Data Quality-Aware Graph Machine Learning

Yu Wang (yu.wang.1@vanderbilt.edu)

Recent years have witnessed a significant shift from just model-centric AI, which focuses on developing top-performing models, to data-centric AI, which emphasizes the quality and refinement of the data used in AI models. Concurrently, graph machine learning (GML), such as Graph Neural Networks (GNNs) has emerged as a rising approach for analyzing graph-structured data by fusing the topological information via message-passing and the feature information via neural transformation. Despite GML's unprecedented success in pushing the boundary of state-of-the-art performance in graph-based real-world applications such as recommender systems, drug discovery, and information retrieval [1, 2, 3], its strong dependence on node features and graph topology also makes it vulnerable to data-quality issues, which can catastrophically impair GML performance. On the one hand, graph-structured data, like many other data modalities, suffer from conventional data-quality issues, e.g., imbalance and bias. On the other hand, the intrinsic complex and abstract nature of graph topology can potentially exacerbate the aforementioned issues and bring up new ones. For example, the imbalance issue that initially happens at the quantitative level could also occur at the topological level [4]. The inherent bias encoded in the sensitive feature space might get amplified after the message passing in GNNs [5]. These issues severely impair downstream task performance and are challenging to diagnose due to graphs' inherent complexity and abstraction.

Given the criticality and ubiquity of the graph-data quality issues in compromising GML performance, **my research strives to establish the Data Quality-Aware Graph Machine Learning framework, which identifies data-quality issues on graph-structured data, diagnoses the problems of existing GML methods when facing data-quality issues, and propose practical solutions to mitigate their negative impacts.** Following this paradigm, I systematically study the graph data-quality issues and propose their corresponding solutions from four perspectives: (1) **Topology** [1, 3, 6, 7, 8, 9], i.e., global positional, ill local topology issues, and missing topology issues; (2) **Imbalance** [1, 4, 7, 8, 10], i.e., node-level imbalance, graph-level imbalance, and edge-level imbalance; (3) **Bias** [5, 7, 8, 11, 12, 13, 14], i.e., bias issues on sensitive group, explainability, and node degree; (4) **Weak Supervision** [15], i.e., semi-supervised learning and self-supervised learning.

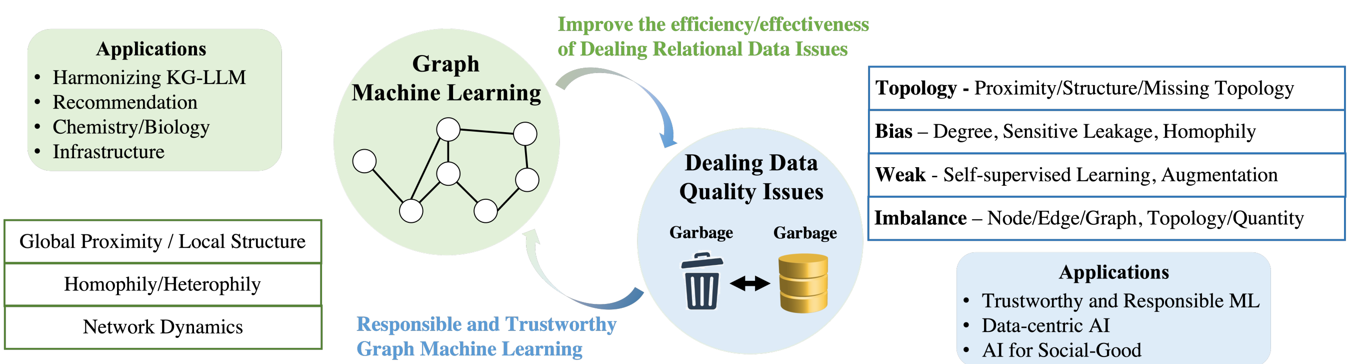

Figure 1: Data-quality-aware Graph Machine Learning.

