Advancing Graph Neural Networks: A Data-Centric View

Wei Jin (jinwei2@msu.edu)

Many learning tasks in Artificial Intelligence (AI) require dealing with graph data, ranging from biology and chemistry to finance and education. As powerful learning tools for graph inputs, graph neural networks (GNNs) have demonstrated remarkable performance in various graph-related applications. Recent research efforts have overwhelmingly been made on developing model-centric approaches to improve the performance of GNNs, which revolve around changing the model while holding the dataset fixed. Nevertheless, these approaches tend to yield less satisfying performance when there is a lack of a sufficiently large amount of high-quality data. Moreover, training GNNs is often computationally expensive on large-scale data; and such cost becomes even prohibitive when we need to train numerous models on the same dataset, e.g., searching for the best hyper-parameters and architectures. The presence of these issues raises a question at the heart of my research: *Can we develop a different perspective to empower graph neural networks?*

Data-Centric AI has shed light on overcoming the aforementioned limitations. While holding the model fixed, data-centric approaches directly optimize the given dataset to improve the imperfect model. **I strive to develop data-centric approaches to enhance graph neural networks**, where I deliver a set of techniques for graph dataset optimization to boost the model's effectiveness and efficiency. On the one hand, I seek to improve the quality of a graph dataset such that the underlying GNNs can be robust to severe noise and attacks. On the other hand, I design methods to reduce the size of a graph dataset while preserving its information such that the training cost can be significantly lowered. Since the outcome of data-centric approaches is a dataset, it allows different models to get improved while model-centric approaches are generally specific to a single model.

My research majorly probes the intersection between Data-Centric AI and Graph Representation Learning and has led to numerous publications in top-tier AI and data science conferences (e.g., ICLR, KDD, ICML, WSDM, and NeurIPS). My publications have received extensive attention and have been widely recognized. For instance, my work ProGNN [1] is selected as **Most Influential Paper** in KDD'20 by Paper Digest; and my work ElasticGNN [2] is accepted as an oral paper in ICML'21 **with an acceptance rate of 3% (166/5513)**. The impact of my work extends beyond theories. I am a passionate open-source contributor and have received **1.7k+ GitHub stars** (top 0.03% among 40M GitHub users). As the leader, I have contributed to the development of several open-source libraries including DeepRobust [3] (for secure AI) and DANCE [4] (for computational biology). Particularly, DeepRobust has attracted 20k+ downloads and **700+ stars**. I also translated the English book *Deep Learning on Graphs* to Chinese and it has been **the top seller** in China. In addition, I am committed to conducting interdisciplinary research, especially in computational biology, and I won **first place** in the task of modality prediction at NeurIPS'21 Single-Cell Multimodal Data Integration Competition.

1 Research Contributions

To expand the horizon of GNN research from models to datasets, I developed a set of techniques to optimize the graph datasets for training GNNs. My research contributions can be summarized as (1) **securing GNNs from a data-centric view**, where I improved the quality of the input graph data to boost the robustness of GNNs; and (2) **scaling up GNNs from a data-centric view**, where I developed methods of reducing the size of graph datasets to promote the scalability of GNNs. *My research work has a wonderful blend of elegant theory and high-impact practice*, and I will demonstrate two directions in the following.

Securing GNNs From A Data-Centric View. Despite the prosperity of GNNs, they have also exposed critical vulnerabilities as summarized in my survey paper [5]. As shown in Figure 1, an attacker can inject small perturbation to the input graph, which is referred to as *adversarial attack*, and mislead the GNN model into giving wrong predictions. The lack of robustness can lead to severe consequences for safety-critical applications such as financial systems and risk management. To resolve

the safety concern, a natural solution is to develop model-centric approaches by altering the architectures of GNNs such that they can be robust to adversarial attacks, as done in my previous work [6, 2, 7]. However, these approaches are generally tailored for a specific GNN model on a certain downstream task. In real-world scenarios, different applications could adopt different GNN variants and it is costly to adapt these methods to different GNNs. Given these drawbacks, it is highly desired to develop data-centric approaches to secure GNN models by directly improving the quality of input graph data. Revolving around this goal, I seek to eliminate the malicious behaviors in the perturbed graph and thus facilitate the correct prediction of GNNs, as shown in Figure 2.

