On the Volatility of Shapley-Based Contribution Metrics in Federated Learning

1st Arno Geimer
SnT, University of Luxembourg
arno.geimer@uni.lu

2nd Beltran Fiz
SnT, University of Luxembourg
beltran.fiz@uni.lu

3rd Radu State
SnT, University of Luxembourg
radu.state@uni.lu

Abstract—**Federated learning (FL) is a collaborative and privacy-preserving Machine Learning paradigm, allowing the development of robust models without the need to centralize sensitive data. A critical challenge in FL lies in fairly and accurately allocating contributions from diverse participants. Inaccurate allocation can undermine trust, lead to unfair compensation, and thus participants may lack the incentive to join or actively contribute to the federation. Various remuneration strategies have been proposed to date, including auction-based approaches and Shapley-value-based methods, the latter offering a means to quantify the contribution of each participant. However, little to no work has studied the stability of these contribution evaluation methods. In this paper, we evaluate participant contributions in federated learning using gradient-based model reconstruction techniques with Shapley values and compare the round-based contributions to a classic data contribution measurement scheme. We provide an extensive analysis of the discrepancies of Shapley values across a set of aggregation strategies and examine them on an overall and a per-client level. We show that, between different aggregation techniques, Shapley values lead to unstable reward allocations among participants. Our analysis spans various data heterogeneity distributions, including independent and identically distributed (IID) and non-IID scenarios.**

Index Terms—**Federated Learning, Incentive Mechanism, Shapley Value**

I. INTRODUCTION

Federated Learning (FL), a collaborative method for training Machine Learning models without the entities involved having to share their datasets, has gained a lot of traction in recent years as a prevalent form of collaborative learning [1]. The growing focus on privacy preservation, driven by the enactment of regulations such as the General Data Protection Regulation (GDPR), emphasizes the importance of FL in addressing data privacy concerns.

It has seen adoption in a wide range of domains, from financial corporations and medical institutions to IoT devices, with major industry players such as Nvidia offering in-house platforms and solutions [2]. An important aspect of a federation is the strategy, defining the approach in which the central server computes model updates. Since the inception of Federated Learning, numerous strategies have been introduced to manage specific problems, such as poor performance on heterogeneous datasets [3], attacks by malicious clients [4], and the convergence speed of the global model [5].

In addition to the performance of the model, the evaluation and fair distribution of rewards to participants in a federation is of utmost importance, particularly in commercial settings where financial incentives might be needed for the participation of the client [6]. Each participant would expect to be fairly compensated for their participation, regardless of the underlying aggregation strategy used.

In this work, we use cumulative round-based Shapley values as percentages to represent the contribution of each client. Shapley values, a concept from cooperative game theory, are employed to quantify the contribution of each participant in a federation by evaluating their marginal impact on the model's performance in each communication round. This approach is a popular contribution evaluation method [7] and reflects a practical industry use case where, in each round, the central server assesses the contribution of each client based on their local model updates. Over time, the server aggregates these round-based contributions to determine the overall contribution of each participant, allowing a fair reward distribution. We analyze, across an array of popular aggregation strategies, the overall, as well as per-client, reliability of these generated contribution values.

Focusing on image recognition tasks, our results treat a vast and diverse set of possible Federated Learning scenarios. In addition to using multiple vision datasets popular in FL literature, we simulate data distributions among participants using splits generated using a Dirichlet distribution to control the degree of heterogeneity. We show that, across most tested strategies, round-based Shapley values perform similarly in terms of closeness to the ground truth.

Finally, we study the stability of contributions on a per-client basis and demonstrate that round-based Shapley values cause instability on client-level contributions, rendering them potentially unfit for real-world usage.

The contributions of this paper can be summarized as follows:

(i) We analyze the properties and effects of the weight assignment parameter in round-based contribution calculation.
(ii) We perform extensive experiments studying Federated Learning contribution allocation across a broad range of popular aggregation strategies, providing a novel comprehensive comparative analysis.

(iii) We discover an inherent instability of Shapley-based contribution evaluation methods across popular aggregation strategies and examine this instability in a cross-silo federation.

II. BACKGROUND

A. Federated Learning

A basic Federated Learning architecture consists of a group of participants, or clients, who each own a private local dataset, along with a central server or aggregator. Before training starts, they agree on the model architecture and hyperparameters such as learning rates, the aggregation strategy, and local epoch numbers.

Subsequently, the central server starts the federation by initializing and distributing the first global model. At the onset of each round, the aggregator selects a subset of clients and transmits them the current global model. After receiving the current global model, users train their respective data for the e epochs and upload their new model to the server. The number of local epochs, e, has a serious impact on the federation [8]: A low value requires more rounds of communication until convergence is achieved, which can be problematic when participants are low-resource IoT devices. However, choosing a value that is too high negatively impacts the performance of the final model, leading to the need to find a balance between performance and communication overhead.

After receiving all chosen client models, the server uses the agreed-upon strategy as an aggregation function to compute the new global model. Repeating this process until satisfactory performance has been achieved allows the collaborating entities to obtain a model incorporating knowledge of all clients, which has been shown to outperform locally trained models [8].

Due to the variety of applications in Federated Learning, two primary setups have emerged. In a **cross-device** federation, a large number of distributed clients, usually with small volumes of data, collaborate to build a robust model. On the other hand, **cross-silo** setups involve few participants with large amounts of data working together to create a global model that incorporates the specific knowledge of each entity. In an environment comprised of profit-driven entities, correctly assessing contributions is of utmost importance, as participants may invest money in infrastructure and compliance in hopes of receiving their equitable part of the payout. We focus on cross-silo federations, as they are a more appropriate use case for contribution calculations.

