI & Trustworthy Knowledge Discovery workshops (CIKM 2025), and a Responsible AI panel (ECML-PKDD 2025). Numerous recent contributions from the data and web communities have further advanced the field [13, 14, 17, 20, 39, 50].

Graph neural networks are a prominent class of deep learning models for graph-structured data [46]. While many surveys and benchmarks have examined GNN explainability methods [4, 23, 27, 28, 30, 32, 37, 41, 43, 51], relatively few tutorials explored the topic [21, 22]. This tutorial offers a distinct, data- and algorithm-centered perspective on state-of-the-art XAI techniques that incorporates user-centric requirements without limiting the scope to HCI approaches [16, 19]. We situate related work within the data and web community, highlight data-driven opportunities for more usable XAI, and aim to stimulate interdisciplinary research that advances explainability for real-world data challenges.

2 Planned Format

The proposed lecture-style tutorial includes some foundational elements (e.g., a categorization of the types of explanations and algorithmic approaches) and some (but not exhaustive) survey elements, focusing on recent work about usable XAI systems for graph neural networks. Additionally, we cover selected demo papers (e.g., [12, 44]) based on data-driven approaches, leading to hands-on

[1] FAccT'25. https://dl.acm.org/doi/proceedings/10.1145/3715275

components, where the audience will explore real-world XAI use cases through live system demonstrations.

3 Target Audience

This tutorial is designed for students, researchers, practitioners, and policymakers wishing to understand the state-of-the-art in explainable and responsible AI from a data-driven and user-centric perspective. Prerequisites include data mining, machine learning, and ideally some familiarity with graph neural networks.

4 Previous Offering and Related Tutorials

We have not yet presented the proposed tutorial in any conference. There are plethora of tutorials on XAI for general AI/ML and deep learning models at data mining and management venues [31, 34, 36, 45, 48] as well as in AI conferences [10, 25]. However, XAI for GNNs has been comparatively underrepresented in recent tutorials [21, 22]. We propose an XAI tutorial for GNNs with a unique twist: we take a user-friendly and data-driven approach, moving beyond "explanation of models" intended mainly for AI practitioners to data-driven "explanations for users", broadening the explanation scope to serve non-technical stakeholders (e.g., policymakers, end-users, or domain experts such as doctors and financial analysts) who need intuitive, actionable, and trustworthy AI insights. This point of view motivates us to present XAI methods that focus on bridging the gap between AI outputs and human expertise using data- and algorithm-centered approaches.

5 Tutorial Scope and Structure

The intended length of our tutorial is half-day (3 hours + breaks).

5.1 Outline

1 Introduction (30 minutes)
  1.1 GNNs and Applications
  1.2 XAI for GNNs
2 GNN Explainers Categorization (15 minutes)
3 User-centric XAI (15 minutes)
4 User-centric and Data-driven XAI Methods for GNNs (90 minutes)
  4.1 Pattern Mining and Concept Hierarchies
  4.2 Counterfactual Explanations
  4.3 Explanation by Examples and Rules
  4.4 Natural Language Explanations
  4.5 Declarative Explanatory Queries
  4.6 Robust Explanations
  4.7 Multi-criteria Explanations
  4.8 Efficiency and Interactiveness
  4.9 XAI beyond Classification
5 Future Directions (30 minutes)

5.2 Introduction

GNNs [46] are a mature class of deep learning models that extend traditional neural networks to transform graphs into embedding representations for various downstream tasks in an end-to-end manner. Most GNNs implement a multi-layer message-passing paradigm in which each layer updates a node's representation from its neighbors' representations. Representative variants include graph convolutional networks (GCNs), graph attention networks (GATs), graph isomorphism networks (GINs), APPNP, and GraphSAGE. They have been employed for graph and node classification, link prediction, entity resolution, question answering, graph alignment, and combinatorial optimization problems.