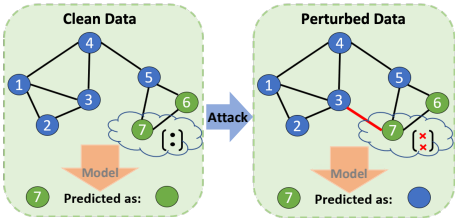

Figure 1: An example of adversarial attacks. The attacker injects one edge (highlighted in red) into the graph and modifies the features of node 7. Such change fools the given GNN model that gives a wrong prediction on node 7.

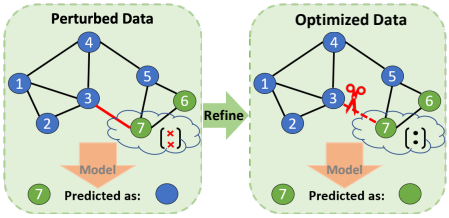

Figure 2: Data-centric approaches can refine the perturbed data by removing the patterns injected by adversarial attacks. Through dataset optimization, the given model can make correct predictions.

I proposed ProGNN [1] for learning to optimize the graph to counteract adversarial attacks. Through extensive empirical studies, I found that training-time adversarial attacks can heavily violate some important graph properties, i.e., low rank, sparsity, and feature smoothness. Inspired by this finding, ProGNN iteratively optimizes the noisy graph by restoring the violated graph properties, as to eliminate the injected adversarial patterns. Specifically, ProGNN models the process of graph reconstruction as a matrix estimation problem and regularizes the estimated matrix with the nuclear norm, ℓ_1 norm, and feature smoothness term, which correspond to the three graph properties, respectively. Since there are two non-differentiable terms, vanilla proximal gradient descent cannot be applied for optimization and I adopted the Incremental Proximal Descent method to effectively optimize the graph structure. Empirically, ProGNN achieves **up to 22% improvement** under different adversarial attacks and outperforms model-centric defenses by large margins. I further showed that ProGNN is able to remove a substantial amount of attacked edges. Notably, ProGNN was selected as **the Most Influential Paper** in KDD 2020 by PaperDigest due to its wide influence in the research community. However, I found that the robustness of ProGNN rapidly drops when the labeled data is limited. Thus, I proposed RS-GNN [8] to improve ProGNN under label sparsity. RS-GNN learns a link predictor to down-weight noisy edges to alleviate the impact of adversarial attacks and connects nodes with high similarity to fully take advantage of the information from unlabeled data. Given limited labels, it improves the robust accuracy of ProGNN by **over 10%**.

Different from the aforementioned works which target defending training-time attacks, I proposed GTrans [9] to counteract test-time attacks by transforming the graph structure as well as node features to remove adversarial patterns. It is **the first work** to transform the graph at test time. However, designing a learning objective to guide the transformation process is immensely challenging since we do not have access to the ground-truth labels of test samples. In the absence of test labels, I provided **theoretical analysis** that sheds insights on what losses should be used during test-time graph transformation. Based on the analysis, GTrans takes advantage of self-supervision from the test samples and optimizes a parameter-free contrastive loss to transform the test graph. In this regard, GTrans does not alter the model training process and can be readily used with a wide variety of pre-trained models and test settings. I demonstrated that the graph learned by GTrans can be employed

