F2ED-LEARNING: Good Fences Make Good Neighbors

Lun Wang, University of California, Berkeley
Qi Pang, Hong Kong University of Science and Technology
Shuai Wang, Hong Kong University of Science and Technology
Dawn Song, University of California, Berkeley

NeurIPS-SpicyFL 2020
Thirty-fourth Conference on Neural Information Processing Systems
December 5th-12th, 2020 - Virtual Only

Bi-directional security concern in FL

Despite the prosperity of FL, multiple privacy or security concern still exists in today's FL pipeline. These concerns can be roughly captured by two threats as shown in Figure 1.

1. **Semi-honest server**: The centralized server might be interested in the client's data for profitable purpose. However, due to the regulation pressure, server tends to infer information from users' legitimate update instead of actively launching attacks.
2. **Malicious user**: The users, on the other hand, can launch whatever kind of attacks due to the anonymity of the identity. Users might launch attacks due to various reasons like profits, competition or even mischief.

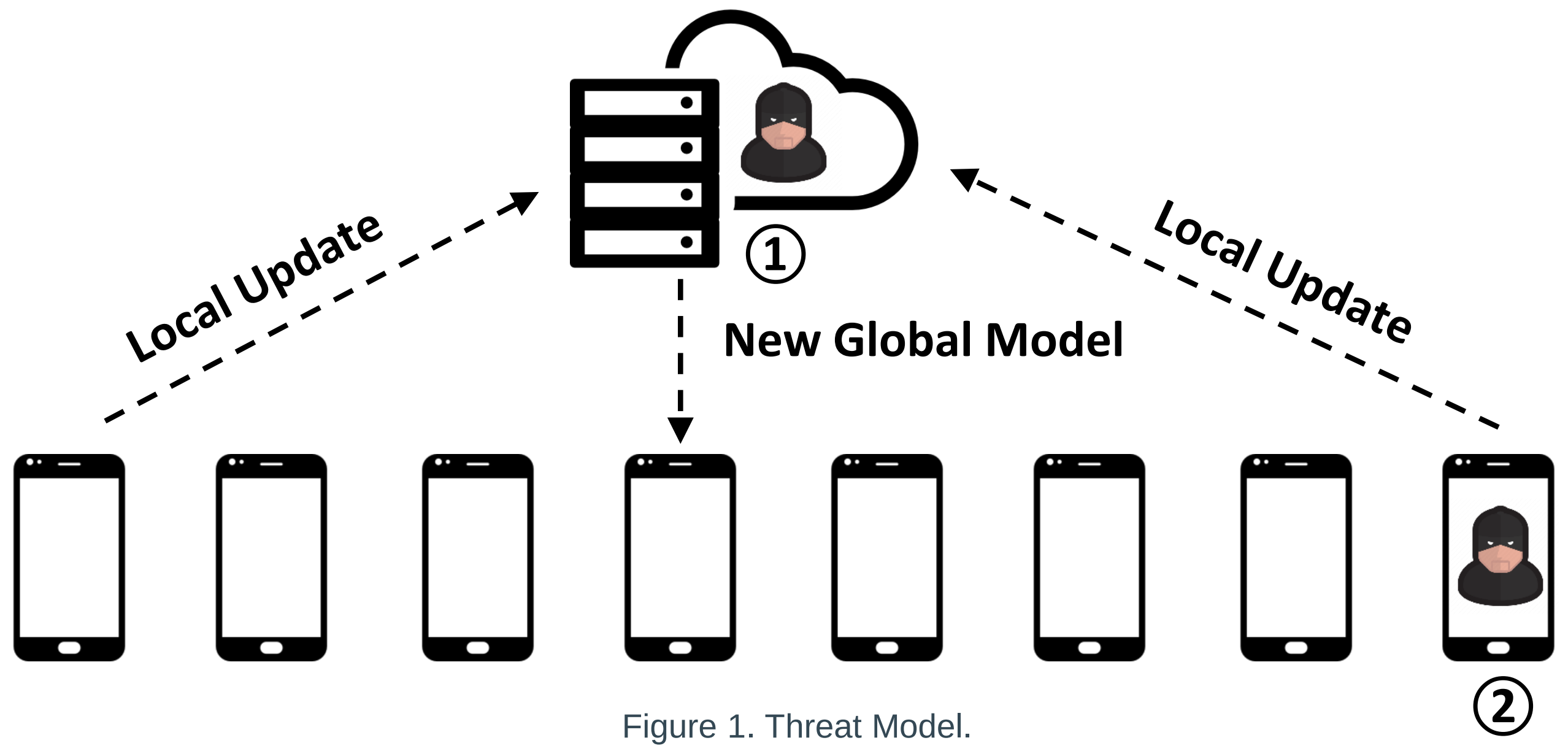


Figure 1. Threat Model.

Secure FL

Defend against semi-honest server in FL

Secure aggregation is developed by [1] to defend against the semi-honest server in FL. Secure aggregation allows the server to obtain the sum of the clients' updates but hides the individual updates cryptographically.

