University of Pittsburgh | School of Computing and Information

Proposal Defense

Doctor of Philosophy in Information Science

"Security Vulnerabilities and Privacy Risks of Graph Neural Networks" by Ying Song

Date: December 4, 2025
Time: 1 – 3 p.m.
Place: Room 302, 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 Computing and Information

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. This widespread deployment is predicated on the assumption of model integrity and data privacy. The GNN's inherent message-passing mechanism—which tightly couples learned node representations with the surrounding topological structure—creates complex hidden attack surfaces. Consequently, the security and privacy implications of GNNs and their rapidly evolving advanced application methodologies, such as graph prompting and graph unlearning, remain critically underexplored. This gap hinders the development of robust and trustworthy GNN-based systems.

This dissertation conducts a rigorous and systematic investigation into security vulnerabilities and privacy risks of GNNs. We propose four novel attack frameworks to expose and quantify the vulnerabilities and risks introduced by both GNN architectures and their cutting-edge application methodologies. First, we introduce a novel graph prompt backdoor attack to expose the inherent vulnerabilities of crafting normal prompts as triggers, where solely poisoning 2 training nodes to achieve 100% attack success rates without sacrificing clean accuracy. Second, we measure the adverse risks of model stealing and demonstrate its feasibility and efficacy against standard GNNs. Third, we reveal the critical privacy risks of reconstructing not only a deleted individual's information and personal links but also sensitive content from their connections, subverting the regulatory guarantees expected from graph unlearning mechanisms. Finally, we quantify the sensitive attribute exposure of trained GNNs by developing novel attribute inference attacks.

The insights derived from this adversarial investigation provide a new security and privacy blueprint for future GNN development. They articulate the foundational requirements for building certifiably robust and privacy-preserving graph machine learning systems. This dissertation, therefore, does not just identify inner weakness of GNNs and their application methodologies, but establishes the critical design principles necessary to develop the next generation of effective and reliable defense mechanisms that can safeguard GNNs especially in high-stake environments.