Explainable Graph Neural Network Recommenders; Challenges and Opportunities

Amir Reza Mohammadi
Universität Innsbruck
Innsbruck, Austria
amir.reza@uibk.ac.at

ABSTRACT

Graph Neural Networks (GNNs) have demonstrated significant potential in recommendation tasks by effectively capturing intricate connections among users, items, and their associated features. Given the escalating demand for interpretability, current research endeavors in the domain of GNNs for Recommender Systems (RecSys) necessitate the development of explainer methodologies to elucidate the decision-making process underlying GNN-based recommendations. In this work, we aim to present our research focused on techniques to extend beyond the existing approaches for addressing interpretability in GNN-based RecSys.

ACM Reference Format:
Amir Reza Mohammadi. 2023. Explainable Graph Neural Network Recommenders; Challenges and Opportunities . In *Seventeenth ACM Conference on Recommender Systems (RecSys '23), September 18–22, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3604915.3608875

1 INTRODUCTION

The increasing need for machine learning models that can effectively handle and process graph-structured data, prevalent in various practical domains like social networks, knowledge graphs, and molecular graphs, has fostered the emergence and advancement of GNNs. GNNs have demonstrated impressive potential in recommendation tasks by understanding intricate connections between users, items, and their characteristics [18, 46]. They are particularly well-suited for modeling recommendation tasks as they can handle large and sparse graphs, which are typical in recommendation scenarios [52].

The success of GNN-based recommenders can be attributed to three main factors: the utilization of structural data, the incorporation of high-order connectivity, and the exploitation of supervision signals [11]. (1) Structural data factor that is encompassing diverse information such as user-item interactions, user profiles, and item attributes, presents a challenge for traditional RecSys that often focus on specific data sources, resulting in suboptimal performance. GNNs provide a unified approach by representing the data as nodes and edges on a graph, enabling comprehensive utilization of available data. In contrast to images and texts, graphs deviate from grid-like data structures and encompass significant structural information. Notably, subgraphs serve as elementary components within graphs and exhibit a strong association with graph functionalities, rendering them valuable tools for graph explanation [56]. (2) The consideration of high-order connectivity is essential for accurate recommendations. The collaborative filtering effect, where the preferences of users with similar tastes impact recommendations, is crucial but typically overlooked in traditional approaches, which primarily rely on directly connected items. In contrast, GNN-based models effectively capture high-order connectivity by representing the collaborative filtering effect as multi-hop neighbors on the graph, integrating it into the learned representations through embedding propagation and aggregation. The preservation of high-order connectivity data also proves beneficial for comprehensively explaining the graph at the model-level (see section 2). (3) Sparse supervision signals pose a challenge in RecSys where GNN-based models address this by leveraging semi-supervised signals in the representation learning process. By encoding non-target behaviors (e.g., search, add to cart) as semi-supervised signals over the graph, GNNs improve recommendation performance [11].

However, recent research has indicated that GNNs encounter a similar problem to other deep neural networks, namely their susceptibility to adversarial attacks [4, 5, 30, 50, 59]. More specifically, attackers can create graph adversarial perturbations by manipulating the graph's structure or node characteristics in order to mislead the GNN model and cause it to produce inaccurate predictions [5, 58]. The enhanced interpretability is thought to provide a feeling of assurance by engaging humans in the decision-making procedure [35, 54]. Nevertheless, due to its reliance on data, interpretability itself is prone to potential malicious manipulations [11]. It should be noted that there are additional endeavors focused on linking these two subjects, whereby interpretation methods on non-graph-structured data are targeted for attack [13, 19, 27, 57].

The explanation of deep learning (DL) models on text or images has received significant attention [37, 42], but explaining DL models on graphs remains relatively unexplored. This is a more difficult task due to several reasons [55]. Firstly, the adjacency matrix, which represents the graph's topology, consists of discrete values that cannot be directly optimized using gradient-based methods [6]. Secondly, in certain domains, a graph is considered valid only if it adheres to a set of specific rules, making the generation of a valid explanatory graph for underlying decision-making processes of GNNs a complex endeavor. Lastly, graph data structures are heterogeneous, containing various types of node and edge features, thereby making the development of a universal explanation method for GNNs even more challenging. The Shapley value [39], a Game

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
RecSys '23, September 18–22, 2023, Singapore, Singapore
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0241-9/23/09.
https://doi.org/10.1145/3604915.3608875

Theory technique that describes the equitable allocation of aggregate gains among players based on their individual contributions, has been employed in addressing the first and third challenges under consideration. This approach has been expanded to elucidate the predictions of machine learning models on tabular data [6] by considering each attribute of the explained instance as a player engaged in a game, where the prediction represents the corresponding payout.