With GNNs deployed across diverse applications, explainability techniques–initiated by GNNExplainer in 2019–have rapidly proliferated to address their black-box nature. Here, we briefly introduce the most prominent methods developed in the past five years, including GNNExplainer, PGExplainer, GraphMask, SubgraphX, PGMExplainer, RelEx, GraphLime, RCExplainer, GCFExplainer, CF2, GNN-LRP, XGNN, ProtGNN, and others [21, 49–51]. We will discuss their usefulness in deriving insights about the model and data, as well as recurring challenges such as complex data, bias, redundant evidence, weak or misaligned GNN models, and the frequent absence of ground truth for explanations [17].

5.3 GNN Explainers Categorization

GNN explainability methods are categorized by their design and the type of explanations they produce [51]. Intrinsic methods incorporate interpretability into the GNN architecture (e.g., graph attention networks), while post-hoc methods fit a separate explainer to an already trained model. Explainability can be global–revealing model-wide structures, parameters, or prototypical graph patterns, or local–explaining individual predictions. Forward, model-agnostic approaches recover evidence via perturbation or surrogate models, whereas backward, model-specific approaches attribute importance through gradients or decomposition. Factual explanations identify subgraphs that preserve the model's output, and counterfactual explanations find subgraphs whose removal flips the prediction. A surrogate model is a simple, interpretable model trained to approximate the input-output behavior of a complex GNN. Finally, mechanistic interpretability probes internal model components such as neurons, while observational explanations are data-driven and typically model-agnostic, inferring model logic by observing outputs to varied inputs.

5.4 User-centric Paradigm in XAI

While SOTA GNN explainability methods offer valuable insights, they exhibit several key limitations on usability aspects: (1) They are not user-centric, emphasizing numerical logits or feature scores over intuitive, domain-specific structures that align with end-user needs, making explanations hard to interpret, access, and adapt for downstream use. (2) They lack robustness, with explanations that change drastically under slight graph perturbations, undermining reliability. (3) They typically optimize a single pre-selected metric such as fidelity or conciseness, producing biased and incomplete views of model behavior. (4) They provide static, one-off explanations of final outputs rather than progressive, layer-by-layer accounts, limiting understanding of internal reasoning and opportunities for targeted debugging and improvement. (5) They focus primarily on node and graph classification, leaving a gap in explainability techniques for the broader range of GNN tasks. (6) They do not support interactive, user-driven exploration via natural language or high-level structural queries, nor do they routinely deliver explanations as natural-language descriptions or illustrative examples preferred by users.

What makes an explanation acceptable to users and how explanations affect user perceptions and actions remain open questions. However, HCI, cognitive science, and social psychology suggest

important criteria for high-quality explanations: (i) Right context, meaning explanations should be tailored to different stakeholders [15]; (ii) right quantity, i.e., concise and informative without overload; (iii) comprehension, favoring higher-level patterns, rules, contrastive examples, and concepts to reduce cognitive load and increase trust; (iv) usefulness, by offering actionable algorithmic recourse; (v) interactivity, enabling queryable rather than static explanations; and (vi) stability, providing consistent explanations for similar inputs. We will use such principles to discuss usable explanations leveraging data mining and algorithmic approaches.

5.5 User-centric and Data-driven XAI Methods

This tutorial reviews recent advances in user-centric, data-driven GNN explainability methods–termed GNN Explainers 2.0.

Pattern Mining and Concept Hierarchies. GVEX [13] introduces a two-tier explanation framework for GNN classification: lower-tier subgraphs provide factual and counterfactual evidence for predictions, while higher-tier patterns abstract these subgraphs into common motifs to support efficient search, exploration, and domain alignment. GNN-Dissect [47] maps GNN neurons to interpretable concepts–logical compositions of node and neighborhood properties–yielding model-level explanations that expose functional groups driving outcomes such as mutagenicity, thereby enhancing transparency. GRAPHSHAP [35], a Shapley-based approach, generates motif-level explanations for identity-aware graph classifiers by assigning importance scores to motifs defined by domain experts or extracted from data.

Counterfactual Explanations. Counterfactual (CF) explanations [20] offer actionable recourse for fairness and interpretability by describing how alternative outcomes could arise if some premises were different, with node- and graph-level CF studied in [1, 2] and [39], respectively. Lanciano et al. [24] design contrast subgraphs–node sets whose induced subgraph is dense in one class and sparse in another–to produce transparent and self-explanatory classifiers.