B. Aggregation strategies

In Federated Learning, different aggregation strategies may be employed based on the specific circumstances and objectives of the participants.

- FEDAVG [8] is the baseline Federated Learning aggregation technique.

- Federated Averaging with Momentum [9], or FEDAVGM, improves the poor performance of FEDAVG on heterogeneous data by adding momentum when updating model weights.

- FEDADAGRAD, FEDADAM, FEDYOGI, all three proposed in [5], employ advanced gradient-based optimization methods to improve the convergence speed of the federation. With clients' weight updates considered a pseudo-gradient, they use ADAGRAD, ADAM and YOGI, respectively.

- To reduce potentially adversarial updates, FEDMEDIAN, FEDTRIMAVG [10] update the global model as the median and trimmed mean of client updates, as opposed to a weighted average in FEDAVG. In adversarial settings, this has been shown to guarantee substantial improvements over the baseline.

- Finally, KRUM [11], similarly to the two previous methods, is a modified aggregation to counter adversarial clients. It excludes client updates that are too far away from the other clients.

All proposed strategies share the global objective function

$$f(w) = \frac{1}{K} \sum_{k=1}^{K} F_k(w).$$

Although our investigation includes aggregation strategies aimed at improving defenses against model attacks, we do not include adversarial participants.

C. Shapley values

In game theory, a cooperative game of n persons is a game in which subsets of players, called coalitions, can cooperate to obtain a utility v [12], which they may distribute among themselves. To determine the exact contribution of each player in the coalition to v, Shapley values [13] can be used. Specifically, let S be a coalition of players, then the contribution ϕ of player $i \in S$ to v can be determined as

$$\phi_i(v) = \sum_{S \subseteq N \setminus i} \frac{|S|!(n - |S| - 1)!}{n!} \Big[v(S \cup i) - v(S) \Big] \quad (1)$$

Shapley values have a range of desirable qualities, among which the property that players who bring no value to a coalition have Shapley value zero. In addition, players who bring the same value also receive the same contribution values. In fact, it has been shown in [14] that the Shapley value is the only contribution measure that satisfies these properties. This is particularly important in an FL setting, where adversarial parties might try to unfairly increase their perceived contribution by joining the federation as multiple participants or artificially inflating their data.

III. RELATED WORK

Several approaches have been devised to evaluate participants' contributions during Federated Learning training. We give a brief overview of the most common approaches, which

are either based on numeric computations or on self-disclosed information and estimations by the central server.

Self-reported: The most direct approach seen in previous work is to have participants perform their own reporting, providing information regarding their local model training process [15], [16]. The information provided may encompass aspects such as the quality and quantity of data provided by the participant; costs associated with data collection, processing, and communication in the federation. A limitation of these approaches is the assumption of honesty and the ability of clients to evaluate their own conditions. Reputation-based client selection techniques have been proposed to mitigate this issue [17]

Influence: Another approach to quantifying the contribution of participants are influence-based methods. These apply a systematic method for quantifying the impact of individual data points on the global model [18]. Influence can be computed by assessing the disparity in models when trained with and without specific data points. This approach was used by [19] to provide incentives for clients to provide high-quality data during the training process. Although the influence metric offers detailed insights, its computation can be challenging due to the expensive nature of retraining models.

Auctions: In diverse areas of Federated Learning, auctions serve as a fundamental economic mechanism to allocate resources such as training data and computational power, setting prices through bidding [20]. Although contribution estimation is an important aspect, other key concepts in auction theory include valuation, utility, and social welfare, with the ultimate goal being to enhance the efficiency and effectiveness of resource allocation in FL environments.

Shapley values: Used in game theory to evaluate the contribution of players to a common task, Shapley values are the most common approach to numerically assess individual clients' contribution to the shared model in FL.

Shapley values in Federated Learning

Although it has been proposed to calculate Shapley values in FL by evaluating the performance of fully trained federated models on subsets of participants [21], a more common approach, called One-Round Reconstruction, is to use gradient-based model reconstruction [22] which assesses contributions at each federation round, possibly using Monte Carlo methods [23] to speed up the calculation. We will use this method to calculate per-round contributions of all clients, with final contribution values derived from the sum of each client's Shapley values. A detailed calculation can be found in Algorithm 1. One-Round Reconstruction is of particular interest in continuous learning tasks, where payouts may occur periodically instead of just once.

Note that the calculation of ω_S depends on the chosen aggregation technique and may not be a simple average as shown above. Although the literature extensively discusses various evaluation methods and their respective advantages and disadvantages from complexity and communication stand-

---

Algorithm 1: One-round reconstruction of Shapley values with FEDAVG; n_k is the Shapley value and dataset size of client k. C is the set of all clients, with size K.

Server executes:
for each round $t = 1, 2, \dots$ **do**
    for each client $k \in C$ **do**
        $\phi_k^t = \sum_{S \subseteq C \backslash k} \frac{|S|!(K-|S|-1)!}{K!} F_k^S$
        where $F_k^S = \left[F_k(\omega_{S \cup k}) - F_k(\omega_S) \right]$
        with $\omega_S = \sum_{i \in S} \frac{n_i}{n_S} \omega_i^t$
        and $n_S = \sum_{i \in S} n_i$
    end
end

---

points, to the best of our knowledge, no prior research has specifically addressed the stability of contributions.