Through an in-depth analysis of pertinent literature in the domains of GNNs and explainable RecSys, we have discerned notable deficiencies in GNN-based recommendation methods. In the context of the forthcoming doctoral research project (currently in its first year), we aim to address the following crucial research gaps that we have identified:

(1) There is a need to explore the vulnerabilities of explainable GNNs to adversarial attacks specifically targeted at explanations [58] considering the employing of GNNs in security-critical applications [49, 50] . Current research on adversarial attacks in GNNs focuses mainly on the input data or the model's predictions, but there is limited understanding of how adversarial attacks can manipulate or deceive the explanations generated by explainable GNNs [8].
(2) There is a lack of research on developing efficient algorithms for subgraph extraction that is one of the leading explanation methods for GNNs [29, 53, 56]. Existing approaches often rely on exhaustive search or heuristics, which can be computationally expensive and limit scalability [53]. Finding innovative techniques to extract important subgraphs efficiently without sacrificing the quality of explanations is an important research challenge.
(3) There is a research gap in evaluating the reliability of local or single-instance explanations [41, 44]. It is crucial to develop metrics or methods to assess the trustworthiness and consistency of local explanations and move toward global and concept-based explanations [31].
(4) The Shapley value [39] exhibits considerable potential as a methodology for assessing the individual contributions of various components within a graph towards predictions, primarily relying on its fundamental characteristics [6, 56]. However, there exists an unaddressed disparity in the utilization of more robust variants of this approach [14, 16].

In summary, while GNN-based RecSys have shown promising results, there are still several challenges that need to be addressed. Developing more efficient GNN architectures and leveraging explainable AI techniques can help bridge this gap and pave the way for more effective and interpretable recommendation systems.

2 BACKGROUND

Within this section, we delve into significant studies pertaining to the acquisition of GNNs, alongside the domain of explainable recommendation. Given that there exist distinct prior studies pertinent to each research question, we will delve into them individually on next section. However, this section primarily focuses on a more comprehensive overview of the background work to the topic.

The objective of explanation techniques for deep models is to investigate the intrinsic connections that underlie the predictions generated by such models. These techniques aim to offer explanations that are interpretable to human understanding, thereby enhancing the trustworthiness of deep models. Depending on the nature of the explanations provided, these techniques can be broadly classified into two primary categories: instance-level methods and model-level methods [48]. Instance-level techniques offer explanations that are specific to each input graph, providing input-dependent explanations. When presented with an input graph, these methods elucidate the workings of deep models by identifying the crucial input features that contribute to its prediction. A prominent example is the model-agnostic technique GNNExplainer [53]. This approach endeavors to maximize the mutual information between the distribution of potential subgraphs and the predictions of a GNN, thereby identifying the subgraph that exerts the most substantial influence on the prediction. However, a limitation of GNNExplainer is its requirement for retraining in each prediction scenario. Moreover, although GNNExplainer strives to generate global explanations for a specific class by employing graph alignment on the subgraph explanations using ten instances of the class, the efficacy of confirming the global explanation is constrained by the computational efficiency of graph alignment, an NP-Hard problem [53]. To tackle these challenges, PGExplainer [29] aims to address the drawbacks associated with GNNExplainer. PGExplainer is a model-agnostic explanation method that shares the same optimization objective, albeit with a fundamental distinction in its utilization of a deep neural network (DNN) to parameterize the process of generating explanations. Although PGExplainer claims to furnish global explanations, it is important to note that these explanations are not genuinely global but rather pertain to multi-instance explanations. Similarly, PGM-Explainer [44] employs the extraction of pertinent subgraphs for a given prediction, with the additional advantage of indicating feature dependencies through conditional probabilities. Diverging from the approach taken by instance-level methods, model-level methods concentrate on offering broad insights and a high-level comprehension to elucidate deep graph models. Their main objective is to investigate the types of input graph patterns that result in specific behaviors exhibited by GNNs, such as the maximization of a target prediction. Input optimization [42] techniques have been widely explored as a means to attain model-level explanations for image classifiers. However, directly applying these techniques to graph models poses challenges due to the discrete nature of graph topology information, rendering the task of explaining GNNs at the model-level considerably more complex. Consequently, this area remains important but relatively under-explored in current research. To the best of our knowledge, the extant literature comprises solely two model-level approaches for explicating GNNs, namely XGNN [54] and GNNInterpreter [47]. XGNN presents a methodology for explaining GNNs by means of graph generation. Rather than directly optimizing the input graph, it trains a graph generator to generate graphs that maximize a specified target graph prediction. These generated graphs are then considered as explanations for the target prediction, expected to encompass distinctive graph patterns. On the other hand, GNNInterpreter employs a numerical optimization technique to acquire the explanation graph through continuous relaxation. In the subsequent section, specifically in RQ2, we will delve into our proposed approach aimed at achieving the same objective.