Explanation by Examples and Rules. GLGEXPLAINER [7] generates global explanations for GNNs by expressing subgraph-level concepts as Boolean formulas. GraphTrail [5] leverages Shapley values to identify discriminative subgraph concepts and employs symbolic regression to translate predictions into human-interpretable Boolean rules. GnnXemplar [6] selects representative nodes in the embedding space and derives natural-language rules from their neighborhoods to clarify model predictions.

Natural Language Explanations. Large language model (LLM)-powered methods are increasingly being used to generate natural language explanations enhancing GNN interpretability [8, 18, 33].

Declarative Explanatory Queries. Exploring GNN inference with explanations often involves ad-hoc queries–for example, requesting a factual explanation at an intermediate layer with constraints on subgraph size–which are cumbersome to implement manually. This motivates a declarative framework that enables users to express and customize explanatory queries through a rich logic and query language. SliceGXQ [44] realizes this vision as an end-to-end, SPARQL-like system supporting interactive, layer-wise explanatory queries for GNNs.

Robust Explanations. Recent advances in robust explainable GNNs integrate interpretability with adversarial defense to preserve explanation quality under worst-case perturbations. XGNNCert [26] delivers the first certifiable robustness guarantee for graph-level tasks, ensuring stable explanations without compromising predictive performance at the node and edge level. k-RCW [40] introduces resilient counterfactual witnesses that remain valid under structural disturbances, while additional methods for robust counterfactual explanations have been investigated in [9].

Multi-criteria Explanations. Multiple metrics evaluate explanation quality–fidelity, sparsity, stability, and so on–each measuring a different aspect, so optimizing a single metric can produce biased, less diverse explanations; besides metrics also often conflict (e.g., higher fidelity usually requires lower sparsity). Pareto-optimality or skyline queries address such multi-criteria trade-offs. Recent work defines and computes explanation skylines to produce high-quality, diversified explanations across multiple metrics [29, 38].

Efficiency and Interactiveness. Parallel and streaming algorithms and indexing techniques have been developed to scale and speed up GNN explanation methods [13, 39, 40]. XAI systems should combine scalable back-end methods with intuitive interfaces that support interaction, exploration, declarative querying, and personalization to make explanations both efficient and usable. For example, GVEX [13] offers a tunable component that lets users select a desired number of important nodes from different classes to produce class-targeted explanations.

XAI beyond Classification. XAI methods are increasingly applied to explain complex AI/ML outcomes beyond classification. Examples include EAGER [11] which uses feature scoring to surface knowledge-graph facts important to neural question-answering systems. SliceGX identifies explanatory subgraphs at each GNN layer to aid model diagnosis and architecture optimization [52]. NAEx is a framework for explaining GNN-based network alignment [42].

5.6 Conclusion and Open Problems

We conclude the tutorial by encouraging the data mining and web community to pursue the following open problems.

Actionable Recourse. While current methods can generate explanations that capture GNN behavior on graph data, they often face out-of-distribution challenges that yield impractical results. In drug discovery, for instance, a highlighted molecular fragment may be nonexistent or unstable in laboratory conditions. This underscores the need for explanations that are not only accurate but also contextually relevant, practical, and interpretable.

Explanation with Privacy Concern. Although stakeholders increasingly demand greater explainability, they often restrict access to datasets or GNN models due to business and privacy concerns. Explanations therefore require privacy-preserving methods that balance interpretive quality with data protection.

Multi-modal Explanation. In many domains, e.g., healthcare and spatio-temporal, data are multi-modal, spanning text, images, tables, and multimedia. Graph-based models, including knowledge graphs, provide integrated solutions for such complexity, and recent GNN-LLM collaborations can generate outputs across modalities. Yet, multi-modal explainable AI remains largely underexplored.

Explanation for Advanced GNNs. The growing complexity of advanced GNN architectures–such as graph transformers, agent-based models, and graph foundation models–demands explainability approaches that extend beyond traditional message passing. While these models offer greater expressiveness and generalization, their intricacy challenges current XAI techniques. Advancing the field will require scalable, faithful, and generalizable explanation frameworks tailored to next-generation GNNs.

