STEVENS
INSTITUTE OF TECHNOLOGY
1870

Ph.D. DISSERTATION DEFENSE

Candidate:	Da Zhong
Degree:	Doctor of Philosophy
School/Department:	School of Engineering and Science / Computer Science
Date:	Tuesday, April 22nd, 2025
Time/Location:	11:00 am / Gateway North 303
Title:	Privacy Disparities in Machine Learning: Causes, Mitigation, and New Privacy Threats
Chairperson:	Dr. Wendy Hui Wang, Department of Computer Science, School of Engineering & Sciences
Committee Members:	Dr. Yue Ning, Department of Computer Science, School of Engineering & Sciences Dr. Tian Han, Department of Computer Science, School of Engineering & Sciences Dr. Jun Xu, School of Computing. The University of Utah.

ABSTRACT

Machine learning (ML) algorithms are applied across various domains. Despite their success, concerns about privacy leakage remain a significant challenge. In this dissertation, I focus on the disparities in privacy risks associated with ML algorithms. I specifically concentrate on Membership Inference Attacks (MIAs), and I demonstrate that privacy leakage varies at both the sample level and the model level.

Aiming to explore the disparity in privacy risks of machine learning models, I focus on three research questions: (1) How can privacy leakage disparity be mitigated? (2) How to design attacks tailored for data or models that exhibit low privacy risks? (3) How can the reasoning behind this disparity be leveraged to strengthen privacy inference attacks?

To address the first research question, I focus on conventional classification models and Graph Neuron Networks (GNNs). Specifically, I examine privacy disparities among data subgroups, and identify model memorization and data distribution as the primary factors contributing to privacy leakage disparities. Furthermore, I demonstrate that existing defense mechanisms against MIAs can effectively reduce these disparities. I then extend this reasoning to GNNs, where I define subgroups based on edges that connect node pairs with different combinations of demographic attributes. My research shows that the structural properties of graph data significantly influence the privacy leakage of edges.

To address the second research question, I move the focus to those data samples and machine learning models that have less privacy vulnerability. In particular, I aim to investigate whether they become exposed under stronger attacks, even though they remain safe under conventional MIAs. At sample level, I enhance attacks on interactions in recommender systems. At model level, I introduce TS-MIA, the first black-box MIA specifically designed to target well-generalized time series classification models.

Finally, I address the third research question by exploring the possibility of leveraging the underlying causes of disparity to enhance attack effectiveness. In particular, I identify that different edges in a graph exhibit varying levels of privacy leakage depending on their structural similarities. Building on this insight, I propose Poisoning Link Membership Inference Attack (PMIA), a novel black-box LMIA that exploits graph poisoning to manipulate node similarities and improve attack success rates.