On the Fragility of Contribution Evaluation in Federated Learning

B Pejó, M Frank, K Varga, P Veliczky, G Biczók
BME / CrySyS / E-Group / HUN-REN

TLDR

Problem: Contribution scoring is critical for FL fairness but highly vulnerable to distortion.

Two Key Threats:

- **Architectural Sensitivity**: Advanced model aggregation methods (beyond simple averaging) unintentionally skew final scores.
- **Intentional Manipulation**: Malicious poisoning attacks can be used to artificially inflate an attacker's score or deflate others'.

Evidence: Extensive experiments via the Flower framework prove that both aggregation choice and attackers significantly bias results.

Take Away: Current evaluation schemes are unreliable; more robust metrics are essential for fair Reward Allocation in Federated Learning.

Introduction

Federated Learning: Solves data sharing issues while increasing data size but comes with issues:

- **Client Heterogeneity** (aka non-IID clients)
- **Adversarial Risks** (aka Byzantine clients)

The Incentive Problem: In projects like MELLODDY [1] (drug discovery), co-opetition only works if participants are rewarded fairly.

- **The Shapley Value**: The SV is the only "fair" metric but is computationally prohibitive.
- **The Reality**: Real-world systems use approximations for Contribution Evaluation (CE).

Small score shifts have massive economic consequences, i.e., a score change of 0.03 results in a $30,000 redistribution with a $1M reward pool.

Contribution

- **Systematic Sensitivity Analysis**: Conducted the first rigorous study of how advanced aggregators (handling client heterogeneity, Byzantine faults, and data noise) unintentionally distort contribution scores.
- **New Adversarial Attacks**: Designed and implemented two new poisoning attacks targeting the reward system rather than the model's utility:
  - **Self-Improvement (SI)**: Artificially inflates the attacker's own contribution score.
  - **Targeted Decrease (TD)**: Maliciously reduces the score of a specific competitor.
- **Open-Source Framework**: All aggregations, attacks, and evaluation schemes are integrated into the Flower [9] framework.

Aggregation Methods

FedAvg: Computes a weighted average of client updates based on their local dataset sizes.

FedProx [2]: Adding a "proximal term" to the local objective, stopping local models from deviating too far.

FedNova [3]: Scale client updates to eliminate biases caused by varying local training steps.

Krum [4]: Selects the most "central" update based on Euclidean distances to its neighbors to avoid outliers.

Zeno [5]: Ranks updates by calculating their estimated loss reduction on a trusted validation set.

Contribution Scoring

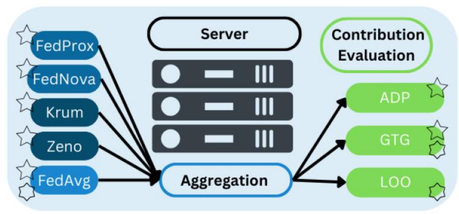
- **Leave-One-Out** [6]: Measures a client's importance by calculating the difference in model utility when that specific client is excluded from the aggregation.
- **Guided Truncated Gradient Shapley** [7]: A scalable approximation of the Shapley Value that uses permutation sampling and performance-based truncation.
- **Adaptive Weighting** [8]: A distance-based metric that rewards clients whose updates consistently align with the global gradient using cosine similarity smoothed over time.

Experimental Setup

	Architectural Sensitivity	Intentional Manipulation
Data	CIFAR-10	ADULT Fashion-MNIST
Dist.	IID, IID + noise non-IID high & low imb.	IID non-IID low imbalance
Arch.	2-layer CNN	3-layer MLP 2-layer CNN
Agg.	Fed -Avg, -Prox, -Nova Krum, Zeno	FedAvg
Adv.	Only benign Label & Gradient flip	Self Improvement Targeted Decrease
CE	GTG, ADP	GTG, LOO

Experimental Settings for Architetural Sensitivity & Intentional Manpulation.