IV. METHODOLOGY

Having established the motivation behind our work, we present the proposed methodology applied to the experiments.

A. Experimental Setup

An open source repository containing an implementation of Shapley value calculations in FL, used in our experiments, is available at https://github.com/arnogeimer/flwr-contributions.

1) Training data: The experiments include image classification on CIFAR-10 and CIFAR-100 [24], MNIST [25] and Fashion-MNIST [26], four of the most common tasks in the Federated Learning literature.

Dataset	Clients	Train examples	Train steps
CIFAR-10	3	50,000	600
CIFAR-100	3	50,000	1000
MNIST	5	60,000	120
FMNIST	5	60,000	160

TABLE I
DATASET STATISTICS.

Following [27], we use a Dirichlet distribution-based size split with concentration parameter α to control the degree of non-IID data distribution. Specifically, we use $\alpha = 1$ for heavily non-IID data, $\alpha = 10$ for slightly non-IID data, and $\alpha = 100$ for an almost uniform distribution across all datasets. Using these values allows us to effectively study the effect of different levels of data heterogeneity.

2) Model and training pipeline: Since the focus of our work is not on model architecture, we deploy a straightforward convolutional neural network on all tasks, without hyperparameter fine-tuning. The model consists of two 5x5 convolution layers, 2 dense layers with ReLu activation and 16x5x5x120 and 120x84 units, respectively, and a dense layer with 84x10 units. Image preprocessing for CIFAR-10 and CIFAR-100 consists of random cropping, horizontal flipping, and random change of brightness, contrast, saturation, and hue.

These models are standard in the Federated Learning literature and their performance is adequate.

3) Strategies: We employ a collection of 8 different aggregation strategies, which have all been widely adopted across their respective application domains. Although aggregation strategies that do not have the same objective function as FEDAVG have been proposed [28], we do not include any of them in our analysis. Performance metrics of strategies, whether related to model accuracy or defense against attacks, are not the primary focus of this paper. We therefore employ out-of-the-box hyperparameters in all strategies.

4) Technical specifications: The models are built with Pytorch 1.13.1 [29], using Flower 1.6 [30] as the Federated Learning framework. All training is executed on Nvidia Tesla V100 16 GB VRAM GPUs.

As we used 4 different datasets and three Dirichlet splitting values, the experiments included 12 different use cases. With 70 seeds and 3 epoch values $e \in (2, 5, 10)$ per use case and a total of 8 aggregation strategies, our experiments encompass more than 20.000 full Federated Learning runs.

B. Contribution calculation

1) Round-based aggregation: We will call the contribution of the client k the normalized weighted sum of their per-round Shapley values ϕ_k^t. The total contribution of a participant is thus represented on a percentage scale, allowing better comparisons between runs. Although the numerical values of the Shapley values may vary, the proportions remain comparable. Shapley values are aggregated using an inverse linear factor for rounds, up to a maximum contribution halting round R. For example, at $R = 10$, the Shapley values are added using weight $\frac{10}{10}$ in round 1, $\frac{9}{10}$ in round 2, etc:

$$\phi_k(R) = \sum_{t=0}^{R} \frac{R-t}{R} \phi_k^t \qquad (2)$$

The final contribution of client k, as a percentage value, is $\phi_k \times (\sum_{i \in C} \phi_i)^{-1}$.

2) Ground truth: Determining which aggregation strategy produces the values closest to a ground truth is of the utmost importance. To this end, we establish the ground-truth contribution relative to the sizes of the participants' datasets. Since we do not operate in an adversarial environment, where actors may lie about their dataset statistics, and the distribution of classes across different splittings is nonheterogeneous, we deem the size of splits to be a reliable contribution assessment for clients. This method has previously been used in the literature [31] [32]. The correctness of a contribution is calculated using the squared Euclidean distance to the ground truth, in line with previous studies. To better understand our results, we establish a lower bound using an equal payout contribution evaluation approach.

Lemma 1. *Let a data set D be split by size into n different subsets $D_1, ..., D_n$ following a Dirichlet distribution* $Dir((\alpha, ..., \alpha))$. *Then, an equal payout differs from a size-based payout, on average, by $d = \frac{n-1}{n^2\alpha+n}$ under the squared Euclidean distance.*

Proof. Representing the size-based payout by a Dirichlet-random variable $X \sim Dir((\alpha, ..., \alpha))$, the expected squared Euclidean difference to an equal payout is

$$d = \mathbb{E}\left[\left\|\left(X - \left(\frac{1}{n}, ..., \frac{1}{n}\right)\right)\right\|_2^2\right]$$
$$= \mathbb{E}\left[\sum_{i=1}^{n}(X_i - \mu(X_i))^2\right]$$
$$= \sum_{i=1}^{n}\mathbb{E}\left[(X_i - \mu(X_i))^2\right]$$
$$= \sum_{i=1}^{n}\text{var}(X_i)$$
$$= n \times \frac{n-1}{n^3\alpha + n^2}$$

□

The lemma gives an upper bound on the performance of any contribution method, provided that size is the ground truth. Given any Dirichlet-based data split, if a contribution method's result does not, on average, lie within $\frac{n-1}{n^2\alpha+n}$ of the ground truth, it is more sensible to use an equal payout instead of that contribution method.