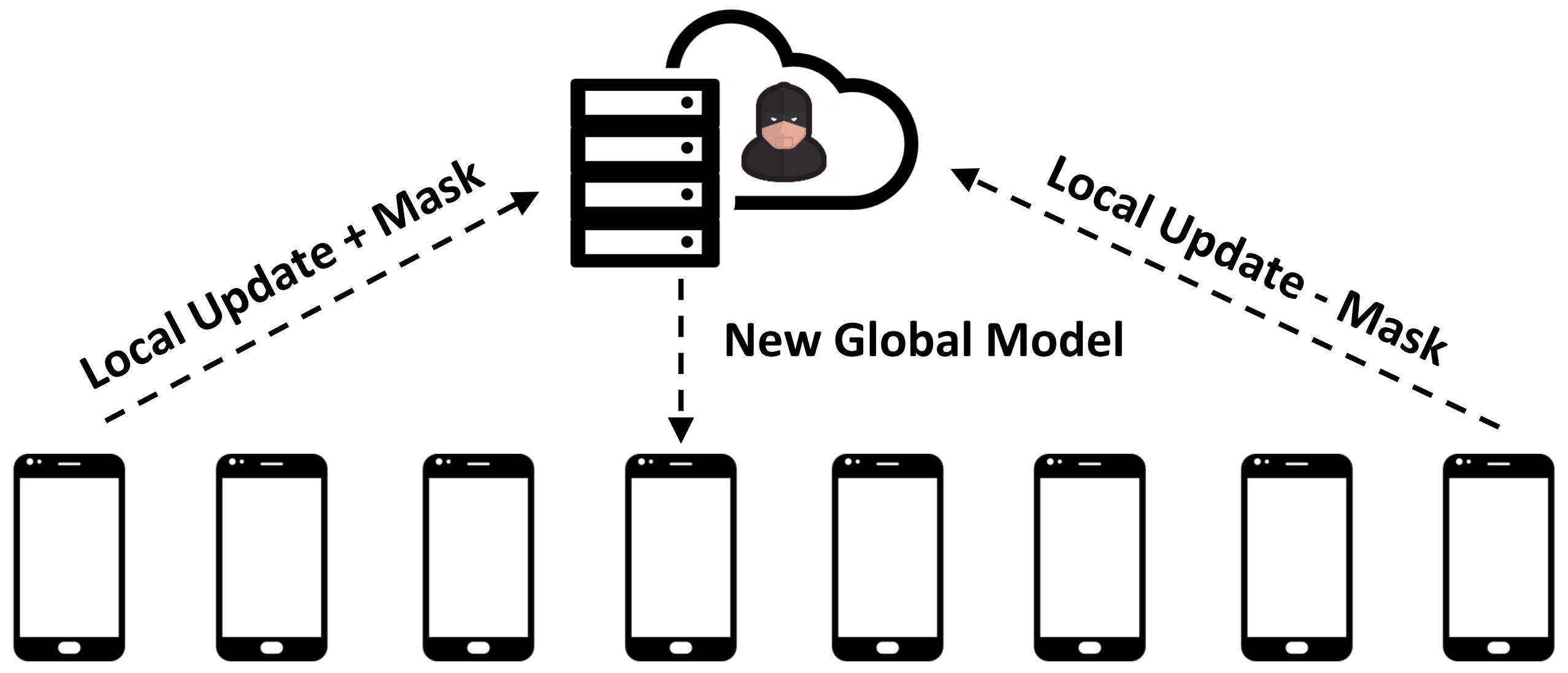


Figure 2. Secure Aggregation.

However, in our threat model, vanilla secure aggregation is insufficient since it provides no protection for the server. As the individual updates are completely hidden, there is no way that the server can identify the malicious clients even after detecting the attack.

Robust FL

Defend against malicious users in FL

The core step in federated learning is to estimate the true mean of the benign updates as accurate as possible even with malicious clients.

- The most commonly used aggregator, averaging, is proven to be vulnerable to even only one malicious client.
- All other works addressing the issue such as [3, 4, 5] suffer from a dimension-dependent estimation error. Such error is unacceptable even for training a 3-layer MLP on MNIST, not to mention more complicated tasks and models such as VGG16 or ResNet50.