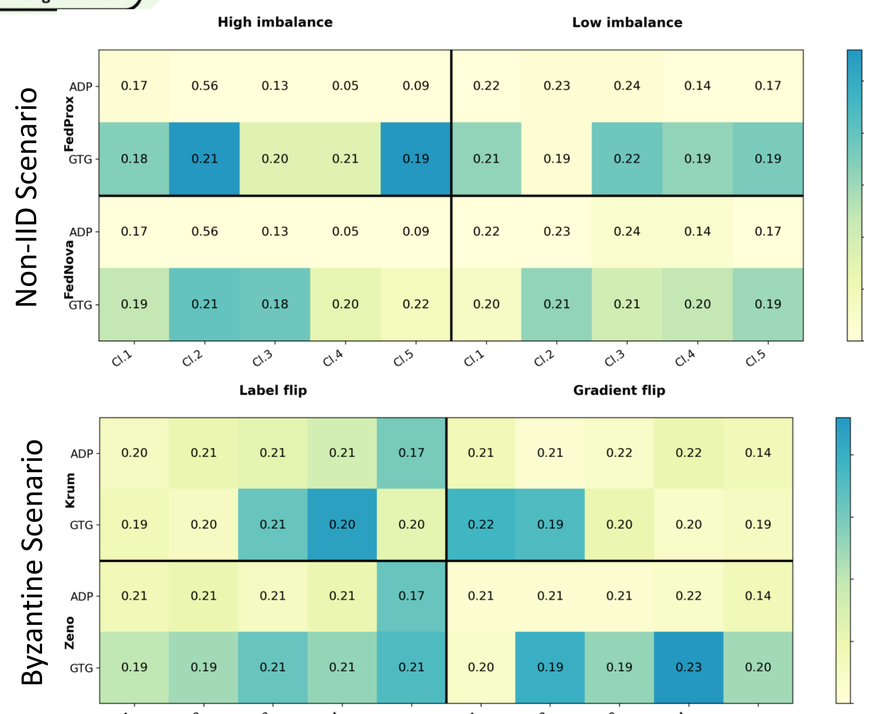
Architectural Sensitivity (marked with a 5-star) uses all aggregation techniques and GTG and ADP as CE.



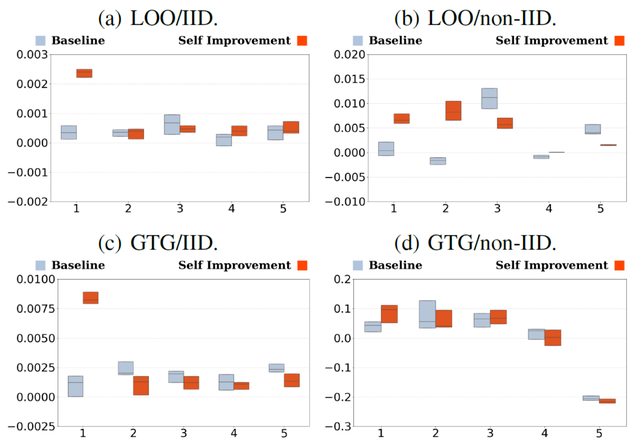
Intentional Manipulation (marked with a 6-star) relies on FedAvg and considers GTG and LOO.

AS Results

Client's CE scores (*values*) and their relation / difference (*color*) to FedAVG-based scores.