3 RESEARCH OBJECTIVES

To extend the motivation introduced above, our research will seek to address the following research questions:

3.1 RQ1: How can we incorporate adversarial attacks to explain GNN-based Recommender Systems?

Adversarial attacks are a modified input to a machine learning model that is designed to cause the model to produce an incorrect output. The purpose of adversarial attacks is typically to evaluate the robustness of a model [22, 23] or to improve its accuracy under adversarial conditions [32, 40]. Generally, adversarial attacks are not used to make a model more interpretable. That being said, recent research has explored the possibility of using explainability of GNNs for adversarially attacking ML models [8, 26]. We hypothesize that we can look at this scenario from another perspective, which is how to use adversarial attacks to prune noisy nodes or edges of the graph to improve interpretability of the model [21]. One potential approach entails employing targeted adversarial attacks, such as "Nettack" [58] to identify suitable node and edge modifications that lead to class changes, aligning with the objective of providing counterfactual explanations [10, 26, 34]. The particular focus could center on Counterfactual Explanations (CE) pertaining to recommenders, particularly exemplified by the "Prince" [12] algorithm. In this regard, the top-n recommender's output can be treated as the label for the node classification task within the Nettack framework, allowing for a comparative analysis of the outcomes.

Adversarial examples are generated by intentionally adding small perturbations to the original input that are imperceptible to humans but can cause the model to make incorrect predictions. In the context of GNNs, an adversarial attack refers to the process of perturbing the nodes or edges of a graph to generate adversarial examples that can cause the GNN to produce incorrect outputs or predictions. There are several methods for generating adversarial examples for GNNs, including node injection [43], edge perturbation [51], and targeted attacks [22], among others. The goal of these attacks is to identify the most important nodes or edges in the input graph that are responsible for the GNN's output and perturb them in a way that maximizes the model's prediction error. In essence, this is highly similar to well-known GNN explainability approaches [2, 28, 53, 56] that aim for finding effective subgraphs and counterfactual examples [2, 28].

At the current stage of the research, there are some efforts to use adversarial attacks for making more interpretable models. For instance, [1, 20] proposed a method for using adversarial examples to highlight important features in a model. The method involves creating adversarial examples that maximize the difference in output between two subsets of inputs, and then using these examples to identify the features that are most responsible for the difference in output. The authors demonstrated the effectiveness of their method on several image classification tasks. Alvarez-Meris et al. [1] proposed a method for using adversarial attacks to generate explanations for model predictions. The method involves perturbing the input image to generate an adversarial example that causes the model to make a different prediction, and then analyzing the difference between the two predictions to generate an explanation. The authors demonstrated the effectiveness of their method on several image classification tasks. Li et al. [24] proposed a method for visualizing the loss landscape of deep neural networks. The authors use adversarial attacks to generate a large number of input examples and then use these examples to explore the loss landscape. Zugner et al. [58] proposed a method for evaluating the robustness of neural networks for graph data. The authors use adversarial attacks to identify the most important nodes and edges in the input graph and then use this information to improve the model's robustness. This paper proposes a method for regularizing the training of deep neural networks to improve their interpretability. The authors use adversarial attacks to identify the most important features in the input and then use this information to guide the regularization process. These instances are only a small subset of the expanding research investigating the efficacy of adversarial attacks to enhance the interpretability of models. Nevertheless, no such research has been conducted in the RecSys community, and as far as we are aware, there exist no analogous works that specifically account for the unique characteristics of RecSys in this research approach. While the approach is still in its early stages, it has the potential to be a valuable tool for gaining insights into how machine learning models are making their predictions and identifying areas for improvement. Through this RQ we will address the first gap introduced earlier.

