# Analysis on the Vulnerability of Multi-Server Federated Learning Against Model Poisoning Attacks

Lazar Nenovski, TU Delft

## Background

- **Federated Learning** - A more collaborative approach to machine learning.
- **FedMes[1]** - A multi-server model for federated learning focusing on efficiency.
- **Untargeted attack** - An attack where one or more malicious participant aims to bring the accuracy of the aggregate model down.

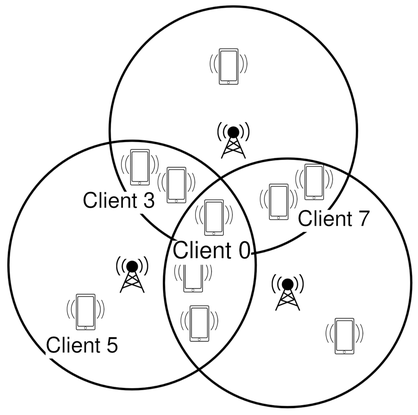


Figure 1: An example of a Fedmes network.

## Research Questions

- Are existing attacks, designed around single-server, effective on a multi-server network, and if so by how much?
- Is the effectiveness of a malicious client in a multi-server network dependent on their location?
- Does the data sharing necessary for multi-server network to operate provides necessary information for conducting an effective attack, without the need to compromise additional parties?

## Methodology

In order to investigate the effects of an untargeted attack a simulation of a FedMes multi-server network was carried out, where some of the simulated clients are malicious. The results are compared if changing the location had an influence on the effectiveness of the attack. Then two novel attacks where tested in the same manner.

## Results


Figure 2: Comparison of the effects of an multi-server adaptation of the single-server attack MinMax[2] carried on a network from differently located malicious client. The malicious client in an overlapped region is able to be very effective, while those within a reach of a single server can barley affect the accuracy.

## References

[1] Dong-Jun Han, Minseok Choi, Jungwuk Park, and Jaekyun Moon. Fedmes: Speeding up federated learning with multiple edge servers. *IEEE Journal on Selected Areas in Communications*, 39(12):3870–3885, 2021.

[2] Virat Shejwalkar and Amir Houmansadr. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *NDSS*, 2021.


Figure 3: The effects of the proposed two novel attacks. We are able to achieve notably better results, by utilizing the information being sent to the malicious clients, without needing a method to obtain the data of benign clients.

## Conclusions

If a malicious party is able to also compromise the communication channel of every client across the network, then existing single-server attacks are sufficient to bring down the accuracy of the model. If the malicious party is only able to compromise neighbouring clients, then the effect is dependent on the amount of cells within reach. In the case that the adversary is aware that the compromised clients are part of a multi-server FL network, they can exploit the data passing through such clients in order to launch a more severe attack.