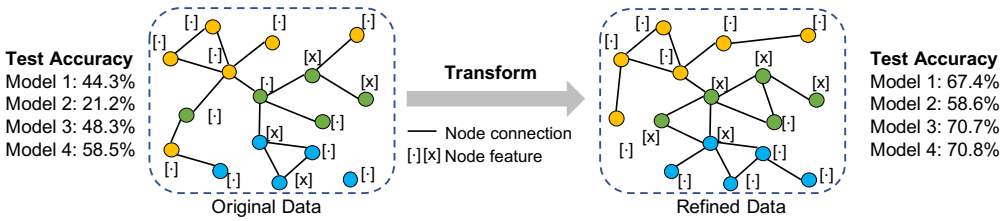

Figure 3: The refined data can benefit multiple model architectures.

by different models to improve their robustness **over 20%**. As shown in Figure 3, the refined graph data can benefit multiple model architectures and thus save the cost of modifying pre-trained models.

To deepen our understanding and immensely foster the research field of robust graph representation learning, I developed DeepRobust [3], **a comprehensive Python toolkit** for generating adversarial attacks and building robust models. In addition to the methods I introduced above, other popular data-centric defenses and model-centric defenses for graph data are included in DeepRobust to help researchers and practitioners secure their models. I believe this line of research can tremendously promote the reliability of graph machine learning and facilitate safety-critical applications.

Scaling Up GNNs From A Data-Centric View. As large-scale graphs are prevalent in real-world scenarios (often on the scale of millions of nodes and edges), it poses significant challenges in storing datasets and training GNNs on them. More dramatically, the computational cost continues to increase when we need to retrain the models multiple times, e.g., under incremental learning settings and neural architecture search. To tackle these issues, a natural idea is to properly simplify, or reduce the graph dataset while providing sufficient information to sufficiently train GNN models. To fulfill this need, I collaborated with researchers at Snap Inc. and proposed GCond [10], **the first approach** for synthesizing a small but informative graph with which we can train machine learning models sufficiently and efficiently, as shown in Figure 4. Given a distribution of random initializations of model parameters, GCond optimizes the condensed graph by matching the gradients of model parameters w.r.t. large-real and small-synthetic training data for multiple training steps. In this way, the GNN trained on the condensed graph can mimic the training trajectory of that on real data. Further, we carefully design the strategy for parametrizations for the condensed graph. In particular, we introduce the strategy of parameterizing the condensed features as free parameters and model the synthetic graph structure as a function of features. It takes advantage of the implicit relationship between structure and node features, consumes less number of parameters, and offers better performance. Remarkably, we are able to **reduce the graph size by 99.9%** while **approximating the original test accuracy by 99.8%**, and the condensed graphs can be used to train various model architectures and speed up the search process of model hyper-parameters. Furthermore, I show that the condensed graph provides a plausible interpretation of the original dataset and in some cases, it preserves the key properties of the original dataset.

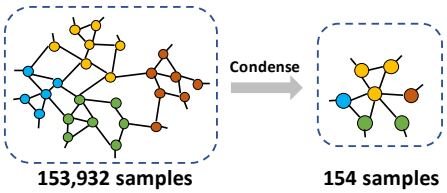

Figure 4: Graph Dataset Condensation.

Despite the promise of GCond, it has two inherent limitations. First, the condensation process in GCond is computationally expensive as it requires multiple steps of gradient matching for each set of parameters in each given initialized model. Second, it produces continuous edge weights to form the graph structure which consumes much more storage than a discrete structure. In collaboration with researchers at Amazon, I proposed DosCond [11] to address these limitations and facilitate the application of graph condensation in large datasets. On the one hand, it simplifies the optimization in GCond by performing gradient matching for only one single step. **With theoretical guarantees**, I showed that this strategy is able to learn synthetic graphs that lead to a small classification loss on real graphs. On the other hand, it models the graph structure as a probabilistic graph model and optimizes the discrete structures in a differentiable manner. With this approach, we can obtain discrete structures to facilitate the storage of condensed graphs. In the experiments, DosCond is **up to 40x faster** than GCond while still effectively approximating the original accuracy. Impressively, I deployed DosCond on **the real data at Amazon** and it also realizes effective reduction. I believe that this line of research can significantly reduce the training time of graph machine learning models and thus save lots of computational resources for both academia and industry.