3.2 RQ2: Toward concept-based explanation: How reliable are the local explanation methods based on individual instances?

Extensive research has been conducted on explanations at the local level. A recent survey [48] revealed that the majority of these explanations can be classified into six distinct categories: gradient-based methods [3], perturbation-based methods [44, 53, 56], decomposition methods [36], surrogate methods [6, 44], generation-based methods [25, 38], and counterfactual-based methods [28]. As GNN models become more complex and the data being analyzed becomes more varied, it may become more difficult to accurately interpret the model using single-instance explanation methods. There have been studies that shows local or single-instance explanations of machine learning models can be unreliable due to several factors such as context dependency and vulnerability to overfitting [7, 41]. Furthermore, incomplete information is another issue, as local explanations may not consider all relevant information that led to the model's decision and may be susceptible to adversarial attacks, where an attacker deliberately manipulates the input to produce a misleading or incorrect explanation (RQ1) [47]. Therefore, it is crucial to complement local explanations with other types of explanations and consider the limitations of each explanation method when interpreting model decisions.

Concept-based methods have several advantages over local explanations in machine learning interpretability. Firstly, they provide a more holistic view of the model's decision-making process by identifying important concepts or features that influence the model's output [15]. This results in a broader perspective that allows users to better understand the model's behavior. Secondly, concept-based methods are context-independent, making it easier to compare explanations across different instances and generalize insights gained from them [31]. They also produce more comprehensible explanations by identifying important features or concepts that contribute to the model's decision. Lastly, concept-based methods are more robust to noise and perturbations (RQ1) in the input data, compared to local explanations, which can vary depending on the specific subset of data being analyzed. Therefore, a potential avenue for future exploration involves incorporating concept-based explanations while taking into account the unique characteristics of the RecSys graph data. Building upon the ideas presented in [15], our focus will be on extracting frequently occurring motifs or subgraph patterns from the input data. Additionally, we will consider structural attributes such as node degree or centrality measures, which provide insights into the significance of nodes within the graph. Through this RQ we will address the second gap introduced earlier.

3.3 RQ3: How to effectively and efficiently identify significant subgraphs for explaining the GNNs?

Various recent research endeavors that have devised explanation methods for GNNs, consistently emphasize the interpretability aspects at the levels of nodes, edges, or node features [29, 44, 53]. However, the inclusion of subgraphs is typically approached indirectly through the incorporation of regularization terms. Moreover, explanations at the subgraph level are deemed more intuitive and valuable, as subgraphs serve as fundamental building blocks within complex graphs and are closely tied to the graph's functionalities [56]. Subgraph analysis entails the identification and examination of smaller subgraphs within a larger graph. By focusing on these more manageable subgraphs, valuable insights can be obtained regarding the underlying data processing mechanisms of the GNN. [53, 56]. Current subgraph explainability methods have optimization task that maximizes the mutual information between a GNN's prediction and distribution of possible subgraph structures that is intractable so they use approximate methods that leads to a local minimum. To address this limitation, our strategy involves employing iterative magnitude pruning through loss landscape analysis of the graph training to capture more accurate and explainable subgraphs . The Iterative Magnitude Pruning (IMP) algorithm [9] is a cutting-edge technique that can discover extremely sparse matching subnetworks, referred to as "winning tickets", which can be retrained from an early stage or the initialization phase. IMP functions through successive cycles of training, whereby a proportion of the smallest magnitude weights are masked, the unmasked weights are reset to a prior training epoch, and the process is repeated. Additionally, inspired by [33] we hypothesize that we can use the same approach to find the optimal value which in this case will be most informative subgraph, in a more efficient way. One important consideration when applying iterative magnitude pruning to GNNs is that the graph structure may change after pruning. Therefore, it is important to re-estimate the adjacency matrix and other graph representations after each pruning iteration. Additionally, the pruning process should be carefully tuned to ensure that the pruned graph retains its important structural properties and does not become disconnected or lose important information. Through this RQ we will address the third gap introduced earlier.

3.4 RQ4: How can Shapley Value be effectively integrated into GNN recommenders to enhance their interpretability and performance?