References

[1] Carlo Abrate and Francesco Bonchi. 2021. Counterfactual Graphs for Explainable Classification of Brain Networks. In *KDD*. 2495–2504.

[2] Carlo Abrate, Giulia Preti, and Francesco Bonchi. 2023. Counterfactual Explanations for Graph Classification Through the Lenses of Density. In *xAI*. 324–348.

[3] Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. 2022. OpenXAI: Towards a Transparent Evaluation of Model Explanations. In *NeurIPS*.

[4] Chirag Agarwal, Owen Queen, Himabindu Lakkaraju, and Marinka Zitnik. 2023. Evaluating Explainability for Graph Neural Networks. *Sci Data* 10, 1 (2023).

[5] Burouj Armgaan, Manthan Dalmia, Sourav Medya, and Sayan Ranu. 2024. GraphTrail: Translating GNN Predictions into Human-Interpretable Logical Rules. In *NeurIPS*.

[6] Burouj Armgaan, Eshan Jain, Harsh Pandey, Mahesh Chandran, and Sayan Ranu. 2025. GnnXemplar: Exemplars to Explanations - Natural Language Rules for Global GNN Interpretability. In *NeurIPS*.

[7] Steve Azzolin, Antonio Longa, Pietro Barbiero, Pietro Liò, and Andrea Passerini. 2023. Global Explainability of GNNs via Logic Combination of Learned Concepts. In *ICLR*.

[8] Peyman Baghershahi, Gregoire Fournier, Pranav Nyati, and Sourav Medya. 2025. From Nodes to Narratives: Explaining Graph Neural Networks with LLMs and Graph Context. *CoRR* abs/2508.07117 (2025).

[9] Mohit Bajaj, Lingyang Chu, Zi Yu Xue, Jian Pei, Lanjun Wang, Peter Cho-Ho Lam, and Yong Zhang. 2021. Robust Counterfactual Explanations on Graph Neural Networks. In *NeurIPS*. 5644–5655.

[10] Oana-Maria Camburu and Zeynep Akata. 2021. Natural-XAI: Explainable AI with Natural Language Explanations. In *Tutorial@ICML*.

[11] Andrew Chai, Alireza Vezvaei, Lukasz Golab, Mehdi Kargar, Divesh Srivastava, Jaroslaw Szlichta, and Morteza Zihayat. 2023. EAGER: Explainable Question Answering Using Knowledge Graphs. In *Joint Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)*. 4:1–4:5.

[12] Tingyang Chen, Dazhuo Qiu, Yinghui Wu, Arijit Khan, Xiangyu Ke, and Yunjun Gao. 2024. User-friendly, Interactive, and Configurable Explanations for Graph Neural Networks with Graph Views. In *SIGMOD/PODS*. 512–515.

[13] Tingyang Chen, Dazhuo Qiu, Yinghui Wu, Arijit Khan, Xiangyu Ke, and Yunjun Gao. 2024. View-based Explanations for Graph Neural Networks. *Proc. ACM Manag. Data* 2, 1 (2024), 40:1–40:27.

[14] Vargha Dadvar, Lukasz Golab, and Divesh Srivastava. 2023. POEM: Pattern-Oriented Explanations of Convolutional Neural Networks. *PVLDB* 16, 11 (2023), 3192–3200.

[15] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael J. Muller, and Mark O. Riedl. 2024. The Who in XAI: How AI Background Shapes Perceptions of AI Explanations. In *CHI*. 316:1–316:32.

[16] Upol Ehsan, Philipp Wintersberger, Elizabeth Anne Watkins, Carina Manger, Gonzalo A. Ramos, Justin D. Weisz, Hal Daumé III, Andreas Riener, and Mark O. Riedl. 2023. Human-Centered Explainable AI (HCXAI): Coming of Age. In *CHI*. 353:1–353:7.

[17] Lukas Faber, Amin K. Moghaddam, and Roger Wattenhofer. 2021. When Comparing to Ground Truth is Wrong: On Evaluating GNN Explanation Methods. In *KDD*. 332–341.

