Privacy Under Attack: Securing Federated Clustering

A Practical Analysis of Membership Inference Attacks and Differential Privacy

Presented by:
Urvi Gupta (22B1006) Dept. of CSE
Ojas Maheshwari (22B0965) Dept. of CSE

Course: CS6007: Multi-Agent Machine Learning

Instructor: Prof. Avishek Ghosh

Indian Institute of Technology Bombay

September 30, 2025

The Problem: The Hidden Risk in Collaboration

Standard Federated K-Means is Vulnerable

Federated K-Means (‘FedKMeans‘) allows collaborative clustering on private data, but the model updates themselves are not secure [2].

- The updates sent from a client to the server are a direct reflection of that client’s local data.
- This creates a subtle information leak.
- **The Threat:** An adversary can analyze these updates to infer whether a specific person’s data was used in training. This is a **Membership Inference Attack (MIA)** [2].

This vulnerability undermines the core privacy promise of Federated Learning.

The Solution: Differential Privacy

FedDP-KMeans: A Provably Private Defense

The ‘FedDP-KMeans‘ algorithm directly counters this threat by integrating Differential Privacy [1].

- **Privacy via Noise:** Before sending updates, clients add a carefully calibrated amount of statistical noise.
- **Plausible Deniability:** This noise masks the exact contribution of any single data point. An attacker can no longer tell if a change in the model was due to a specific person’s data or just the random noise.
- **The Goal:** To make the attacker’s inference no better than a random guess.

Key Assumptions & Experimental Setup

Theoretical Assumptions

- Client data is generated from a **mixture of Gaussians**.
- The underlying clusters are **well-separated**.
- Server data contains at least one sample from **every cluster**.

Our Experimental Setup

Following the paper, we will adopt these practical approaches:

- **Simulating OOD Server Data:** Server data will be 2/3 in-distribution (Gaussian) and 1/3 out-of-distribution (uniform noise) to test robustness and show algorithm works even with imperfect data at server. This data is used for appropriate initialization.
- **Clipping Data Norms:** We will clip data point norms to a fixed value (Δ) to enforce sensitivity bounds, a standard practice in applied DP.

Project Plan: 1. Implementation & Setup

Algorithm and Attacker Implementation

We will implement two algorithms and one attack model in Python:

- **‘FedKMeans‘:** The standard, vulnerable federated clustering algorithm.
- **‘FedDP-KMeans‘:** The privacy-preserving version that uses Differential Privacy as a defense [1].
- **Membership Inference Attacker:** A model designed to analyze global cluster updates and predict membership [2].

Project Plan: 2. Experiments

We will conduct two primary experiments:

- **Experiment A: Privacy-Utility Trade-off Analysis**
  - Replicate the core experiments from the 'FedDP-KMeans' paper [1].
  - We will evaluate clustering performance (k-means cost) across a range of privacy budgets (ϵ) to analyze the trade-off between privacy and model utility.
- **Experiment B: Simulating Membership Inference Attack**
  - We will launch a practical MIA against both 'FedKMeans' and 'FedDP-KMeans'.
  - This will show the real-world consequence of privacy by testing if an attacker can identify training data in a live simulation.

Experiments will use both synthetic data for clear visualization and the FEMNIST dataset for a realistic scenario.

Project Plan: 3. Analysis & Evaluation

Quantifying Performance and Vulnerability

Each experiment will have distinct success metrics:

- For **Experiment A (Privacy-Utility Analysis)**, we will plot the k-means cost vs. epsilon (ϵ). Lower cost indicates better clustering. We will generate Pareto-optimal curves to benchmark performance, as done in the paper [1].
- For **Experiment B (Attack Simulation)**, success will be measured using standard classification metrics:
  - **Attacker Accuracy:** How often the attacker's guess is correct.
  - **Precision & Recall:** To evaluate the reliability of the attack.

This dual analysis provides both a performance benchmark and a practical demonstration of security.

Expected Outcomes & Significance

- **Quantitative Proof of Vulnerability:** We expect the MIA to achieve high accuracy against 'FedKMeans', providing concrete evidence of the privacy risk.
- **Demonstration of Defense:** We predict the attacker's accuracy against 'FedDP-KMeans' will drop to near 50% (a random guess), proving the effectiveness of Differential Privacy.
- A reproducible, open-source simulation of a critical privacy attack and its countermeasure.
- **Broader Impact:** Our findings will highlight that true privacy in FL requires more than just keeping data local; it demands provable guarantees like Differential Privacy.

References I

J. Scott, C. H. Lampert, and D. Saulpic.
Differentially Private Federated k-Means Clustering with Server-Side Data.
arXiv preprint arXiv:2506.05408, 2025.

R. Shokri, M. Stronati, C. Song, and V. Shmatikov.
Membership Inference Attacks Against Machine Learning Models.
In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017 R. Shokri, M. Stronati, C. Song, and V. Shmatikov.
Membership Inference Attacks Against Machine Learning Models.
In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017.