Shapley Value (SV) [39] is a concept in cooperative game theory that measures the contribution of each player in a coalition game. SV provides a way to distribute the payoff of a coalition game among the players fairly, based on their individual contributions. Recently, SV has attracted a lot of attention in ML community. Various researches have demonstrated the effectiveness of SV on capturing the fair contribution of feature/nodes of a graph [6] which can also be considered for subgraphs as described in RQ3. SV have proven to be effective also in finding important neurons for backdoor defense towards adversarial attacks [17] which could be potentially used to link this approach to RQ1. Wang et al. [45] have used shapley value to build bidirectional associations between neurons and hierarchical concepts to explains whether and how the neurons learn the high-level hierarchical relationships of concepts which directly connects the importance of SV for RQ2. It is our assertion that there exists a need for heightened attention to be paid towards the comprehensive incorporation of the Shapley value (SV) in machine learning model interpretability, through the exploration of distributional Shapley approaches [14]. Drawing from the work of Ghorbani and Kousathanas [14], we intend to pursue a more universal approach to model explanation, by factoring in the influence of graph nodes and features through a distribution over their Shapley values. This approach aims to increase the flexibility of the model's interpretability towards the addition of new nodes in the network. By means of this research question, we intend to investigate the fourth identified gap that was presented in the introduction.

4 CONCLUSION AND NEXT STEPS

In this paper, we conducted a comprehensive analysis of recent advancements in explainable GNN-based RecSys and identified several noteworthy concerns and research gaps within the context of an ongoing PhD project. With the aim of making substantive contributions to this domain, we have formulated research inquiries and are committed to addressing them through thorough and rigorous investigations.

ACKNOWLEDGMENTS

Special thanks to Dr. Eva Zangerle for the feedback and guidance in this project.

REFERENCES

[1] David Alvarez-Melis and Tommi S. Jaakkola. 2018. Towards Robust Interpretability with Self-Explaining Neural Networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and

Roman Garnett (Eds.). 7786–7795. https://proceedings.neurips.cc/paper/2018/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html

[2] 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 *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 5644–5655. https://proceedings.neurips.cc/paper/2021/hash/2c8c3a57383c63caef6724343eb62257-Abstract.html

[3] Federico Baldassarre and Hossein Azizpour. 2019. Explainability Techniques for Graph Convolutional Networks. *CoRR* abs/1905.13686 (2019). arXiv:1905.13686 http://arxiv.org/abs/1905.13686

[4] Heng Chang, Yu Rong, Tingyang Xu, Wenbing Huang, Honglei Zhang, Peng Cui, Wenwu Zhu, and Junzhou Huang. 2020. A Restricted Black-Box Adversarial Framework Towards Attacking Graph Embedding Models. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020.* AAAI Press, 3389–3396. https://ojs.aaai.org/index.php/AAAI/article/view/5741

[5] Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song. 2018. Adversarial Attack on Graph Structured Data. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 1123–1132. http://proceedings.mlr.press/v80/dai18b.html

[6] Alexandre Duval and Fragkiskos D. Malliaros. 2021. GraphSVX: Shapley Value Explanations for Graph Neural Networks. In *Machine Learning and Knowledge Discovery in Databases. Research Track - European Conference, ECML PKDD 2021, Bilbao, Spain, September 13-17, 2021, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12976)*, Nuria Oliver, Fernando Pérez-Cruz, Stefan Kramer, Jesse Read, and José Antonio Lozano (Eds.). Springer, 302–318. https://doi.org/10.1007/978-3-030-86520-7_19

[7] Lukas Faber, Amin K. Moghaddam, and Roger Wattenhofer. 2021. When Comparing to Ground Truth is Wrong: On Evaluating GNN Explanation Methods. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, Feida Zhu, Beng Chin Ooi, and Chunyan Miao (Eds.). ACM, 332–341. https://doi.org/10.1145/3447548.3467283

[8] Wenqi Fan, Wei Jin, Xiaorui Liu, Han Xu, Xianfeng Tang, Suhang Wang, Qing Li, Jiliang Tang, Jianping Wang, and Charu C. Aggarwal. 2021. Jointly Attacking Graph Neural Network and its Explanations. *CoRR* abs/2108.03388 (2021). arXiv:2108.03388 https://arxiv.org/abs/2108.03388

[9] Junfeng Fang, Xiang Wang, An Zhang, Zemin Liu, Xiangnan He, and Tat-Seng Chua. 2023. Cooperative Explanations of Graph Neural Networks. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining* (Singapore, Singapore) *(WSDM '23)*. Association for Computing Machinery, New York, NY, USA, 616–624. https://doi.org/10.1145/3539597.3570378

