Date & Time: 5:00 PM, 11 Aug 2025.

Venue: Room No: KD-101, Kadim Diwan Building, Dept of CSE

Speaker: Dr. Sumanta Sarkar, Lecturer, University of Essex, United Kingdom.

Title: Privacy-Preserving Data Deduplication for Enhancing Federated Learning of Language Models

Abstract:

Deduplication is a vital preprocessing step that enhances machine learning model performance and saves training time and energy. However, enhancing federated learning through deduplication poses challenges, especially regarding scalability and potential privacy violations if deduplication involves sharing all clients' data. In this talk, we present a pioneering protocol, Efficient Privacy-Preserving Multi-Party Deduplication (EP-MPD), that addresses the problem of deduplication in a federated setup. It efficiently removes duplicates from multiple clients' datasets without compromising data privacy. EP-MPD is constructed in a modular fashion, utilizing two novel variants of the Private Set Intersection protocol. Our extensive experiments demonstrate the significant benefits of deduplication in federated learning of large language models. For instance, we observe up to 19.62% improvement in perplexity and up to 27.95% reduction in running time while varying the duplication level between 10% and 30%. EP-MPD effectively balances privacy and performance in federated learning, making it a valuable solution for large-scale applications. This work was accepted for presentation at NDSS 2025.

Bio:

Dr Sumanta Sarkar is a Lecturer in the School of Computer Science and Electronic Engineering at the University of Essex in the United Kingdom. His research lies at the intersection of post-quantum cryptography, privacy-preserving machine learning, and lightweight cryptographic design—areas that are critical to securing the next generation of digital systems.

He brings a unique combination of industry and academic research experience. He spent over five years at TCS Innovation Labs in Hyderabad.

After completing his PhD at the Indian Statistical Institute, Kolkata, he held postdoctoral positions at INRIA Paris and the University of Calgary. He later served as an Assistant Professor (Research Focused) at the University of Warwick. Dr Sarkar also contributes to national standards development as a member of the British Standards Institution's IST/33/2 committee on cryptography and security mechanisms.