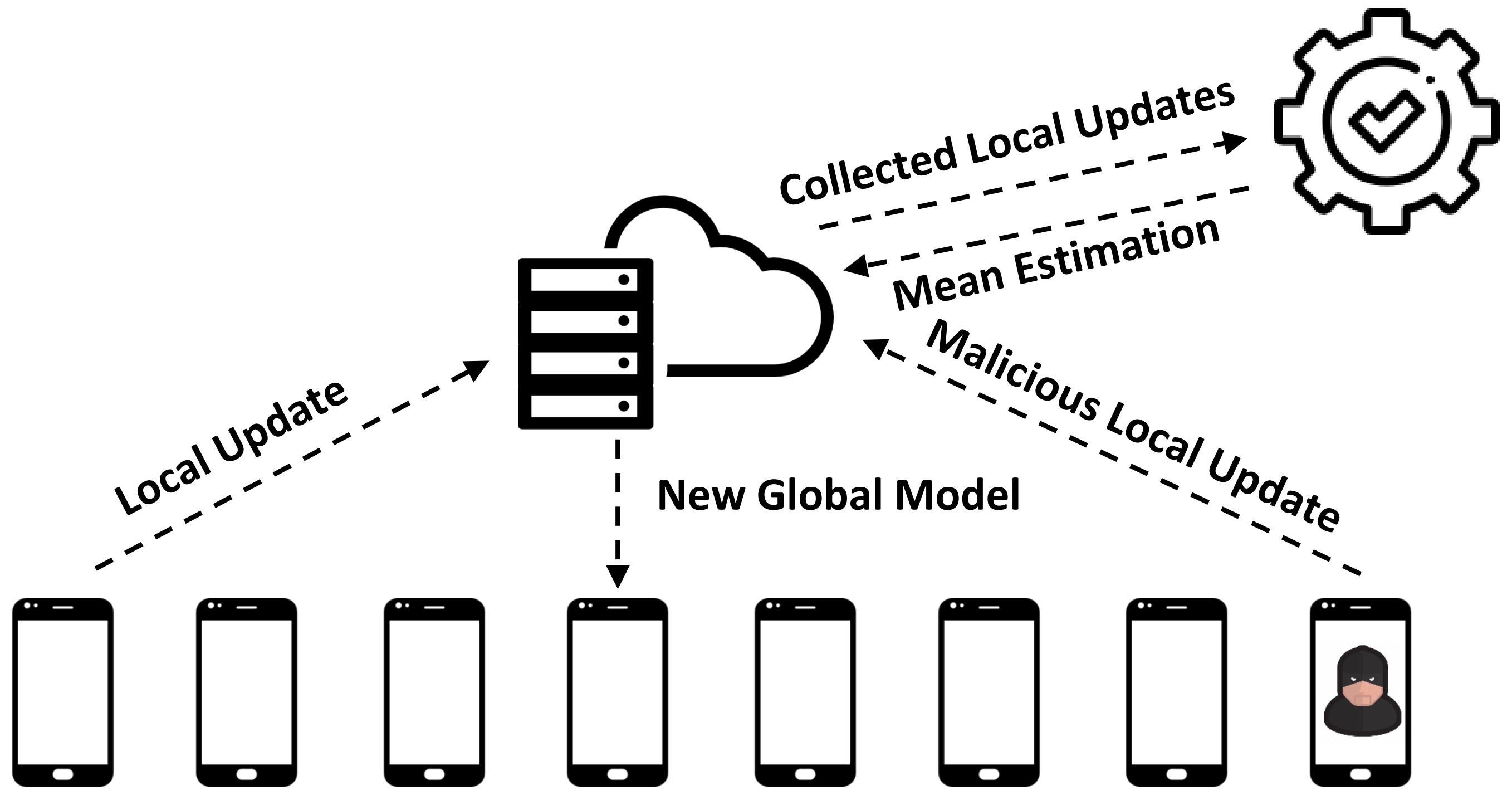


Figure 3.Robust Aggregation.

Actually, the problem is well studied in statistics under the name "robust mean estimation" and there already exist several robust mean estimators with dimension-free estimation error ([2]). Therefore, instead of reinventing the wheel, we choose to leverage a representative robust mean estimator: FilterL2 [2].

