Defense against Local Model Poisoning Attacks to Byzantine-Robust Federated Learning

Shiwei LU, Ruihu LI, Xuan CHEN, Yuena MA

Frontiers of Computer Science, DOI: 10.1007/s11704-021-1067-4

Problems： Local model poisoning attacks can iteratively attack Byzantine-robust aggregation rules successfully, and thus make these secure aggregation rules failure in Federated Learning.

Existing defenses have three limitations:

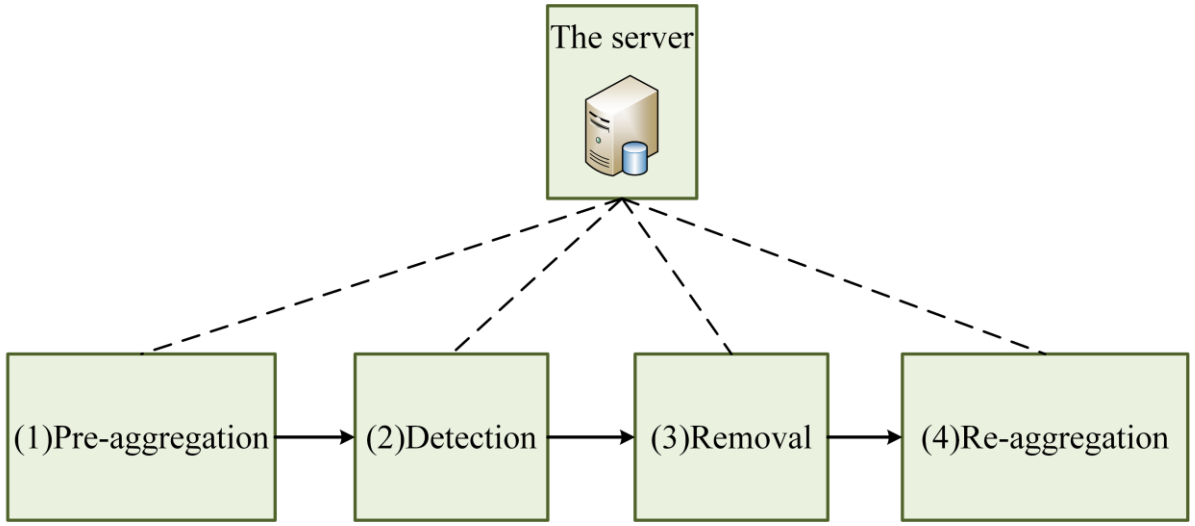
(1)The auxiliary dataset is needed to verify error rate effect or loss function effect of each local model on the aggregated model.

(2)The detection accuracy needs to be further improved.

(3)High time cost.

Ideas: To defend against local model poisoning attack, we propose a defense paradigm (PDRR). According to the vulnerability of local model poisoning attack, targeted defense is deployed.

By pre-aggregating all local models and calculating the similarity between pre-aggregated model and local model, we can distinguish malicious model from benign models. The paradigm of our defense is shown as follows,

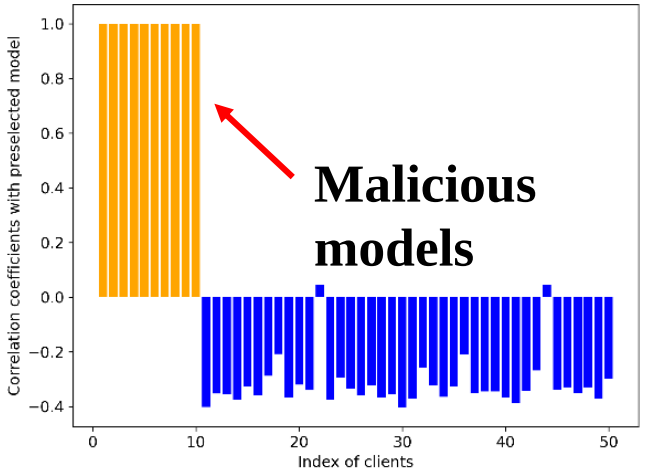


The defense paradigm (PDRR) against local model poisoning attack

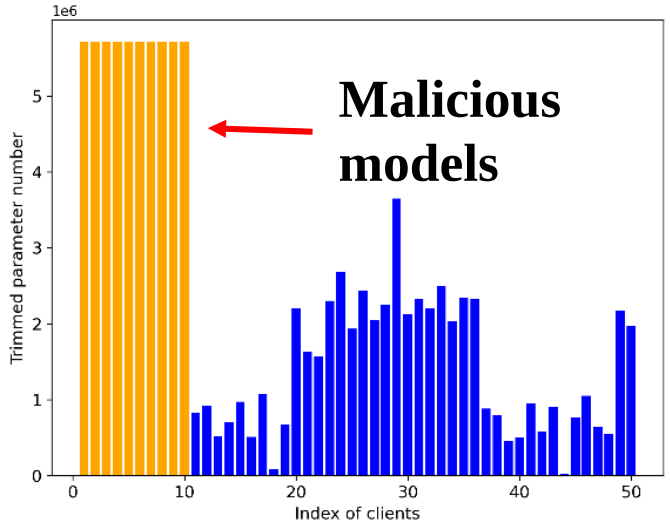
Main Contribution:

- The similarity measurements of each defense are shown as follows,

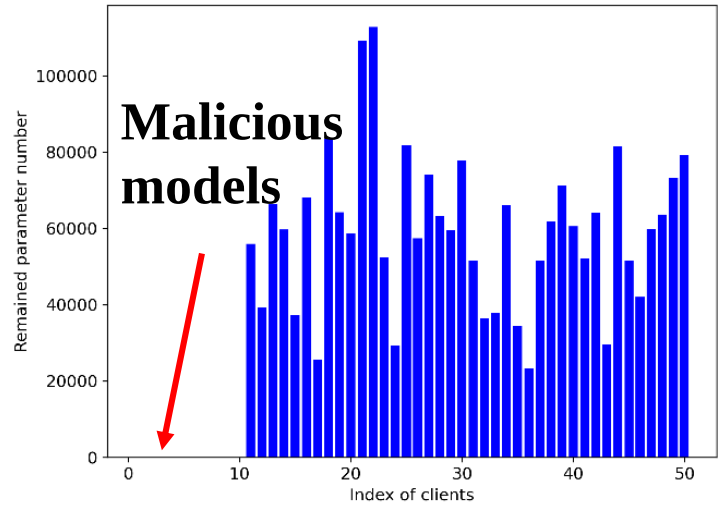
Detection for Krum attack



Detection for Trimmed-mean attack



Detection for Median attack



We can observe that the detection scheme of our defense for three attacks can distinguish malicious models from benign models successfully.