My research on Data Quality-Aware Graph Machine Learning intersects between Data-centric AI and Graph Representation Learning, which has led to numerous publications in top-tier AI and data science conferences (e.g., KDD, AAAI, WWW, WSDM, CIKM, ICDMW, LOG) and two of them have been recognized respectively as the Top-10 most influential papers in CIKM'22 and WWW'23. The innovation of my research can be recognized through numerous prestigious awards, such as the Best Paper award in the 2020 Smokey Mountain Data Challenge, first-author of Vanderbilt's C.F.Chen Best Paper Award in 2022, and I was selected as the sole graduate student recipient of Vanderbilt's Graduate Leadership Anchor Award for Research in 2023. In addition to studying graph data quality issues, I am eagerly committed to conducting interdisciplinary research and applying the outcome to practical realms such as Science/Recommender Systems/Documents/Infrastructures. Specifically, my contributed open-source drug pair scoring Github Repository, ChemicalX, has garnered 650+ stars [16] with a total stars over 800+ for all my contributed projects. My internship at The Home Depot resulted in a knowledge-graph-enhanced session recommendation framework, elevating the offline session recommendation performance [9] on the million-scale product platform. Meanwhile, at Adobe Research, I spearheaded a novel approach to incorporate knowledge graphs into prompting LLMs [3], a contribution that earned recognition from a chief scientist at the Company, Aldecis. Moreover, our team collaborating with students from civil engineering won first place in the 2020 Smokey Mountain Data Computation in analyzing urban mobility patterns and human activities.

Research Contribution

Topology Issues. Graph topology acts as a double-edged sword in influencing GML performance. While the topology can provide additional gleaning patterns to benefit downstream tasks, it may also compromise the quality of the learned node embeddings when misapplied in unsuitable scenarios. Specifically, my works [1, 3, 6, 7, 8, 9, 17] explore the ill local topology and missing topology issues.

GNNs obtain high-quality embedding for each node through iterative message-passing over the local topology centering around that node. When performing node classification on heterophily networks where neighboring nodes primarily belong to different classes or share distinct feature distributions shown in Figure 2, the message-passing may fuse features belonging to different classes, obtain non-informative node embeddings, and compromise the node classification performance. To mitigate the over-smoothing caused by aggregating neighboring information from different classes, I propose a tree decomposition mechanism to separate the message-passing from neighbors of different layers. As shown in Figure 2 and also theoretically analyzed in [6], instead of aggregating 2$^{\text{nd}}$-layer neighbor v_7 information indirectly via 1$^{\text{st}}$-layer neighbor v_4, I extract the subgraphs at different layers and design a Tree-Decomposed Graph Neural Network (TDGNN) to selectively aggregate neighbors information at different layers that are most beneficial to the centering node. Empirically, TDGNN improves the performance over existing baselines and demonstrates the equal importance of capturing multi-hop dependencies and incorporating high-layer neighbor information.

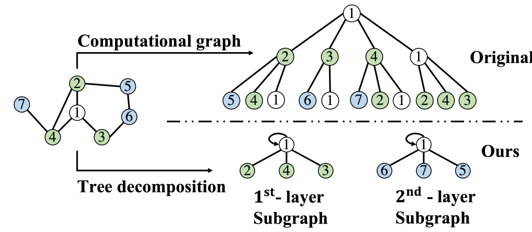

Figure 2: Tree decomposition of node v_1.

In addition to the node classification, I further investigate how local topology impacts each node's link prediction (LP) performance. Our series works [7, 8] theoretically prove the degree-related bias in LP evaluation metrics, motivating us to delve deep into the relationship between the node degree and its performance. Given the empirically observed weak correlation between the degree and LP performance [7], I propose the Topological Concentration (TC) metric, which measures the interactions among the local subgraphs of the neighbors of each node as shown in Figure 3. Remarkably, TC exhibits an 82.10% stronger correlation with LP performance and highlights a 200% increase in the performance disparity between identified under-performed nodes and their counterparts than node degree. In [7], I show that the node LP performance does not increase as the node degree increases while monotonically increasing as the Train-TC increases. **This pioneering insight breaks the prevailing belief that low-degree nodes perform worse and further suggests that cold-start nodes might not necessarily lag in LP.** Moreover, with TC, I first discovered the distribution shift in the topological space, i.e., newly-joined neighbors become less connected to the existing neighbors of one node.

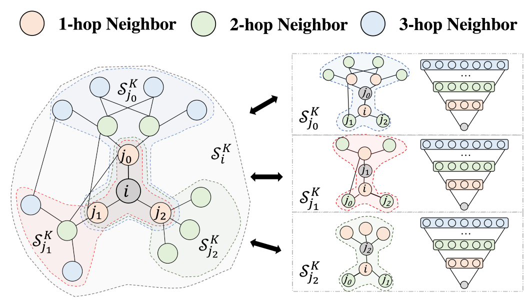

Figure 3: v_i's Topological Concentration.