2 Future Research Agenda

My long-term research goal is to expand the horizon of AI research from model-centric approaches to data-centric approaches. In my past work, I have developed data-centric approaches for advancing secure and scalable machine learning on graph data. Nevertheless, broader applications and a deeper

understanding of data-centric AI still remain under-explored, and thus I am interested in investigating its full potential. In addition, I aim to develop Trustworthy AI to expand my current research. Meanwhile, I am also deeply committed to continuing my efforts in interdisciplinary research.

Data-Centric AI. My prior work has demonstrated the tangible practical impacts of data-centric approaches on the security and scalability of GNNs. ① I am interested in exploring other types of data (e.g., images and text) and more applications, especially in areas of fairness and privacy. Recent studies have shown that many machine learning datasets including graphs, images, and text contain biases. It urges us to optimize the given dataset to directly mitigate the biases such that various machine learning models can provide fair results. Exploring this research direction can promote diversity, equity, and inclusion in our society. On a separate note, recent studies have shown that AI algorithms can carry the risk of disclosing users' private and sensitive information, such as medical records and financial transactions, etc. Thus, I hope to develop data-centric AI to encrypt the dataset to prevent user information leakage while not affecting the performance of models trained on it. ② I also plan to investigate methodologies for dataset creation. As previously stated, data quality is crucial to the success of AI systems. However, existing benchmark datasets are often noisy, biased, and overly simplified, which may not be extended to the practical scenario and can hinder the evaluation of AI models. To address these issues, I plan to construct real-world benchmark datasets that are large-scale and collected from diverse resources to ensure fairness. In fact, I am currently collaborating with Amazon researchers to build such real-world benchmark datasets from industry data. To make dataset creation more efficient, I will develop human-in-the-loop approaches, such as active learning algorithms, to prioritize the most valuable data for humans to annotate.

Trustworthy AI. Trustworthiness is the key to successful AI-enabled product deployment and uptake, and I plan to spend substantial efforts in developing Trustworthy AI systems. I have worked on the security perspectives and will continue to investigate them: I plan to enhance the robustness of my previous work on deep and self-supervised GNNs [12, 13, 14]. Besides, I will further explore the directions of fairness, privacy, and explanation in AI systems. As shown in recent studies, the output of some AI systems has a high correlation with the attributes of individuals, such as gender and ethnicity. For instance, my work [15] demonstrated that some text classification models exhibit biases toward different demographic groups (e.g, people of different ages, genders, and races). I hope to deepen our understanding of fairness in AI and develop algorithms that show no discrimination toward people from any group. Moreover, the interpretability of machine learning models has drawn increasing attention from academic researchers and industrial practitioners. I aim to build AI systems that produce interpretable results such that users can fully trust them and take advantage of them. In my work [16], I found that interpretable models can be used to detect adversarial attacks, which shows the importance of studying interpretable AI.

Interdisciplinary Research. Besides the above-mentioned directions, I plan to continue my efforts in conducting influential interdisciplinary research for biology and social good. In my work [17], I employed GNNs to capture the rich interactions between cells and genes, which benefits three fundamental tasks in single-cell multimodal integration. My package DANCE [4] on single-cell analysis has also prepared me well for continuing working in this direction. I plan to leverage my expertise in more diverse single-cell tasks such as imputation, spatial domain identification, cell-type deconvolution, etc. On a separate note, I have contributed to facilitating transportation management of smart cities [18] and improving recommender systems in online E-commerce platforms [19]. These experiences have prepared me for continuing the efforts on AI research for social good, as to have a positive impact on society. As I expand my research scope to these areas, I look forward to collaborating with experts in Social Science, Biology, and Bioinformatics. I also believe that most insights from my work are transferable to AI applications beyond graph data. Through collaborations with e.g., Human-Computer Interaction, Vision, Robotics, and Security researchers, I am eager to distill shared and unique challenges in shaping future Data-Centric AI.