[10] Timo Freiesleben. 2022. The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples. *Minds Mach.* 32, 1 (2022), 77–109. https://doi.org/10.1007/s11023-021-09580-9

[11] Chen Gao, Yu Zheng, Nian Li, Yinfeng Li, Yingrong Qin, Jinghua Piao, Yuhan Quan, Jianxin Chang, Depeng Jin, Xiangnan He, and Yong Li. 2021. Graph Neural Networks for Recommender Systems: Challenges, Methods, and Directions. *CoRR* abs/2109.12843 (2021). arXiv:2109.12843 https://arxiv.org/abs/2109.12843

[12] Azin Ghazimatin, Oana Balalau, Rishiraj Saha Roy, and Gerhard Weikum. 2020. PRINCE: Provider-side Interpretability with Counterfactual Explanations in Recommender Systems. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, James Caverlee, Xia (Ben) Hu, Mounia Lalmas, and Wei Wang (Eds.). ACM, 196–204. https://doi.org/10.1145/3336191.3371824

[13] Amirata Ghorbani, Abubakar Abid, and James Y. Zou. 2019. Interpretation of Neural Networks Is Fragile. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019.* AAAI Press, 3681–3688. https://doi.org/10.1609/aaai.v33i01.33013681

[14] Amirata Ghorbani, Michael P. Kim, and James Zou. 2020. A Distributional Framework For Data Valuation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 3535–3544. http://proceedings.mlr.press/v119/ghorbani20a.html

[15] Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. 2019. Towards Automatic Concept-based Explanations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 9273–9282. https://proceedings.neurips.cc/paper/2019/hash/77d2afcb31f6493e350fca61764efb9a-Abstract.html

[16] Amirata Ghorbani and James Y. Zou. 2020. Neuron Shapley: Discovering the Responsible Neurons. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/41c542dfe6e4fc3deb251d64cf6ed2e4-Abstract.html

[17] Jiyang Guan, Zhuozhuo Tu, Ran He, and Dacheng Tao. 2022. Few-shot Backdoor Defense Using Shapley Estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022.* IEEE, 13348–13357. https://doi.org/10.1109/CVPR52688.2022.01300

[18] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yong-Dong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 639–648. https://doi.org/10.1145/3397271.3401063

[19] Juyeon Heo, Sunghwan Joo, and Taesup Moon. 2019. Fooling Neural Network Interpretations via Adversarial Model Manipulation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 2921–2932. https://proceedings.neurips.cc/paper/2019/hash/7fea637fd6d02b8f0adf6f7dc36aed93-Abstract.html

[20] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. Adversarial Examples Are Not Bugs, They Are Features. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 125–136. https://proceedings.neurips.cc/paper/2019/hash/e2c420d928d4bf8ce0ff2ec19b371514-Abstract.html

[21] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. 2023. Adversarial Counterfactual Visual Explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 16425–16435.

[22] Wei Jin, Yaxin Li, Han Xu, Yiqi Wang, Shuiwang Ji, Charu Aggarwal, and Jiliang Tang. 2020. Adversarial Attacks and Defenses on Graphs. *SIGKDD Explor.* 22, 2 (2020), 19–34. https://doi.org/10.1145/3447556.3447566

[23] Wei Jin, Yao Ma, Xiaorui Liu, Xianfeng Tang, Suhang Wang, and Jiliang Tang. 2020. Graph Structure Learning for Robust Graph Neural Networks. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (Eds.). ACM, 66–74. https://doi.org/10.1145/3394486.3403049

[24] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. 2018. Visualizing the Loss Landscape of Neural Nets. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). 6391–6401. https://proceedings.neurips.cc/paper/2018/hash/a41b3bb3e6b050b6c9067c67f663b915-Abstract.html

[25] Wanyu Lin, Hao Lan, Hao Wang, and Baochun Li. 2022. OrphicX: A Causality-Inspired Latent Variable Model for Interpreting Graph Neural Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022.* IEEE, 13719–13728. https://doi.org/10.1109/CVPR52688.2022.01336

[26] Ninghao Liu, Mengnan Du, Ruocheng Guo, Huan Liu, and Xia Hu. 2021. Adversarial Attacks and Defenses: An Interpretation Perspective. *SIGKDD Explor.* 23, 1 (2021), 86–99. https://doi.org/10.1145/3468507.3468519