F2ED-LEARNING: Reconcile Robustness and Security via Sharding

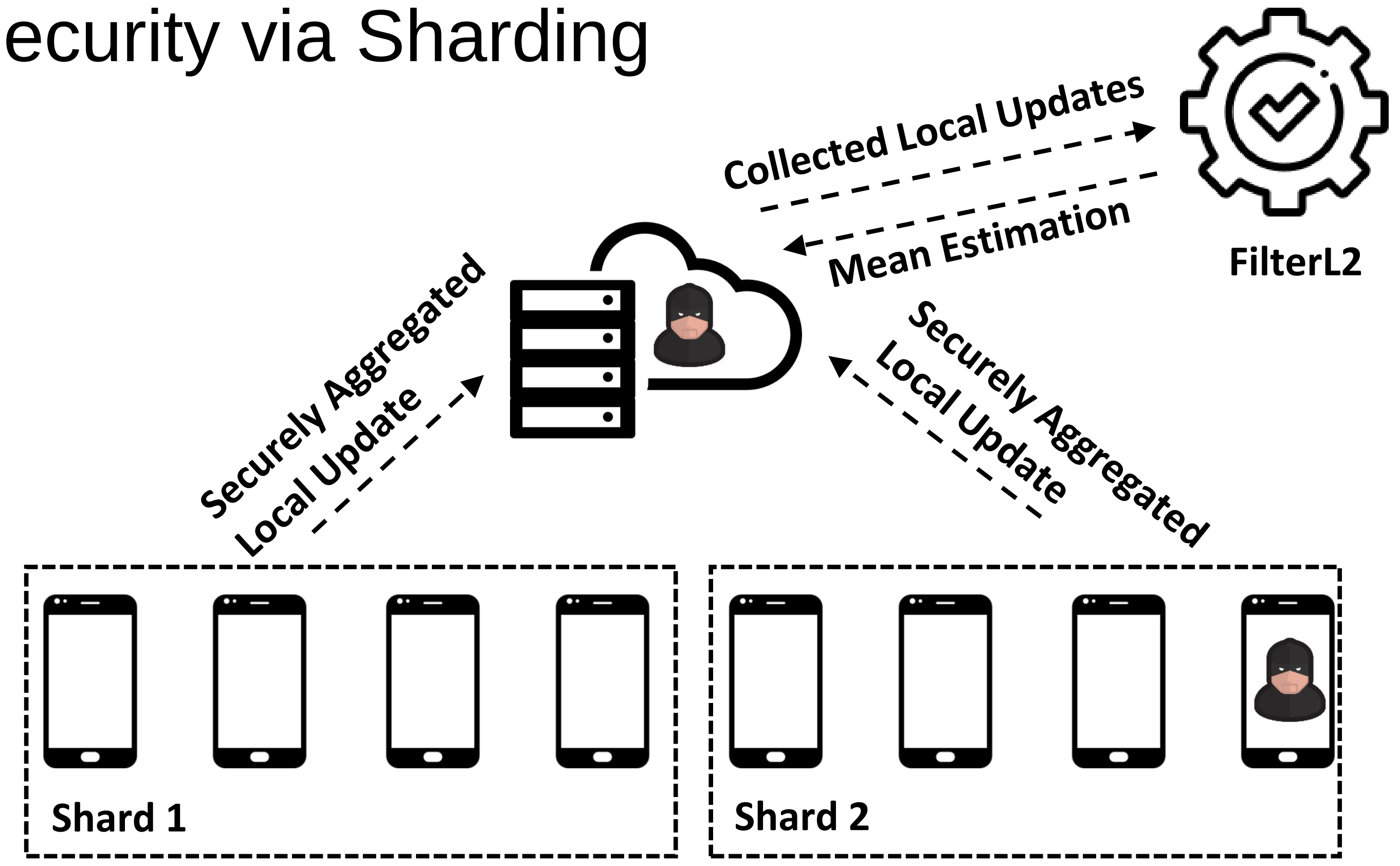
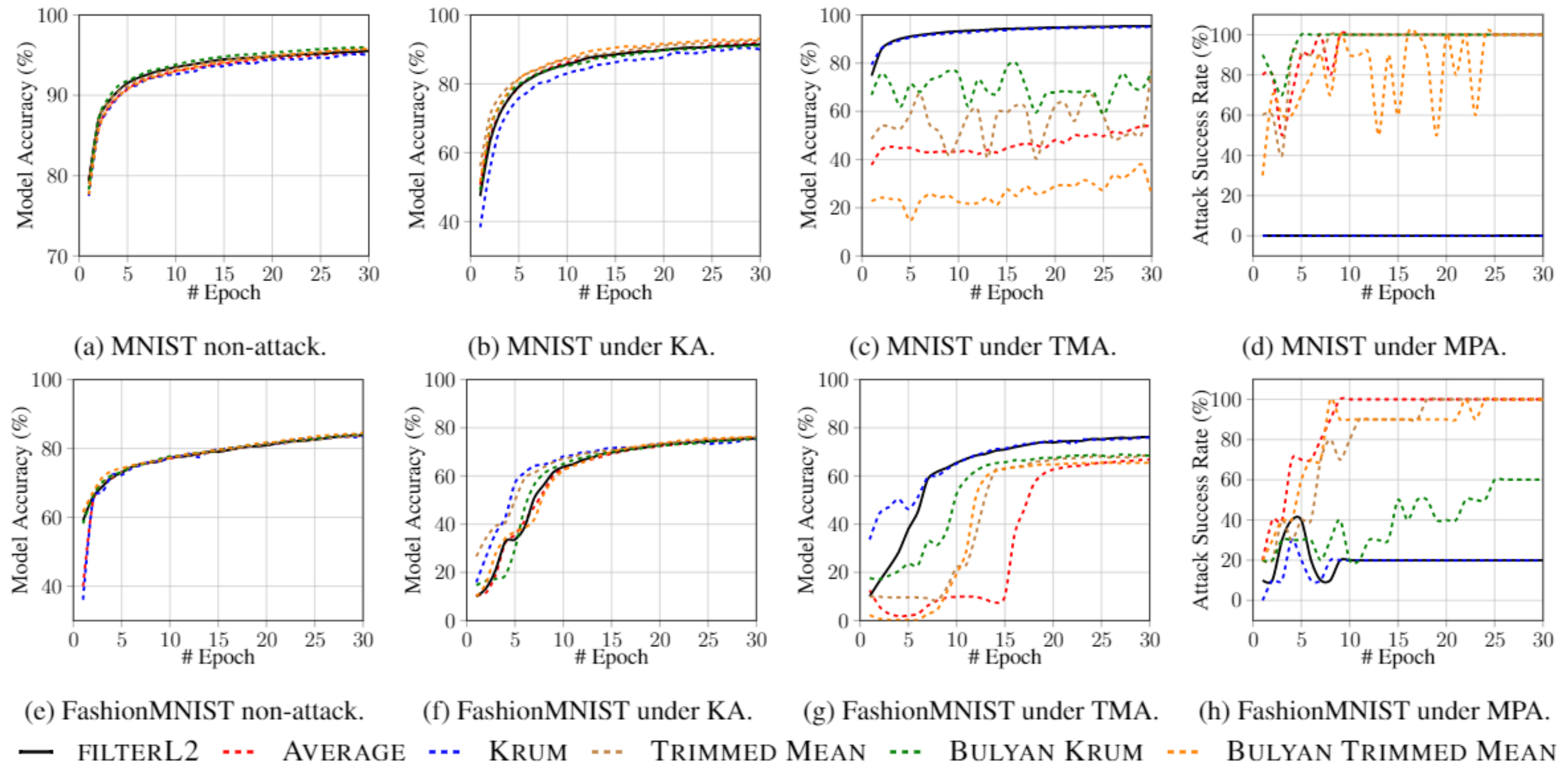

