

Dissertation Defense
Doctor of Philosophy in Information Science

“On the Adversarial Attacks of Graph Neural Networks and Their Emerging Paradigms”
by Ying Song

Date: October 19, 2026

Time: 2 to 4 p.m.

Place: Room 828, Information Sciences Building, 135 N.
Bellefield Ave, Pittsburgh, PA 15260

Committee:

- Dr. Balaji Palanisamy, Associate Professor and Advisor, Department of Informatics and Networked Systems, School of Computing and Information
- Dr. James Joshi, Professor, Department of Informatics and Networked Systems, School of Computing and Information
- Dr. Daqing He, Professor, Department of Informatics and Networked Systems, School of Computing and Information
- Dr. Pengfei Zhou, Assistant Professor, Department of Informatics and Networked Systems, School of Computing and Information
- Dr. Xiaowei Jia, Associate Professor, Department of Computer Science, School of Arts and Sciences, Rutgers University

Abstract:

Graph neural networks (GNNs) have achieved remarkable success in diverse and high-impact real-world applications, such as social network analysis, drug discovery, and financial fraud detection. While their success hinges on propagating and aggregating information across graph topology via message passing, the same mechanism also creates complex and often hidden attack surfaces, leaving GNNs highly susceptible to unique adversarial attacks. Given the intricate and security-critical nature of their application domains, such attacks can trigger severe consequences, ranging from compromised model integrity to significant privacy breaches, underscoring an urgent need to systematically audit and safeguard GNNs and their emerging paradigms.

This dissertation conducts a rigorous and systematic investigation into adversarial attacks on GNNs and their emerging paradigms. We propose four novel attack frameworks to examine their security and privacy threats. First, we introduce a novel graph prompt backdoor attack to exploit the inherent vulnerability of crafting benign prompts as triggers, achieving a 100% attack success rate by poisoning as few as two training nodes without sacrificing clean accuracy. Second, we design a decoupling-based model stealing attack against GNNs under a stricter and more realistic threat model, demonstrating its feasibility and efficacy particularly in settings with sparse graphs and incomplete class coverage. Third, we develop a full graph reconstruction attack revealing that graph unlearning fails to erase deleted entities' information, and even exposes sensitive content from their neighbors, subverting the regulatory guarantees expected from such unlearning paradigms. Finally, we pioneer a query-free, data-only sensitive attribute inference attack that challenges the long-standing assumption that inference relies on active model queries, extending prior single-attribute attacks to simultaneous multi-attribute inference on graphs.

The insights derived from this adversarial investigation provide a security and privacy blueprint for building the next generation of trustworthy GNN-based systems. This dissertation, therefore, does not just identify the inherent weaknesses of GNNs and their emerging paradigms, but also establishes the foundational principles necessary to develop effective and robust defense mechanisms that can secure GNNs, especially in high-stakes environments.