[18] Flavio Giorgi, Cesare Campagnano, Fabrizio Silvestri, and Gabriele Tolomei. 2025. Natural Language Counterfactual Explanations for Graphs Using Large Language Models. In *AISTATS*. 3565–3573.

[19] Pantea Habibi, Peyman Baghershahi, Sourav Medya, and Debaleena Chattopadhyay. 2024. Design Requirements for Human-Centered Graph Neural Network Explanations. *CoRR* abs/2405.06917 (2024).

[20] Zexi Huang, Mert Kosan, Sourav Medya, Sayan Ranu, and Ambuj Singh. 2023. Global Counterfactual Explainer for Graph Neural Networks. In *WSDM*. 141–149.

[21] Arijit Khan and Ehsan Bonabi Mobaraki. 2023. Interpretability Methods for Graph Neural Networks. In *DSAA*. 1–4.

[22] Megha Khosla and Luis Galárraga. 2023. Explainable Graph Machine Learning. In *Tutorial@ECML-PKDD*.

[23] Mert Kosan, Samidha Verma, Burouj Armgaan, Khushbu Pahwa, Ambuj K. Singh, Sourav Medya, and Sayan Ranu. 2024. GNNX-BENCH: Unravelling the Utility of Perturbation-based GNN Explainers through In-depth Benchmarking. In *ICLR*.

[24] Tommaso Lanciano, Francesco Bonchi, and Aristides Gionis. 2020. Explainable Classification of Brain Networks via Contrast Subgraphs. In *KDD*. 3308–3318.

[25] Freddy Lecue, Pasquale Minervini, Riccardo Guidotti, and Fosca Giannotti. 2023. On Explainable AI: From Theory to Motivation, Industrial Applications, XAI Coding & Engineering Practices. In *Tutorial@AAAI*.

[26] Jiate Li, Meng Pang, Yun Dong, Jinyuan Jia, and Binghui Wang. 2025. Provably Robust Explainable Graph Neural Networks against Graph Perturbation Attacks. In *ICLR*.

[27] Xuyan Li, Jie Wang, and Zheng Yan. 2025. Can Graph Neural Networks be Adequately Explained? A Survey. *ACM Comput. Surv.* 57, 5, Article 131 (2025).

[28] Meng Liu, Youzhi Luo, Limei Wang, Yaochen Xie, Hao Yuan, Shurui Gui, Haiyang Yu, Zhao Xu, Jingtun Zhang, Yi Liu, Keqiang Yan, Haoran Liu, Cong Fu, Bora Oztekin, Xuan Zhang, and Shuiwang Ji. 2021. DIG: A Turnkey Library for Diving into Graph Deep Learning Research. *J. Mach. Learn. Res.* 22 (2021), 240:1–240:9.

[29] Yifei Liu, Chao Chen, Yazheng Liu, Xi Zhang, and Sihong Xie. 2021. Multi-objective Explanations of GNN Predictions. In *ICDM*. 409–418.

[30] Antonio Longa, Steve Azzolin, Gabriele Santin, Giulia Cencetti, Pietro Lio, Bruno Lepri, and Andrea Passerini. 2025. Explaining the Explainers in Graph Neural Networks: a Comparative Study. *ACM Comput. Surv.* 57, 5 (2025), 120:1–120:37.

[31] Xuan Luo and Jian Pei. 2024. Applications and Computation of the Shapley Value in Databases and Machine Learning. In *SIGMOD*.

[32] Corrado Monti, Paolo Bajardi, Francesco Bonchi, André Panisson, and Alan Perotti. 2024. A True-to-the-model Axiomatic Benchmark for Graph-based Explainers. *Trans. Mach. Learn. Res.* (2024).

[33] Bo Pan, Zhen Xiong, Guanchen Wu, Zheng Zhang, Yifei Zhang, Yuntong Hu, and Liang Zhao. 2025. GraphNarrator: Generating Textual Explanations for Graph Neural Networks. In *ACL*. 23–42.