[27] Ninghao Liu, Hongxia Yang, and Xia Hu. 2018. Adversarial Detection with Model Interpretation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, Yike Guo and Faisal Farooq (Eds.). ACM, 1803–1811. https://doi.org/10.1145/3219819.3220027

[28] Ana Lucic, Maartje A. ter Hoeve, Gabriele Tolomei, Maarten de Rijke, and Fabrizio Silvestri. 2022. CF-GNNExplainer: Counterfactual Explanations for Graph Neural Networks. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event (Proceedings of Machine Learning Research, Vol. 151)*, Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (Eds.). PMLR, 4499–4511. https://proceedings.mlr.press/v151/lucic22a.html

[29] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. 2020. Parameterized Explainer for Graph Neural Network. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/e37b08dd3015330dcbb5d6663667b8b8-Abstract.html

[30] Jiaqi Ma, Shuangrui Ding, and Qiaozhu Mei. 2020. Towards More Practical Adversarial Attacks on Graph Neural Networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/32bb90e8976aab5298d5da10fe66f21d-Abstract.html

[31] Lucie Charlotte Magister, Dmitry Kazhdan, Vikash Singh, and Pietro Liò. 2021. GCExplainer: Human-in-the-Loop Concept-based Explanations for Graph Neural Networks. *CoRR* abs/2107.11889 (2021). arXiv:2107.11889 https://arxiv.org/abs/2107.11889

[32] Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial Training Methods for Semi-Supervised Text Classification. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=r1X3g2_xl

[33] Mansheej Paul, Feng Chen, Brett W Larsen, Jonathan Frankle, Surya Ganguli, and Gintare Karolina Dziugaite. 2022. Unmasking the Lottery Ticket Hypothesis: What's Encoded in a Winning Ticket's Mask? *arXiv preprint arXiv:2210.03044* (2022).

[34] Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. 2022. Exploring Counterfactual Explanations Through the Lens of Adversarial Examples: A Theoretical and Empirical Analysis. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event (Proceedings of Machine Learning Research, Vol. 151)*, Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (Eds.). PMLR, 4574–4594. https://proceedings.mlr.press/v151/pawelczyk22a.html

[35] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi (Eds.). ACM, 1135–1144. https://doi.org/10.1145/2939672.2939778

[36] Robert Schwarzenberg, Marc Hübner, David Harbecke, Christoph Alt, and Leonhard Hennig. 2019. Layerwise Relevance Visualization in Convolutional Text Graph Classifiers. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing, TextGraphs@EMNLP 2019, Hong Kong, November 4, 2019*, Dmitry Ustalov, Swapna Somasundaran, Peter Jansen, Goran Glavas, Martin Riedl, Mihai Surdeanu, and Michalis Vazirgiannis (Eds.). Association for Computational Linguistics, 58–62. https://doi.org/10.18653/v1/D19-5308

[37] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 618–626. https://doi.org/10.1109/ICCV.2017.74

[38] Caihua Shan, Yifei Shen, Yao Zhang, Xiang Li, and Dongsheng Li. 2021. Reinforcement Learning Enhanced Explainer for Graph Neural Networks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 22523–22533. https://proceedings.neurips.cc/paper/2021/hash/be26abe76fb5c8a4921cf9d3e865b454-Abstract.html

[39] Lloyd S. Shapley. 1952. *A Value for N-Person Games.* RAND Corporation, Santa Monica, CA. https://doi.org/10.7249/P0295

[40] Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. 2020. Robustness Verification for Transformers. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=BJxwPJHFwS

[41] Vivswan Shitole, Fuxin Li, Minsuk Kahng, Prasad Tadepalli, and Alan Fern. 2021. One Explanation is Not Enough: Structured Attention Graphs for Image Classification. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 11352–11363. https://proceedings.neurips.cc/paper_files/paper/2021/file/5e751896e527c862bf67251a474b3819-Paper.pdf

[42] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1312.6034

[43] Yiwei Sun, Suhang Wang, Xianfeng Tang, Tsung-Yu Hsieh, and Vasant G. Honavar. 2020. Adversarial Attacks on Graph Neural Networks via Node Injections: A Hierarchical Reinforcement Learning Approach. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen (Eds.). ACM / IW3C2, 673–683. https://doi.org/10.1145/3366423.3380149