C. Optimizing the Contribution Halting Round R

Figure 1 contains the density plots of the optimal values for R over all experiments with the respective aggregation strategy. It shows that determining an optimal value for R is not straightforward: No strategy displays a clear, easy-to-analyze distribution. In all cases, an optimal value appears across all rounds, with a skew towards the beginning, or the end, of the Federated Learning process. We will use the mean optimal R of the respective strategy in our experiments. In applications, a relatively early stopping point for contribution evaluation proves beneficial, as it requires less computational expenses on the server side. However, this will be decided on a case-by-case basis.

V. CONTRIBUTION PERFORMANCE

Table II shows the squared Euclidean distance from the ground truth in all experiments. We note that most strategies outperform an equal payout when dataset sizes are heavily heterogeneous. The same can be observed in the slightly non-iid case. In the case of $\alpha = 100$, equal contributions beat all strategies as expected, as Dirichlet random values with such a high concentration parameter have extremely low variance. We can conclude that, for most practical purposes, employing an actual contribution calculation strongly surpasses an equal payout. These results demonstrate the validity of the implementation, and of round-wise Shapley values.

e	Strategy	CIFAR-10			CIFAR-100			MNIST			FMNIST		
	α	1	10	100	1	10	100	1	10	100	1	10	100
2	FedAvg	<u>3.71</u>	<u>0.86</u>	**0.77**	**0.62**	0.76	0.96	<u>3.42</u>	<u>1.3</u>	1.18	<u>3.63</u>	<u>1.12</u>	0.65
	FedAvgM	<u>3.66</u>	<u>0.9</u>	1.05	<u>0.7</u>	**0.73**	1.08	<u>3.53</u>	<u>1.11</u>	1.2	<u>3.59</u>	<u>1.03</u>	0.68
	FedAdagrad	8.21	<u>1.53</u>	0.93	<u>3.02</u>	<u>1.18</u>	**0.74**	<u>4.06</u>	<u>1.51</u>	1.27	<u>4.22</u>	<u>1.07</u>	0.85
	FedAdam	<u>3.5</u>	2.39	1.13	<u>1.04</u>	1.19	1.15	**2.43**	1.85	1.63	3.98	1.62	1.21
	FedYogi	**2.4**	1.34	0.82	1.09	1.55	1.37	4.58	<u>1.37</u>	**0.5**	5.77	1.51	0.64
	FedMedian	<u>3.3</u>	**0.79**	0.97	<u>0.69</u>	0.92	1.05	<u>2.74</u>	<u>1.33</u>	1.05	**3.49**	**0.95**	**0.63**
	FedTrimAvg	<u>3.48</u>	<u>0.83</u>	0.96	<u>0.74</u>	0.81	0.75	<u>3.08</u>	**<u>1.1</u>**	1.07	<u>3.56</u>	<u>1.01</u>	0.67
	Krum	22.71	7.14	2.3	41.23	7.76	2.79	<u>5.93</u>	1.85	1.23	<u>6.38</u>	2.35	1.05
5	FedAvg	<u>3.71</u>	<u>0.9</u>	1.15	<u>0.82</u>	**0.64**	0.67	<u>3.51</u>	1.65	1.78	<u>3.37</u>	<u>1.0</u>	0.87
	FedAvgM	<u>3.66</u>	<u>0.92</u>	1.2	<u>0.75</u>	<u>0.71</u>	0.88	<u>3.39</u>	1.62	1.57	**3.23**	<u>0.96</u>	1.09
	FedAdagrad	7.78	<u>1.56</u>	**0.9**	3.19	<u>0.93</u>	1.08	4.46	1.81	1.36	3.79	**0.88**	1.13
	FedAdam	4.16	3.55	2.56	2.04	2.11	2.49	4.22	3.11	3.11	3.92	2.38	1.78
	FedYogi	**1.98**	1.01	0.77	**0.67**	1.01	**0.5**	3.48	**<u>0.81</u>**	0.53	3.92	1.24	**0.63**
	FedMedian	<u>3.31</u>	**0.73**	0.92	<u>0.88</u>	0.66	0.94	**<u>2.96</u>**	<u>1.31</u>	1.75	<u>3.27</u>	0.99	1.05
	FedTrimAvg	<u>3.91</u>	<u>0.89</u>	0.92	<u>0.9</u>	<u>0.73</u>	0.9	<u>3.2</u>	1.69	1.78	<u>3.3</u>	<u>0.92</u>	0.97
	Krum	20.14	7.35	2.54	26.03	4.9	1.84	<u>6.09</u>	2.37	2.25	<u>6.19</u>	2.48	1.61
10	FedAvg	4.32	**0.85**	**1.17**	<u>1.12</u>	**0.56**	1.1	<u>3.63</u>	2.24	2.2	<u>3.14</u>	**1.09**	1.35
	FedAvgM	4.04	<u>0.86</u>	1.36	<u>1.23</u>	<u>0.63</u>	**0.8**	<u>3.43</u>	1.89	2.39	<u>3.31</u>	<u>1.12</u>	1.58
	FedAdagrad	7.82	<u>1.42</u>	1.53	<u>3.49</u>	<u>1.45</u>	0.95	<u>3.43</u>	**1.43**	**1.12**	<u>3.28</u>	<u>1.42</u>	1.48
	FedAdam	4.55	3.15	1.64	2.68	<u>1.12</u>	1.4	4.71	2.99	2.95	4.16	3.31	2.85
	FedYogi	**2.28**	1.54	1.27	<u>1.34</u>	0.8	1.34	4.63	3.14	1.87	<u>3.13</u>	1.83	1.51
	FedMedian	<u>3.16</u>	<u>1.17</u>	1.74	<u>1.0</u>	0.97	1.33	**<u>2.94</u>**	2.12	2.24	<u>3.28</u>	1.67	1.45
	FedTrimAvg	<u>4.1</u>	<u>0.93</u>	1.27	**0.88**	0.64	0.94	<u>3.11</u>	2.02	2.31	**3.04**	<u>1.11</u>	**1.19**
	Krum	<u>15.85</u>	7.07	4.04	17.63	4.57	2.11	<u>6.02</u>	3.8	3.67	<u>5.94</u>	3.36	2.37
	Expected equal error	16.67	2.15	0.22	16.67	2.15	0.22	13.33	1.57	0.16	13.33	1.57	0.16