[34] Eliana Pastor, Eleonora Poeta, André Panisson, Alan Perotti, and Gabriele Ciravegna. 2025. Beyond Input Attribution: A Hands-On Tutorial to Concept-Based Explainable AI and Mechanistic Interpretability. In *KDD*. 6247–6248.

[35] Alan Perotti, Paolo Bajardi, Francesco Bonchi, and André Panisson. 2023. Explaining Identity-aware Graph Classifiers through the Language of Motifs. In *IJCNN*. 1–8.

[36] Romila Pradhan, Aditya Lahiri, Sainyam Galhotra, and Babak Salimi. 2022. Explainable AI: Foundations, Applications, Opportunities for Data Management Research. In *SIGMOD*.

[37] Mario Alfonso Prado-Romero, Bardh Prenkaj, Giovanni Stilo, and Fosca Giannotti. 2024. A Survey on Graph Counterfactual Explanations: Definitions, Methods, Evaluation, and Research Challenges. *ACM Comput. Surv.* 56, 7, Article 171 (2024).

[38] Dazhuo Qiu, Haolai Che, Arijit Khan, and Yinghui Wu. 2026. Interpreting Graph Inference with Skyline Explanations. In *ICDE*.

[39] Dazhuo Qiu, Jinwen Chen, Arijit Khan, Yan Zhao, and Francesco Bonchi. 2025. Finding Counterfactual Evidences for Node Classification. In *KDD*. 2362–2373.

[40] Dazhuo Qiu, Mengying Wang, Arijit Khan, and Yinghui Wu. 2024. Generating Robust Counterfactual Witnesses for Graph Neural Networks. In *ICDE*.

[41] Benjamin Sanchez-Lengeling, Jennifer Wei, Brian Lee, Emily Reif, Peter Y. Wang, Wesley Wei Qian, Kevin McCloskey, Lucy Colwell, and Alexander Wiltschko. 2020. Evaluating Attribution for Graph Neural Networks. In *NeurIPS*.

[42] Shruti Saxena, Arijit Khan, and Joydeep Chandra. 2025. NAEx: A Plug-and-Play Framework for Explaining Network Alignment. arXiv:2508.04731

[43] Kenneth Teo Tian Shun, Eko Edita Limanta, and Arijit Khan. 2020. An Evaluation of Backpropagation Interpretability for Graph Classification with Deep Learning. In *IEEE BigData*. 561–570.

[44] Haitong Tang, Yinghui Wu, Arijit Khan, Tingting Zhu, Tingyang Chen, and Xiangyu Ke. 2025. Declarative Explanations for Graph Neural Networks: A Demonstration. https://github.com/Hai0709/SliceGXQ.

[45] Cong Wang, Xiao-Hui Li, Haocheng Han, Shendi Wang, Luning Wang, Caleb Chen Cao, and Lei Chen. 2021. Counterfactual Explanations in Explainable AI: A Tutorial. In *KDD*. 4080–4081.

[46] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. 2021. A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems* 32, 1 (2021), 4–24.

[47] Han Xuanyuan, Pietro Barbiero, Dobrik Georgiev, Lucie Charlotte Magister, and Pietro Liò. 2023. Global Concept-Based Interpretability for Graph Neural Networks via Neuron Analysis. In *AAAI*. 10675–10683.

[48] Zhou Yang, Ninghao Liu, Xia Ben Hu, and Fang Jin. 2022. Tutorial on Deep Learning Interpretation: A Data Perspective. In *CIKM*. 5156–5159.

[49] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. GNNExplainer: Generating Explanations for Graph Neural Networks. In *NeurIPS*. 9240–9251.

[50] Hao Yuan, Jiliang Tang, Xia Hu, and Shuiwang Ji. 2020. XGNN: Towards Model-Level Explanations of Graph Neural Networks. In *KDD*. 430–438.

[51] Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. 2023. Explainability in Graph Neural Networks: A Taxonomic Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 5 (2023), 5782–5799.

[52] Tingting Zhu, Tingyang Chen, Yinghui Wu, Arijit Khan, and Xiangyu Ke. 2025. SliceGX: Layer-wise GNN Explanation with Model-slicing. arXiv:2506.17977