[44] Minh N. Vu and My T. Thai. 2020. PGM-Explainer: Probabilistic Graphical Model Explanations for Graph Neural Networks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/8fb134f258b1f7865a6ab2d935a897c9-Abstract.html

[45] Andong Wang, Wei-Ning Lee, and Xiaojuan Qi. 2022. HINT: Hierarchical Neuron Concept Explainer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 10244–10254. https://doi.org/10.1109/CVPR52688.2022.01001

[46] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural Graph Collaborative Filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 165–174. https://doi.org/10.1145/3331184.3331267

[47] Xiaoqi Wang and Han-Wei Shen. 2022. GNNInterpreter: A Probabilistic Generative Model-Level Explanation for Graph Neural Networks. *CoRR* abs/2209.07924 (2022). https://doi.org/10.48550/arXiv.2209.07924 arXiv:2209.07924

[48] Bingzhe Wu, Jintang Li, Junchi Yu, Yatao Bian, Hengtong Zhang, Chaochao Chen, Chengbin Hou, Guoji Fu, Liang Chen, Tingyang Xu, Yu Rong, Xiaolin Zheng, Junzhou Huang, Ran He, Baoyuan Wu, Guangyu Sun, Peng Cui, Zibin Zheng, Zhe Liu, and Peilin Zhao. 2022. A Survey of Trustworthy Graph Learning: Reliability, Explainability, and Privacy Protection. *CoRR* abs/2205.10014 (2022). https://doi.org/10.48550/arXiv.2205.10014 arXiv:2205.10014

[49] Huijun Wu, Chen Wang, Yuriy Tyshetskiy, Andrew Docherty, Kai Lu, and Liming Zhu. 2019. Adversarial Examples for Graph Data: Deep Insights into Attack and Defense. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, Sarit Kraus (Ed.). ijcai.org, 4816–4823. https://doi.org/10.24963/ijcai.2019/669

[50] Han Xu, Yao Ma, Haochen Liu, Debayan Deb, Hui Liu, Jiliang Tang, and Anil K. Jain. 2020. Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. *Int. J. Autom. Comput.* 17, 2 (2020), 151–178. https://doi.org/10.1007/s11633-019-1211-x

[51] Kaidi Xu, Hongge Chen, Sijia Liu, Pin-Yu Chen, Tsui-Wei Weng, Mingyi Hong, and Xue Lin. 2019. Topology Attack and Defense for Graph Neural Networks: An Optimization Perspective. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, Sarit Kraus (Ed.). ijcai.org, 3961–3967. https://doi.org/10.24963/ijcai.2019/550

[52] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. 2018. Graph Convolutional Neural Networks for Web-Scale Recommender Systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, Yike Guo and Faisal Farooq (Eds.). ACM, 974–983. https://doi.org/10.1145/3219819.3219890

[53] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. GNNExplainer: Generating Explanations for Graph Neural Networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 9240–9251. https://proceedings.neurips.cc/paper/2019/hash/d80b7040b773199015de6d3b4293c8ff-Abstract.html

[54] Hao Yuan, Jiliang Tang, Xia Hu, and Shuiwang Ji. 2020. XGNN: Towards Model-Level Explanations of Graph Neural Networks. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (Eds.). ACM, 430–438. https://doi.org/10.1145/3394486.3403085

[55] Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. 2020. Explainability in Graph Neural Networks: A Taxonomic Survey. *CoRR* abs/2012.15445 (2020). arXiv:2012.15445 https://arxiv.org/abs/2012.15445

[56] Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. 2021. On Explainability of Graph Neural Networks via Subgraph Explorations. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 12241–12252. http://proceedings.mlr.press/v139/yuan21c.html

[57] Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. 2020. Interpretable Deep Learning under Fire. In *29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020*, Srdjan Capkun and Franziska Roesner (Eds.). USENIX Association, 1659–1676. https://www.usenix.org/conference/usenixsecurity20/presentation/zhang-xinyang

[58] Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. 2018. Adversarial Attacks on Neural Networks for Graph Data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, Yike Guo and Faisal Farooq (Eds.). ACM,

2847–2856. https://doi.org/10.1145/3219819.3220078
[59] Daniel Zügner and Stephan Günnemann. 2019. Adversarial Attacks on Graph Neural Networks via Meta Learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=Bylnx209YX