Collected Local Updates
Mean Estimation
FilterL2
Securely Aggregated Local Update
Securely Aggregated Local Update
Shard 1
Shard 2

Figure 4.Sharded Secure Aggregation w/ Robust Aggregation.

The biggest challenge is that robust FL protocols have incompatible implementation with secure aggregation techniques. The robust estimators have to access local updates while secure aggregation hides them from the server.

To address the issue, we propose **FED**ERATED **LEARNING** WITH **F**ENCE, abbreviated **F2ED-LEARNING**. F2ED-LEARNING integrates FilterL2 [2] and secure aggregation [1] to defend against both the Byzantine malicious clients and the semi-honest server. Specifically the clients are split into multiple shards, the local updates from the same shard are securely aggregated at the centralized server, and the robust estimator is run on the aggregated local updates from different shards (Figure 4).



(a) MNIST non-attack. (b) MNIST under KA. (c) MNIST under TMA. (d) MNIST under MPA.

(e) FashionMNIST non-attack. (f) FashionMNIST under KA. (g) FashionMNIST under TMA. (h) FashionMNIST under MPA.

Figure 5.F2ED-LEARNING evaluation.

The evaluation (Figure 5) shows that F2ED-LEARNING consistently achieves optimal or close-to-optimal performance under three attacks among five robust FL protocols.

References

[1] Bonawitz, Keith, et al. "Practical secure aggregation for privacy-preserving machine learning." *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017.

[2] Steinhardt, Jacob. *Robust learning: Information theory and algorithms*. Diss. Stanford University, 2018.

[3] Blanchard, Peva, Rachid Guerraoui, and Julien Stainer. "Machine learning with adversaries: Byzantine tolerant gradient descent." *Advances in Neural Information Processing Systems*. 2017.

[4] Yin, Dong, et al. "Byzantine-robust distributed learning: Towards optimal statistical rates." *arXiv preprint arXiv:1803.01498* (2018).

[5] Mhamdi, El Mahdi El, Rachid Guerraoui, and Sébastien Rouault. "The hidden vulnerability of distributed learning in byzantium." *arXiv preprint arXiv:1802.07927* (2018).

[6] Fang, Minghong, et al. "Local model poisoning attacks to Byzantine-robust federated learning." *29th {USENIX} Security Symposium ({USENIX} Security 20)*. 2020.

[7] Bhagoji, Arjun Nitin, et al. "Analyzing federated learning through an adversarial lens." *International Conference on Machine Learning*. PMLR, 2019.