STEVENS
INSTITUTE OF TECHNOLOGY
1870

Ph.D. DISSERTATION DEFENSE

Candidate:	Kun Wu
Degree:	Doctor of Philosophy
School/Department:	The Charles V. Schaefer, Jr. School of Engineering and Science / Department of Computer Science
Date:	December 9, 2025
Time/Location:	1:30 – 3:30 PM (GN 421)
Title:	Towards Efficient and Verifiable Machine Unlearning
Chairperson:	Dr. Wendy Hui Wang, Department of Computer Science, Stevens Institute of Technology
Committee Members:	Dr. Yue Ning, Department of Computer Science, Stevens Institute of Technology Dr. Jie Shen, Department of Computer Science, Stevens Institute of Technology Dr. Yuan Hong, School of Computing, University of Connecticut

ABSTRACT

As machine learning systems become increasingly integrated into socially and privacy-sensitive applications, ensuring their compliance with data protection regulations has become critical. In particular, it is essential to efficiently remove the influence of specific data from trained models and to verify the success of such removal. This dissertation addresses these challenges by developing frameworks for efficient and verifiable machine unlearning (MU).

First, we propose Certified Edge Unlearning (CEU) for Graph Neural Networks (GNNs) [1], an efficient unlearning framework that approximates retraining while offering formal guarantees on edge removal. CEU achieves scalable, effective, and provably correct unlearning without compromising model performance.

Second, we introduce PANDA [2], the first probabilistic verification framework for detecting incomplete unlearning in GNNs. PANDA employs adversarial perturbations to construct verifiable challenge edges, providing probabilistic guarantees on detecting unfaithful unlearning with negligible model distortion.

Third, we present Quantization-Aided Machine Unlearning (QAMU), which integrates model quantization into the unlearning process to reduce the retrain-unlearn gap of the MU models and improve unlearning efficiency.

Beyond machine unlearning, this dissertation also explores complementary directions

toward trustworthy machine learning. Specifically, we propose a framework to enhance individual fairness in recommender systems by modeling latent user similarity through second-order proximity [3], demonstrating that fairness can be improved without sacrificing accuracy. Furthermore, we develop a knowledge-enhanced approach for misinformation detection, leveraging graph-based reasoning to improve robustness against incomplete or misleading content [4].

Together, these contributions advance a holistic vision of trustworthy machine learning, encompassing efficient and verifiable unlearning, fairness, and robustness.

REFERENCE:

[1] Kun Wu, Jie Shen, Yue Ning, Ting Wang, and Wendy Hui Wang. Certified edge unlearning for graph neural networks. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023.

[2] Kun Wu and Wendy Hui Wang. Verification of Incomplete Graph Unlearning through Adversarial Perturbations. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2025.

[3] Kun Wu, Jacob Erickson, Wendy Hui Wang, and Yue Ning. Equipping recommender systems with individual fairness via second-order proximity embedding. In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2022.

[4] Kun Wu, Xu Yuan, and Yue Ning. Incorporating Relational Knowledge in Explainable Fake News Detection. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2021.