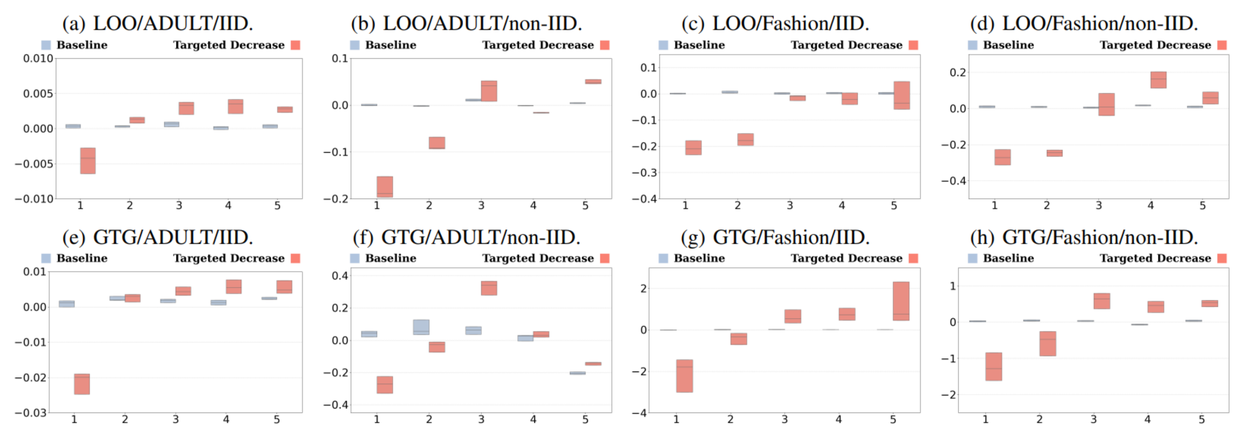


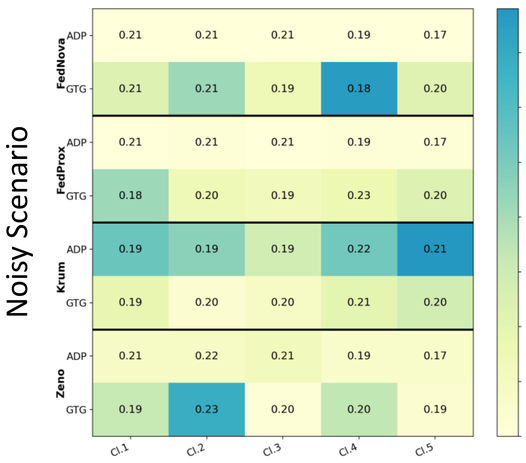
IM Results



Clients' LOO and GTG scores where the 1st client either self improves or not.

Clients' contribution score differences (LOO - blue or GTG - red) when the 1st client is decreasing the 2nd's score.





Statistical Tests

Setting	CE	1	2	3	4	5
High	ADP	0.00	0.25	0.00	0.01	0.01
	GTG	0.25	0.25	0.25	0.25	0.25
Low	ADP	0.00	0.25	0.25	0.25	0.01
	GTG	0.25	0.25	0.25	0.25	0.25
Label	ADP	0.00	0.00	0.00	0.00	0.00
	GTG	0.25	0.25	0.25	0.25	0.25
Grad.	ADP	0.00	0.00	0.00	0.00	0.00
	GTG	0.25	0.25	0.25	0.25	0.25
Noise	ADP	0.00	0.00	0.00	0.00	0.00
	GTG	0.21	0.19	0.25	0.04	0.11

One-sided *t*-test to for the SI attack's success.

		R2	R3	R4	All
IID	LOO	0.00	0.00	0.01	0.00
	GTG	0.00	0.00	0.00	0.00
nonn-IID	LOO	0.01	0.01	0.00	0.00
	GTG	0.08	0.01	0.02	0.00

Andreson - Darling test to check score distribution shift.

Two-sided *t*-test to determine the TD attack's success.

			R2	R3	R4	All
ADULT	IID	LOO	0.05	0.29	0.50	0.83
		GTG	0.02	0.75	0.88	0.01
	nIID	LOO	0.00	0.00	0.00	0.00
		GTG	0.01	0.00	0.02	0.01
Fashion	IID	LOO	0.00	0.00	0.00	0.00
		GTG	0.01	0.06	0.01	0.01
	nIID	LOO	0.00	0.00	0.00	0.00
		GTG	0.00	0.00	0.00	0.00

Conclusions

- The choice of aggregation dictates how a client is valued.
- An attacker can improve the global model while corrupting the rewards.
- Competitors can be silenced, yet the attack is crashing the performance.
- Without robust, attack-resistant reward schemes, the federation could collapse.

Contact

pejo@crysys.hu

Paper

https://tinyurl.com/Frag-ContEval

Acknowledgement

This work was supported by Project No. 145832, implemented with the support provided by the Ministry of Innovation and Technology from the NRDI Fund, financed under the PD23 funding scheme.

References

1. Heyndrickx, Wouter, et al. "MELLODDY: cross-pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information." Journal of chemical information and modeling 64.7 (2023)
2. Li, Tian, et al. "Federated optimization in heterogeneous networks." Proceedings of Machine learning and systems 2 (2020)
3. Wang, Jianyu, et al. "Tackling the objective inconsistency problem in heterogeneous federated optimization." Advances in neural information processing systems 33 (2020)
4. Blanchard, Peva, et al. "Machine learning with adversaries: Byzantine tolerant gradient descent." Advances in neural information processing systems 30 (2017).
5. Xie, Cong, Sanmi Koyejo, and Indranil Gupta. "Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance." International conference on machine learning. PMLR, 2019.
6. Evgeniou, Theodoros, Massimiliano Pontil, and André Elisseeff. "Leave one out error, stability, and generalization of voting combinations of classifiers." Machine learning 55.1 (2004)
7. Liu, Zelei, et al. "Gtg-shapley: Efficient and accurate participant contribution evaluation in federated learning." ACM Transactions on intelligent Systems and Technology (TIST) 13.4 (2022)
8. Wu, Hongda, and Ping Wang. "Fast-convergent federated learning with adaptive weighting." IEEE Transactions on Cognitive Communications and Networking 7.4 (2021)
9. Beutel, Daniel J., et al. "Flower: A friendly federated learning research framework." arXiv preprint arXiv:2007.14390 (2020)