As TC closely correlates to node LP performance and it essentially measures the contribution of the interactions of each neighbor with the whole neighborhood of a central node, I further design the Collaboration-aware Graph Convolution(CAGC) [1], to synthetically augment node TC by aggregating more information from neighbors with higher interactions to the whole neighborhoods. For example, in Figure 3, our designed CAGC would aggregate more information from j_0, j_1 to i rather than j_2 since j_2 has fewer interactions with the whole neighborhoods of i and might be an outlier compromising the learned embedding of i. I empirically demonstrate that the designed CAGC fused with LightGCN achieves 10% performance improvement with 80% speed up over existing baselines and theoretically prove the CAGC surpasses the 1-WL test. **Notably, this research is the first to showcase the effectiveness of graph convolution surpassing the 1-WL test in LP performance and has been selected as the 9$^{\text{th}}$ most influential paper in WWW 2023 according to Paper Digest by 10/15/2023.**

While addressing the issue of ill local topology is essential, the more pressing concern is the underperformance or even the complete failure of GML models in the absence of graph topology. Unfortunately, many real-world applications do not come along with a natural graph by itself. For example, e-commerce platforms typically only have customer-product interactions without any structured relation among products. Recognizing this challenge, two of my prior intern projects aim to create knowledge graphs from existing data that serve downstream tasks. In the first intern project with The Home Depot [9], I constructed the product knowledge graph by connecting two co-interacted products with the same customer in the same session and enhanced the session recommendation performance during off-line evaluation over 60 million sessions involving million-scale products.

In the second intern project at Adobe Research [3], I constructed a knowledge graph over multiple documents by connecting passages sharing higher semantic/lexical similarity. Furthermore, I add table/page nodes to denote document structures and devise an LLM-guided graph traversal to navigate the KG, rationalize, and collect relevant evidence for prompting LLMs to answer questions over multi-documents. **This work has become the pioneering example of harnessing the retrieval-augmented generation, marrying the strength of LLMs in rationalizing the evidence approaching the questions with the power of knowledge graphs in grounding the factual information.** In addition, my previous research [17, 18] also explores the heuristic rule-based and data-driven way to reconstruct/synthesize interdependent infrastructure networks and provides a unified testbed for vulnerability analysis of critical infrastructure systems.

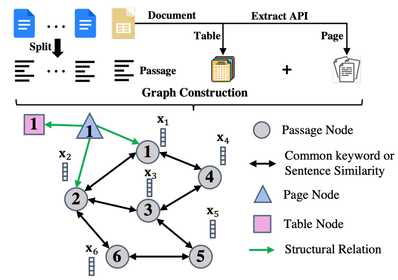

Figure 4: Constructing knowledge graph over documents.

Imbalance Issues. Data imbalance is widespread across various domains (e.g., chemistry/social), manifesting in diverse formats (e.g., quantity/topology), and exhibits different graph granularity (e.g., nodes/graphs/edges). Consequently, my previous research focuses on handling these three granularity of imbalance respectively [2, 4, 10, 11]. The node-level imbalance refers to the imbalanced supervision assigned to nodes in different classes and is categorized into quantitative and topological imbalance. My previous research [10] handles the quantitative imbalance by designing a distance-wise prototypical GNN paired with an imbalanced label propagation mechanism to augment the supervision for nodes in minority classes. The graph-level imbalance issue stands out in graph classification tasks such as drug discovery. For example, high-throughput screening in drug discovery yields a success rate of only 0.05-0.5%, and only about 2% of autism cases demonstrate abnormal brain structures [19]. My previous research [4] handles this graph-level imbalance issue by constructing graph-of-graphs(GoG) connecting graphs sharing higher topological similarity shown in Figure 5(a). The message-passing performed over the constructed GoG essentially imitates the mix-up between topologically similar graphs and hence remedies the imbalanced supervision in the topological space. I theoretically justify the benefit of the proposed GoG framework by relating it to label smoothing. **This is the very first work studying imbalanced graph classification using GNNs and has been ranked 6$^{\text{th}}$ most influential paper in CIKM'22.** In addition to the imbalanced node/graph classification, **I take the initiative to study the topological imbalance in edge classification and generalize the node-level homophily to edge-level homophily.** As shown in Figure 5(b), the minority edges are usually surrounded by edges from different classes (lower homophily), while the majority edges tend to cluster together (higher homophily). This significantly different class distribution in the local subgraph around each edge causes imbalanced performance in edge classification. To mitigate this issue, I propose topological reweight with the wedge-based mix-up to reweigh the contribution of each edge in computing the training loss and derive novel supervision for minority edges.