TABLE II

AVERAGE SQUARE EUCLIDEAN DISTANCE FROM PER-STRATEGY CONTRIBUTIONS TO THE SIZE-BASED GROUND TRUTH, LOWER VALUES ARE BETTER. THE LAST ROW CONTAINS THE EXPECTED DIFFERENCE TO AN EQUAL PAYOUT, AS DETERMINED IN LEMMA 1. <u>UNDERLINED</u> ENTRIES BEAT THE EQUAL ALLOCATION, **BOLD** ENTRIES ARE THE BEST IN THE COLUMN. ALL VALUES WERE MULTIPLIED BY 100 FOR READABILITY.
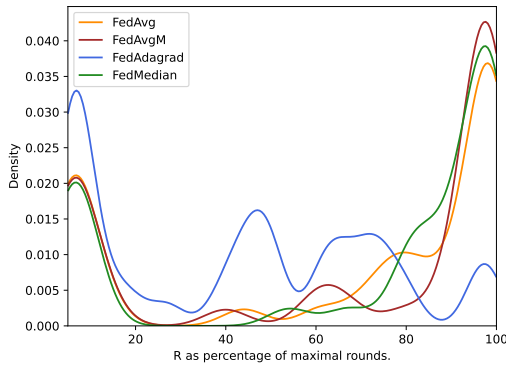

Fig. 1. Density plots of the best value for R, as percentage of total rounds, for select strategies. An optimal value minimizes the distance to the ground truth in a run.

However, we observe performance differences between the aggregation strategies. Although they are not often substantial, being aware that different aggregation methods may not yield the exact same performance in contribution estimation is important when establishing a federation. In particular, we note that no single aggregation strategy clearly outperforms all others. In fact, the best-performing strategy seems to change randomly between experiments. Although FEDYOGI surpasses the other strategies in eight out of 36 scenarios, this is not enough to draw the conclusion that it is a clearly superior aggregation method when it comes to per-round Shapley evaluation. The only conclusion which can be drawn is that KRUM is severely underperforming.

We conclude that all aggregation strategies result in satisfactory contribution evaluations. We suggest that, when proposing new contribution methods for Federated Learning, our results be taken into account by including multiple different aggregation techniques. We have shown that new aggregation mechanisms do not necessarily perform well with any contribution method, as is the case with KRUM in our results. This important observation shows that an out-of-the-box strategy, in combination with a contribution mechanism, may lead to poor results even though both the strategy and the contribution method perform adequately on their own.

Overall, the results in Table II suggest that using per-round Shapley values for contribution evaluation allows the central server to accurately approximate each client's contribution to the shared model and thus reward them fairly. This appears true, as all strategies result in contributions that lie relatively close to the ground truth.

However, Table III shows that a minimal distance from the ground truth does not necessarily mean that all clients are paid fairly. Indeed, we observe that, on average, the clients' rewards are in disagreement by more than 10% in

highly heterogeneous settings for even the best-performing strategies. We conclude that solely minimizing the Euclidean distance does not provide good information with regard to the performance of a contribution mechanism, as an overall decent result does not imply that each client is fairly compensated.

VI. CONTRIBUTION INSTABILITIES BETWEEN AGGREGATION STRATEGIES

Robustness, fairness, and generalization as parts of trustworthiness have become an important aspect of Artificial Intelligence [33]. This is no different in Federated Learning scenarios, where fairness of the global model and trust between participants are crucial to the operation of the federation.
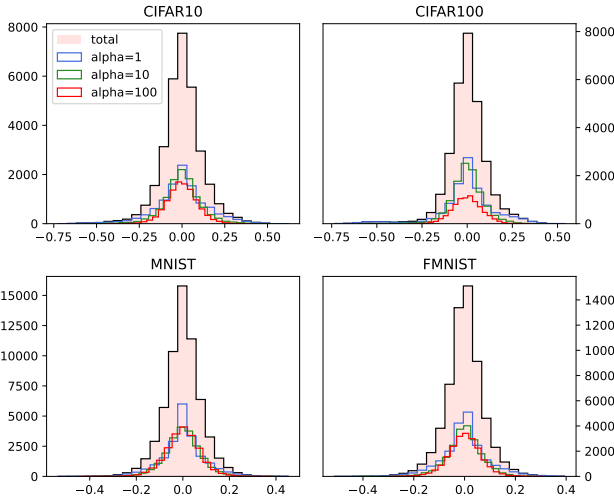

Fig. 2. The distributions of client-wise contribution differences between strategies across all experiments. A value of 0.1 represents a client receiving 10% more of the total reward using one strategy instead of another.