References

[1] W. Jin, Y. Ma, X. Liu, X. Tang, S. Wang, and J. Tang, "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*, 2020.

[2] X. Liu, W. Jin, Y. Ma, Y. Li, H. Liu, Y. Wang, M. Yan, and J. Tang, "Elastic graph neural networks," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, 2021.

[3] Y. Li, W. Jin, H. Xu, and J. Tang, "Deeprobust: a platform for adversarial attacks and defenses," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 18, 2021, pp. 16 078–16 080.

[4] J. Ding, H. Wen, W. Tang, R. Liu, Z. Li, J. Venegas, R. Su, D. Molho, W. Jin, W. Zuo *et al.*, "Dance: A deep learning library and benchmark platform for single-cell analysis," *bioRxiv*, pp. 2022–10, 2022.

[5] W. Jin, Y. Li, H. Xu, Y. Wang, S. Ji, C. Aggarwal, and J. Tang, "Adversarial attacks and defenses on graphs," *ACM SIGKDD Explorations Newsletter*, no. 2, 2021.

[6] W. Jin, T. Derr, Y. Wang, Y. Ma, Z. Liu, and J. Tang, "Node similarity preserving graph convolutional networks," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. ACM, 2021.

[7] X. Liu, J. Ding, W. Jin, H. Xu, Y. Ma, Z. Liu, and J. Tang, "Graph neural networks with adaptive residual," in *Advances in Neural Information Processing Systems*, 2021.

[8] E. Dai, W. Jin, H. Liu, and S. Wang, "Towards robust graph neural networks for noisy graphs with sparse labels," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 2022, pp. 181–191.

[9] W. Jin, T. Zhao, J. Ding, Y. Liu, J. Tang, and N. Shah, "Empowering graph representation learning with test-time graph transformation," in *The Eleventh International Conference on Learning Representations*, 2023.

[10] W. Jin, L. Zhao, S. Zhang, Y. Liu, J. Tang, and N. Shah, "Graph condensation for graph neural networks," in *International Conference on Learning Representations*, 2022.

[11] W. Jin, X. Tang, H. Jiang, Z. Li, D. Zhang, J. Tang, and B. Yin, "Condensing graphs via one-step gradient matching," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

[12] W. Jin, X. Liu, X. Zhao, Y. Ma, N. Shah, and J. Tang, "Automated self-supervised learning for graphs," in *International Conference on Learning Representations*, 2022.

[13] W. Jin, X. Liu, Y. Ma, C. Aggarwal, and J. Tang, "Feature overcorrelation in deep graph neural networks: A new perspective," in *KDD*, 2022.

[14] Y. Wang, W. Jin, and T. Derr, "Graph neural networks: Self-supervised learning," in *Graph Neural Networks: Foundations, Frontiers, and Applications*, 2022.

[15] H. Liu, W. Jin, H. Karimi, Z. Liu, and J. Tang, "The authors matter: Understanding and mitigating implicit bias in deep text classification," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 74–85.

[16] W. Fan, W. Jin, X. Liu, H. Xu, X. Tang, S. Wang, Q. Li, J. Tang, J. Wang, and C. Aggarwal, "Jointly attacking graph neural network and its explanations," *arXiv preprint arXiv:2108.03388*, 2021.

[17] H. Wen, J. Ding, W. Jin, Y. Wang, Y. Xie, and J. Tang, "Graph neural networks for multimodal single-cell data integration," in *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 2022, pp. 4153–4163.

[18] X. Wang, Y. Ma, Y. Wang, W. Jin, X. Wang, J. Tang, C. Jia, and J. Yu, "Traffic flow prediction via spatial temporal graph neural network," in *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, 2020.

[19] W. Fan, X. Liu, W. Jin, X. Zhao, J. Tang, and Q. Li, "Graph trend filtering networks for recommendation," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.