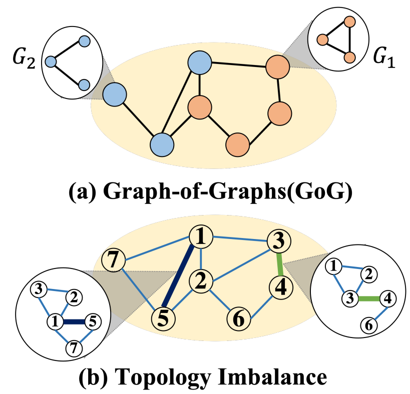

Figure 5: I handle imbalanced graph classification via GoG and observe the topological imbalance in edge classification.

Bias Issues. The topological characteristics of graphs will likely encode biased information, causing graph ML models trained on the biased graphs prone to discrimination. My previous research uncovers the topology-inspired bias mainly from the perspectives of community structure and degree distribution. Specifically, [5, 20] studies the community-induced bias from the feature and spatial perspectives in node classification. [5] discovers that the message-passing varies the correlation of previously non-sensitive features to the sensitive ones, causing the leakage of sensitive information and further exacerbating discrimination in prediction. **Moreover, this work [5] is the first to theoretically establish the connection between network homophily and group fairness in node classification.** [7, 8, 21] studies the degree-induced bias in link prediction/recommendation and empirically verifies that link prediction performance tends to behave worse for nodes with higher interest diversity/degree. In addition, we also explore fairness in explainability [13] and real-world applications such as online-dating recommendations [14].

Weak Supervision Issues. GML models have been demonstrated to require a massive amount of annotated data to reach their full potential, which is typically unrealistic since obtaining this high-quality labeled data demands extensive human annotations. To augment existing limited training signals and derive novel ones, our book chapter [15] systematically reviews recent self-supervised learning and data augmentation techniques used in semi/self-supervised settings. Following some of these design recipes, we devise self-consistency paired with graph augmentation in [4] and the label-smoothing pretext tasks in [10] to boost GNNs' performance when only limited data is provided.

Future Research Blueprint

Marrying Power of AI and Network Science. As the power of any graph machine learning task heavily relies on its underlying network structure, delving deeper into the sophisticated realms of NS would catalyze the evolution of avant-garde GML techniques. Previously, I had applied my NS knowledge in designing model architectures/deriving novel insights in handling/understanding graph data-quality issues, e.g., designing a Breadth-First-Search-based tree decomposition algorithm to enhance node classification on heterophily networks [6] or devising a topological concentration metric to better characterize the node LP performance [7]. Following this research principle, I hope to continuously bridge the profound knowledge of NS into tailoring state-of-the-art machine learning techniques. Concretely, I plan to (1) equip Artificial Intelligence Generated Content (AIGC) with NS/GT by fusing topology-based regularization constraining the generation process of existing graph diffusion methods (e.g., I deeply collaborate with my labmate in molecular ML and plan to follow-up work on enhancing imbalance drug discovery [2, 19] by diffusion-based molecular generation); (2) design novel topological encodings to make large-language models(LLMs) fully aware of the complex network structure (e.g., I am currently collaborating with Adobe researchers in designing position/role-based topological encoding techniques to augmenting the LLMs' capability for the textual generation.); (3) investigate the applications of network dynamics in designing lifelong GML (e.g., my work [7] identifies a topological distribution shift that newly-joined neighbors become less connective with existing neighbors of a node);

Harmonizing Knowledge Graph (KG) and Large Language Models (LLMs). Large language models (LLMs), such as LLaMA2 and GPT4, are making new waves in natural language processing and artificial intelligence due to their human-like capability and domain-agnostic generalizability. However, LLMs are black-box models, falling short of capturing and accessing factual knowledge. In contrast, Knowledge Graphs (KGs), such as Wikipedia Knowledge Base, are structured databases explicitly storing interpretable factual knowledge. KGs can enhance LLMs by providing external knowledge for grounding, rationalizing, and interpreting. However, KGs are hard to construct, usually domain-specific, and consistently evolve by nature, which limits their long-term benefits to broad real-world applications. Therefore, it is complementary to bridge LLMs and KGs together and simultaneously harness the strength of both. My forward-looking roadmap for this research field is bipartite: (1) LLMs-augmented KG: leverage LLMs to improve/create novel KG signals for enhancing/completing KG-based tasks such as knowledge graph completion/question-answering. (2) KG-augmented LLMs: incorporate KG during the training/inference phase of LLMs to ground/constrain the generated information from LLMs. One golden example is my previous intern project [3], which leverages the document KGs to mitigate the hallucination of LLMs (KG-augmented LLMs) and leverage LLMs to guide the graph traversal (LLM-augmented KGs).