In light of this, the previous findings introduce the question of the stability of reward payouts, in general. We have shown that similarity to a ground truth does not imply that client-wise performance is adequate. This leads to the question of how stable the contributions are when there is no ground truth but only different strategies which share an objective function. Indeed, there is no guarantee that the Shapley values display similar distributions across different aggregation strategies, even if they perform equally well with respect to a common ground truth. Assuming a shared environment, two strategies should produce the same contribution values: If neither the data distribution, nor the model or the objective function change, but only the server-side aggregation of models, there is no reason for contribution to behave any differently. This is essential in an industrial setting: If clients know that diverse strategies produce the same global model, but contributions differ, agreeing on the choice of an aggregation strategy seems impossible. Therefore, we analyze and discuss the stability of contributions, not with respect to the ground truth, but among aggregation strategies.

Our experiments show that there are severe discrepancies in contribution allocation between strategies. Figure 2 shows the

distribution of the differences in the client-specific contribution in all experiments. We observe that clients may receive rewards that differ greatly depending on the chosen aggregation strategy. Keeping in mind that the upper row contains the results of CIFAR-10 and CIFAR-100, both using 3 clients, some examples contain discrepancies that reach 50% of the total reward. Naturally, the average difference is zero: One client's loss is another client's gain.

In addition, the figure contains α-specific histograms. We observe that there is no value that performs significantly better than the others; in fact, the error distribution is stable between different values. This illustrates that no matter how heterogeneous the data, the problem persists. This is undesirable in a nonmalicious federation: A client receiving almost half of the total reward more depending on the aggregation technique is not in the interest of anyone involved.
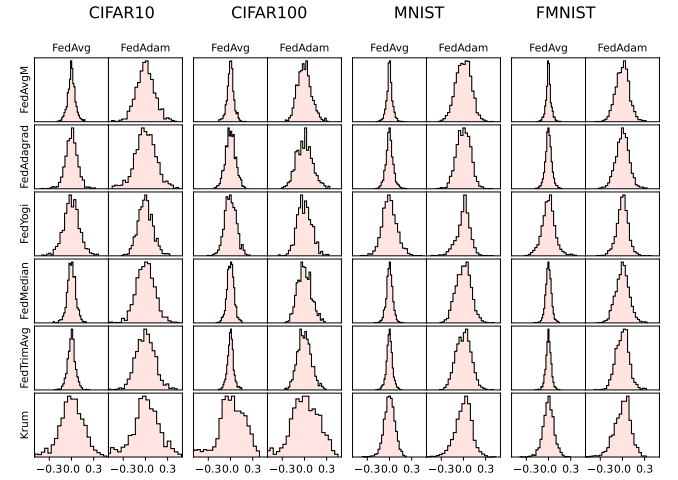

Fig. 3. Histograms show contribution differences between pairs of strategies. X-axes are fixed across all plots.

The scale of the problem is demonstrated in Figure 3, which shows the instability between different strategies. Ideally, all histograms would show no variance, with all differences close to zero. We show that this is not the case: Although the variance between strategies changes, there always exist substantially different contribution allocations. Although there are clearly some strategies that agree with others in most contributions, especially FEDAVG, we show that there are still discrepancies of up to 30% of the total reward when compared to other aggregations. In summary, we have shown that, even if the overall contribution lies close to the ground truth, different aggregation strategies will lead to differing contributions per client. As described above, this can cause substantial problems when establishing federations.

VII. DISCUSSION

To the best of our knowledge, this work is the first to conduct a large-scale empirical analysis with findings that highlight a critical issue in federated learning: the inherent instability in the estimation of the contribution based on the per-round Shapley value across different aggregation strategies.

e	Strategy (α)	CIFAR-10			CIFAR-100			MNIST			FMNIST		
		1	10	100	1	10	100	1	10	100	1	10	100
2	FedAvg	12.63	6.13	**5.99**	**5.31**	5.89	6.88	12.62	7.65	7.5	13.06	7.13	5.71
	FedAvgM	12.55	6.59	7.25	5.64	**5.65**	6.9	12.82	**7.31**	7.73	12.97	6.68	5.48
	FedAdagrad	20.72	8.61	6.44	12.32	7.29	**5.56**	13.84	8.69	7.66	14.37	6.9	6.29
	FedAdam	12.7	10.1	7.14	7.26	7.21	7.09	10.2	9.93	9.31	13.32	8.26	7.23
	FedYogi	**10.39**	8.21	6.49	7.22	9.17	8.18	15.22	8.15	**4.67**	16.53	8.09	5.57
	FedMedian	12.78	6.18	6.73	5.76	6.45	7.35	11.43	7.86	7.07	**12.71**	**6.62**	**5.26**
	FedTrimAvg	12.6	**6.03**	6.5	5.91	6.21	6.05	12.39	7.33	7.46	13.04	6.82	5.43
	Krum	35.77	19.28	10.24	51.03	18.99	11.68	16.23	9.71	7.41	17.73	10.35	7.07
5	FedAvg	13.1	6.49	7.38	6.09	**5.69**	5.8	13.17	8.69	9.44	12.73	6.99	**6.24**
	FedAvgM	12.85	6.37	7.09	5.84	5.84	6.79	12.74	8.7	8.56	**12.31**	**6.3**	7.11
	FedAdagrad	19.54	8.58	6.82	12.98	6.55	7.18	14.52	9.43	7.83	13.64	6.33	7.32
	FedAdam	14.27	13.52	11.65	9.55	9.82	11.27	13.96	12.99	13.37	13.6	10.48	9.05
	FedYogi	**9.7**	6.83	**6.2**	**5.58**	6.94	**4.69**	12.49	**6.04**	**5.18**	13.71	7.27	5.64
	FedMedian	12.47	**5.91**	6.88	6.35	5.83	6.82	**12.1**	7.67	9.18	**12.31**	6.67	7.23
	FedTrimAvg	13.32	6.56	6.71	6.32	5.95	6.58	12.97	8.68	9.36	12.37	6.5	6.78
	Krum	33.57	18.88	10.83	38.61	15.48	9.04	16.8	10.82	10.87	17.4	10.76	8.37
10	FedAvg	13.91	**6.29**	**7.48**	7.29	**5.18**	7.13	13.49	10.47	10.39	12.31	7.21	8.01
	FedAvgM	13.32	6.44	7.82	7.61	5.63	**6.28**	13.0	9.31	10.62	12.48	7.05	8.93
	FedAdagrad	19.28	8.15	8.22	13.31	8.32	6.84	12.67	**8.1**	**7.26**	12.48	8.01	8.31
	FedAdam	14.77	12.65	8.66	11.65	7.39	8.21	14.86	11.35	11.26	13.86	12.45	11.34
	FedYogi	**10.85**	8.64	8.09	7.74	6.54	8.38	15.41	12.58	9.18	12.26	9.63	8.99
	FedMedian	12.44	7.02	9.64	6.77	6.56	8.25	**12.36**	9.8	10.6	12.42	8.65	8.28
	FedTrimAvg	13.31	6.8	7.91	**6.45**	5.52	6.54	12.81	9.54	10.39	**12.11**	**6.77**	**7.5**
	Krum	29.6	18.43	13.73	30.62	15.27	10.45	16.98	13.64	12.93	17.2	12.57	10.61

