Privacy Under Attack: Securing Federated Clustering

A Practical Analysis of Membership Inference Attacks and Differential Privacy

Urvi Gupta (22B1006) **Ojas Maheshwari (22B0965)**
Department of Computer Science & Engineering
Indian Institute of Technology Bombay

Course: CS6007: Multi-Agent Machine Learning
Instructor: Prof. Avishek Ghosh

September 30, 2025

Abstract

Federated Learning (FL) is a transformative paradigm for training machine learning models on decentralized data, preserving user privacy by keeping raw data on-device. Unsupervised tasks, such as clustering, are critical applications within this domain, with Federated K-Means (‘FedKMeans‘) being a standard approach. However, the iterative model updates communicated between clients and the central server in ‘FedKMeans‘ are not inherently private. These updates, while aggregated, can leak subtle information about the underlying client data distributions. This leakage creates a significant vulnerability to privacy attacks, most notably Membership Inference Attacks (MIAs), where a malicious actor can analyze model updates to infer whether a specific data point was included in a client’s local dataset.

This project will conduct a practical security analysis to investigate this vulnerability. We will implement and compare two federated clustering algorithms: the standard ‘FedKMeans‘ and a privacy-preserving alternative, ‘FedDP-KMeans‘, which integrates Differential Privacy (DP). Our methodology involves two key stages. First, we will establish a baseline by evaluating the clustering performance of both algorithms on a benchmark dataset. Second, we will design, simulate, and quantify a Membership Inference Attack aimed at a target client. This attack will analyze the trajectory of global cluster centroids to make inferences, and its success will be measured using standard metrics such as accuracy, precision, and recall.

We hypothesize that the MIA will achieve a high success rate against the standard ‘FedKMeans‘ model, confirming the existence of a significant privacy leak. Conversely, we expect the attack’s performance against ‘FedDP-KMeans‘ to degrade to that of a random guess (approximately 50% accuracy). The calibrated noise introduced by Differential Privacy should effectively mask the influence of individual data points, rendering the attack ineffective. The results of this project will provide a clear, quantitative demonstration of the necessity and efficacy of DP in securing federated clustering systems against privacy violations.