Data-centric AI for Social-Good Applications The artificial intelligence (AI) community has traditionally taken a model-centric perspective and primarily focuses on developing models for refreshing State-of-the-Art performance while keeping the datasets untouched. However, many of these improvements are narrowly domain-specific and have shown the power exclusively on benchmark datasets, which overlooks potential data quality issues such as missing values/anomalies/imbalance/bias/incorrect annotations and behave disastrously in real-world scenarios outside the training domains. Furthermore, much of model-centric AI has been driven by a leaderboard mentality associated with these benchmark datasets, resulting in part of the research community being biased towards more and more complex models that achieve more excellent performance yet more unrealistic utility. This has led to the recent rise in data-centric AI, which emphasizes curating and refining data used within AI models. In my future research endeavors, I am committed to advancing this data-centric AI direction, as already evidenced by my previous research [5, 17, 20], to curate the data from the wild and derive the most appropriate signals for downstream applications. Beyond research on data-centric AI, I am keen on translating my findings into tangible real-world applications for social good. For example, I have constructed the responsible question-answering systems/recommendation platform as exhibited in [3, 9], designed a cost-efficient drug discovery workflow as initially explored in [19], and analyzed human activities and urban mobility patterns in resolving congestion and emission issues [22].

Trustworthy and Responsible AI. Trustworthiness and Responsibility are essential in guaranteeing the successful deployment, long-standing maintenance, and consistent upgrade of AI-related products. I plan to spend substantial efforts on developing Trustworthy and Responsible AI systems. My previous research has uncovered unfair utility [5, 20, 21], uneven explainability [13], and imbalanced link prediction performance [4, 7, 8] among instances in different groups. Following this paradigm, I will continuously enhance existing AI systems regarding their fairness (e.g., design an AI-based disaster response system that ensures fair resource allocation to people in different communities when natural disasters happen following my previous work [23, 17]), privacy (e.g., how to remove any sensitive/privacy information contained in the generated contents in the era of AIGC as proposed in my survey [24]), robustness (e.g., enhance reliability and robustness of infrastructure systems by AI [25, 23]), and explainability (e.g., how to guarantee that the recommender systems deliver not only high-quality recommendation but also faithful interpretation).

References

[1] **Wang, Yu**, Yuying Zhao, Yi Zhang, and Tyler Derr. Collaboration-aware graph convolutional network for recommender systems. In *Proceedings of the ACM Web Conference 2023*, pages 91–101, 2023.

[2] Yunchao Lance Liu, **Wang, Yu**, Oanh Vu, Rocco Moretti, Bobby Bodenheimer, Jens Meiler, and Tyler Derr. Interpretable chirality-aware graph neural network for quantitative structure activity relationship modeling in drug discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14356–14364, 2023.

[3] **Wang, Yu**, Nedim Lipka, Ryan A Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. Knowledge graph prompting for multi-document question answering. *arXiv preprint arXiv:2308.11730*, 2023.

[4] **Wang, Yu**, Yuying Zhao, Neil Shah, and Tyler Derr. Imbalanced graph classification via graph-of-graph neural networks. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2067–2076, 2022.

[5] **Wang, Yu**, Yuying Zhao, Yushun Dong, Huiyuan Chen, Jundong Li, and Tyler Derr. Improving fairness in graph neural networks via mitigating sensitive attribute leakage. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1938–1948, 2022.

[6] **Wang, Yu** and Tyler Derr. Tree decomposed graph neural network. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 2040–2049, 2021.

[7] **Wang, Yu**, Tong Zhao, Yuying Zhao, Yunchao Liu, Xueqi Cheng, Neil Shah, and Tyler Derr. A topological perspective on demystifying gnn-based link prediction performance. *arXiv preprint arXiv:2310.04612*, 2023.

[8] **Wang, Yu** and Tyler Derr. Degree-related bias in link prediction. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 757–758, 2022.