TABLE III

AVERAGE CHEBYCHEV, OR L_∞, DISTANCE TO THE SIZE-BASED GROUND TRUTH, LOWER VALUES ARE BETTER. THE ENTRIES REPRESENT THE MEAN WORST CLIENT PERCENTAGE DIFFERENCE FROM THE GROUND TRUTH; **BOLD** ENTRIES ARE THE BEST IN THE COLUMN.

This instability raises questions about the viability of current methods for determining client contributions in real-world federated learning environments, particularly in a cross-silo setting, where economic incentives are more common. One of the main implications of these results is the potential erosion of trust in the contribution evaluation methods and therefore in the federation at large. Without assurances in place, a participant whose contribution turns out to be far less than expected could argue that the chosen strategy disadvantages them and leaves the federation. This could hurt all participants, as the leaver might have possessed unique data from which the global model, and thus everyone involved, would profit. Trust is a cornerstone of federated learning, and its erosion could significantly impede its adoption in commercial applications.

Hence, we argue that although most aggregation strategies achieve comparable performance in terms of model accuracy, contribution evaluation using round-based Shapley values in Federated Learning lacks robustness, generalization, and fairness. By shedding light on this instability of contribution evaluation between some of the most popular strategies, we hope to provide a foundation for future aggregation strategies to self-evaluate not only their performance under traditional metrics but also the contribution evaluation stability.

Therefore, when configuring federation parameters in a real-world setting, it is crucial to ensure that any transition to an alternative aggregation strategy is accompanied by a comprehensive stability study such as the one presented in this work. Through such a study, federations can ensure that the new aggregation strategy not only performs well in terms of model accuracy but also provides contribution stability, fostering a sustainable and trustworthy continuation of the federation.

In a more general context, we advocate for a reorientation in the design and evaluation of federated learning contribution systems. Beyond traditional metrics like accuracy, future research and development efforts should include the fairness and robustness of contributions in their studies to foster a sustainable and trustworthy federated learning ecosystem.

VIII. CONCLUSIONS & FUTURE WORK

This work has brought to light a concern regarding the stability of popular Federated Learning strategies when evaluating the contribution of participants. Our findings underscore the challenges and risks associated with deploying cross-silo federated learning platforms in industrial settings, given the discrepancy of contribution amongst these strategies. Future proposed FL strategies should take our findings into account, conducting a performative analysis not only on the similarity of the strategies' contributions compared to a ground truth, but also on their stability compared to other strategies.

In a deployed cross-silo federation, a major task is to reduce friction between participants: Unlike what was observed, where a specific aggregation strategy might have been more beneficial to a client, statistically relevant contribution values may be sampled ahead. This would leaves the option to choose the strategy which is most beneficial to the federation as a

whole, eliminating inter-client concurrence. This would lead to participants being rewarded more fairly, greatly improving trust in the remuneration process.

A potential venue of research would be to design an aggregation strategy based on an ensemble of different aggregation strategies designed to reduce the variance of the contribution allocation. This could mitigate the observed instability and maintain acceptable performance.

Finally, even if mitigation strategies are implemented for the observed instability, the computational complexity of Shapley value-based methods presents another hurdle in scaling these approaches to federations with a large number of participants. The computational effort of Shapley values scales exponentially with the number of clients, rendering simulations in high-client environments highly demanding. As federated learning gains traction in industrial and societal applications, such as healthcare and finance, the need for scalable, computationally efficient contribution evaluation techniques becomes paramount. Methods like FASTSHAP [34] or Monte Carlo sampling offer promising avenues for reducing the computational burden, but their impact on stability and fairness requires thorough investigation.