[9] **Wang, Yu**, Amin Javari, Janani Balaji, Walid Shalaby, Tyler Derr, and Xiquan Cui. Knowledge graph-based session recommendation with session-adaptive propagation. *In submission*, 2023.

[10] **Wang, Yu**, Charu Aggarwal, and Tyler Derr. Distance-wise prototypical graph neural network in node imbalance classification. *arXiv preprint arXiv:2110.12035*, 2021.

[11] **Wang, Yu**. Fair graph representation learning with imbalanced and biased data. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1557–1558, 2022.

[12] April Chen, Ryan A Rossi, Namyong Park, Puja Trivedi, **Wang, Yu**, Tong Yu, Sungchul Kim, Franck Dernoncourt, and Nesreen K Ahmed. Fairness-aware graph neural networks: A survey. *arXiv preprint arXiv:2307.03929*, 2023.

[13] Yuying Zhao, **Wang, Yu**, and Tyler Derr. Fairness and explainability: Bridging the gap towards fair model explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11363–11371, 2023.

[14] Yuying Zhao, **Wang, Yu**, Yi Zhang, Pamela Wisniewski, Charu Aggarwal, and Tyler Derr. Fair online dating recommendations for sexually fluid users via leveraging opposite gender interaction ratio. 2018.

[15] **Wang, Yu**, Wei Jin, and Tyler Derr. *Graph neural networks: Self-supervised learning*. Springer, 2022.

[16] Benedek Rozemberczki, Charles Tapley Hoyt, Anna Gogleva, Piotr Grabowski, Klas Karis, Andrej Lamov, Andriy Nikolov, Sebastian Nilsson, Michael Ughetto, **Wang, Yu**, et al. Chemicalx: A deep learning library for drug pair scoring. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3819–3828, 2022.

[17] **Wang, Yu**, Jin-Zhu Yu, and Hiba Baroud. Generating synthetic systems of interdependent critical infrastructure networks. *IEEE Systems Journal*, 16(2):3191–3202, 2021.

[18] **Wang, Yu**, Jin-Zhu Yu, and Hiba Baroud. A bayesian approach to reconstructing interdependent infrastructure networks from cascading failures. *arXiv preprint arXiv:2211.15590*, 2022.

[19] Yunchao Lance Liu, Rocco Moretti, **Wang, Yu**, Bobby Bodenheimer, Tyler Derr, and Jens Meiler. Integrating expert knowledge with deep learning improves quantitative structure-activity relationship(qsar) models for computer-aided-design-and-drafting(cadd) modeling. *bioRxiv*.

[20] Yushun Dong, Song Wang, **Wang, Yu**, Tyler Derr, and Jundong Li. On structural explanation of bias in graph neural networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 316–326, 2022.

[21] Yuying Zhao, **Wang, Yu**, Yunchao Liu, Xueqi Cheng, Charu Aggarwal, and Tyler Derr. Fairness and diversity in recommender systems: A survey. *arXiv preprint arXiv:2307.04644*, 2023.

[22] Ao Qu, **Wang, Yu**, Yue Hu, Yanbing Wang, and Hiba Baroud. A data-integration analysis on road emissions and traffic patterns. In *Driving Scientific and Engineering Discoveries Through the Convergence of HPC, Big Data and AI: 17th Smoky Mountains Computational Sciences and Engineering Conference, SMC 2020, Oak Ridge, TN, USA, August 26-28, 2020, Revised Selected Papers 17*, pages 503–517. Springer, 2020.

[23] **Wang, Yu**, Jin-Zhu Yu, and Hiba Baroud. Quantifying the interdependency strength across critical infrastructure systems using a dynamic network flow redistribution model. In *Proceedings of the 30th European Safety and Reliability Conference and 15th Probabilistic Safety Assessment and Management Conference.*, 2020.

[24] Yi Zhang, Yuying Zhao, Zhaoqing Li, Xueqi Cheng, **Wang, Yu**, Olivera Kotevska, Philip S Yu, and Tyler Derr. A survey on privacy in graph neural networks: Attacks, preservation, and applications. *arXiv preprint arXiv:2308.16375*, 2023.

[25] Qingfei Gao, **Wang, Yu**, Jun Li, Kejian Sheng, and Chenguang Liu. An enhanced percolation method for automatic detection of cracks in concrete bridges. *Advances in Civil Engineering*, 2020:1–23, 2020.