REFERENCES

[1] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. A. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. D'Oliveira, S. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konečný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, M. Raykova, H. Qi, D. Ramage, R. Raskar, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao, "Advances and open problems in federated learning," 2019. [Online]. Available: https://arxiv.org/abs/1912.04977
[2] H. R. Roth, Y. Cheng, Y. Wen, I. Yang, Z. Xu, Y.-T. Hsieh, K. Kersten, A. Harouni, C. Zhao, K. Lu et al., "Nvidia flare: Federated learning from simulation to real-world," arXiv preprint arXiv:2210.13291, 2022.
[3] M. Ye, X. Fang, B. Du, P. C. Yuen, and D. Tao, "Heterogeneous federated learning: State-of-the-art and research challenges," ACM Computing Surveys, vol. 56, no. 3, pp. 1–44, 2023.
[4] S. Li, Y. Cheng, W. Wang, Y. Liu, and T. Chen, "Learning to detect malicious clients for robust federated learning," arXiv preprint arXiv:2002.00211, 2020.
[5] S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," 2021.
[6] R. Zeng, C. Zeng, X. Wang, B. Li, and X. Chu, "A comprehensive survey of incentive mechanism for federated learning," 2021. [Online]. Available: https://arxiv.org/abs/2106.15406
[7] Z. Liu, Y. Chen, H. Yu, Y. Liu, and L. Cui, "Gtg-shapley: Efficient and accurate participant contribution evaluation in federated learning," ACM Transactions on intelligent Systems and Technology (TIST), vol. 13, no. 4, pp. 1–21, 2022.
[8] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," PMLR, pp. 1273–1282, 2017.
[9] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," 2019.
[10] D. Yin, Y. Chen, K. Ramchandran, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," 2021.
[11] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," Advances in neural information processing systems, vol. 30, 2017.

[12] I. Parrachino, A. Dinar, and F. Patrone, "Cooperative game theory and its application to natural, environmental, and water resource issues: 3. application to water resources," Application to Water Resources (November 2006). World Bank Policy Research Working Paper, no. 4074, 2006.
[13] L. S. Shapley, "Notes on the n-person game—ii: The value of an n-person game," 1951.
[14] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gürel, B. Li, C. Zhang, D. Song, and C. J. Spanos, "Towards efficient data valuation based on the shapley value," in The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, 2019, pp. 1167–1176.
[15] Y. Sarikaya and O. Ercetin, "Motivating workers in federated learning: A stackelberg game perspective," IEEE Networking Letters, vol. 2, no. 1, pp. 23–27, 2019.
[16] J. Zhang, C. Li, A. Robles-Kelly, and M. Kankanhalli, "Hierarchically fair federated learning," arXiv preprint arXiv:2004.10386, 2020.
[17] L. Lavaur, P.-M. Lechevalier, Y. Busnel, R. Ludinard, M.-O. Pahl, and G. Texier, "Radar: Model quality assessment for reputation-aware collaborative federated learning," in 2024 43rd International Symposium on Reliable Distributed Systems (SRDS), 2024, pp. 222–234.
[18] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in International conference on machine learning. PMLR, 2017, pp. 1885–1894.
[19] A. Richardson, A. Filos-Ratsikas, and B. Faltings, "Rewarding high-quality data via influence functions," arXiv preprint arXiv:1908.11598, 2019.
[20] X. Tu, K. Zhu, N. C. Luong, D. Niyato, Y. Zhang, and J. Li, "Incentive mechanisms for federated learning: From economic and game theoretic perspective," 2021.
[21] G. Wang, C. X. Dang, and Z. Zhou, "Measure contribution of participants in federated learning," in 2019 IEEE international conference on big data (Big Data). IEEE, 2019, pp. 2597–2604.
[22] T. Song, Y. Tong, and S. Wei, "Profit allocation for federated learning," in 2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019, pp. 2577–2586.
[23] T. Wang, J. Rausch, C. Zhang, R. Jia, and D. Song, "A principled approach to data valuation for federated learning," 2020.
[24] A. Krizhevsky, G. Hinton et al., "Learning multiple layers of features from tiny images," 2009.
[25] L. Deng, "The mnist database of handwritten digit images for machine learning research," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 141–142, 2012.
[26] H. Xiao, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," arXiv preprint arXiv:1708.07747, 2017.
[27] Q. Li, Y. Diao, Q. Chen, and B. He, "Federated learning on non-iid data silos: An experimental study," in 2022 IEEE 38th International Conference on Data Engineering (ICDE). IEEE, 2022, pp. 965–978.
[28] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair resource allocation in federated learning," arXiv preprint arXiv:1905.10497, 2019.
[29] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in Advances in Neural Information Processing Systems 32. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
[30] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, H. L. Kwing, T. Parcollet, P. P. d. Gusmão, and N. D. Lane, "Flower: A friendly federated learning research framework," arXiv preprint arXiv:2007.14390, 2020.
[31] P. Guo, Y. Yang, W. Guo, and Y. Shen, "A fair contribution measurement method for federated learning," Sensors, vol. 24, no. 15, p. 4967, 2024.
[32] S. K. Shyn, D. Kim, and K. Kim, "Empirical measurement of client contribution for federated learning with data size diversification," IEEE Access, vol. 10, pp. 118 563–118 574, 2022.
[33] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou, "Trustworthy ai: From principles to practices," 2022.
[34] N. Jethani, M. Sudarshan, I. Covert, S.-I. Lee, and R. Ranganath, "Fastshap: Real-time shapley value estimation," 2022.