Quality Inference in Federated Learning with Secure Aggregation

Balázs Pejó and Gergely Biczók

Abstract—Federated learning algorithms are developed both for efficiency reasons and to ensure the privacy and confidentiality of personal and business data, respectively. Despite no data being shared explicitly, recent studies showed that the mechanism could still leak sensitive information. Hence, secure aggregation is utilized in many real-world scenarios to prevent attribution to specific participants. In this paper, we focus on the quality (i.e., the ratio of correct labels) of individual training datasets and show that such quality information could be inferred and attributed to specific participants even when secure aggregation is applied. Specifically, through a series of image recognition experiments, we infer the relative quality ordering of participants. Moreover, we apply the inferred quality information to stabilize training performance, measure the individual contribution of participants, and detect misbehavior.

Index Terms—Quality Inference, Federated Learning, Secure Aggregation, Misbehavior Detection, Contribution Score

1 Introduction

For machine learning (ML) tasks, it is widely accepted that more training data leads to a more accurate model. Unfortunately, in reality, the data is scattered among multiple different entities. Thus, data holders could potentially increase the accuracy of their local model accuracy by training a joint model together with others [1]. Several collaborative learning approaches were proposed in the literature, amongst which the least privacy-friendly method is centralized learning, where a server pools the data from all participants together and trains the desired model. On the other end of the privacy spectrum, there are cryptographic techniques such as multi-party computation [2] and homomorphic encryption [3], guaranteeing that only the final model is revealed to legitimate collaborators and nothing more. Neither of these extremes admits most real-world use cases: while the first requires participants to share their datasets directly, the latter requires too much computational resource to be a practical solution for big data scenarios.

Somewhere between these (in terms of privacy protection) stands *federated learning* (FL), which mitigates the communication bottleneck and provides flexible participation by selecting a random subset of participants per round, who compute and send their model updates to the aggregator server [4]. FL provides some privacy protection by design as the actual data never leaves the hardware located within the participants' premises. Yet, there is already rich and growing related literature revealing that from these updates (i.e., gradients) a handful of characteristics can be inferred about the underlying training dataset. Potential attacks include model inversion [5], membership inference [6], reconstruction attack [7], (hyper)parameter inference [8], and property inference [9].

Parallel to these, several techniques have been developed to conceal the participants' updates from the aggregator server, such as differential privacy (DP) [10] and *secure aggregation* (SA) [11]. Although DP comes with a mathematical privacy guarantee, it also results in heavy utility loss, which limits its applicability in many real-world scenarios. On the other hand, *SA* does not affect the aggregated final model, which makes it a suitable candidate for many applications. Essentially, *SA* hides the individual model updates without changing the aggregated model by adding pairwise masks to the participants' gradients in a clever way so that they cancel out during aggregation.

Consequently, *SA* only protects the participants' individual updates and leaves the aggregated model unprotected. Hence, *SA* provides a "hiding in the crowd" type of protection [12], thus, without specific background knowledge, it is unlikely that a privacy attacker could link the leaked information to a specific participant. The lack of attribution severely affects the security of FL as well; we are not aware of any attack detection scheme applicable with *SA* enabled.

In this paper, we study the possibility of *inferring the quality of the individual datasets when SA is in place*. This could be utilized for for attack detection as well. Note, however, that it is different from mere poisoning and backdoor detection [13], as that line of research is only interested in classifying participants as malicious or benign, while our goal is to enable the fine-grained differentiation of FL participants with respect to their data quality. This is fundamentally similar to contribution score computation, which is also an unsolved problem in the *SA* setting.

We are aware that data quality is a complex concept with multiple dimensions [14], and in general it is relative from two aspects: it can only be considered in terms of the proposed use and in relation to other data samples. For this reason (similarly to [15]) we focus on image recognition

Project no. 138903 has been implemented with the support provided by the Ministry of Innovation and Technology from the NRDI Fund, financed under the FK_21 funding scheme. The research was supported by the Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence National Laboratory Program.

CrySyS Lab, Department of Networked Systems and Services Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, and ELKH-BME Information Systems Research Group, Hungary {pejo,biczok}@crysys.hu

tasks with noisy labels, as in this scenario data quality has a straightforward interpretation.

Contributions

We propose a method called *Quality Inference* (QI) which (by utilizing the improvement of the aggregated updates) recovers the relative label quality of the contributing participants' datasets. To obtain this quality information, our method takes advantage of the improvements of the aggregated models across multiple rounds, as well as the known per-round selected subset of participants. QI works by evaluating the aggregated updates in each round and assigning scores to the selected participants based on three simple but novel rules called *The Good*, *The Bad*, and *The Ugly* (as in the movie [16]). As a result, we are able to recover the relative quality ordering (i.e., by label correctness rate) of the participants.

We simulated datasets with different qualities by utilizing unique label-flipping rates for each participant, and conduct experiments on two neural network architectures (MLP and CNN) and two datasets (MNIST and CIFAR10). We consider three FL settings, where 2 out of 5, 5 out of 25, and 10 out of 100 participants are selected in each round to update the model, respectively.

Our experiments show that the three proposed heuristic scoring rules significantly outperform the baseline in determining the participants' data qualities relative to each other (i.e., correct label rates). We find that the accuracy of QI depends on both the complexity of the task and the trained model architecture. We also conduct an ablation study on the hyperparameters of the proposed rules.

Finally, we investigate three potential applications of QI: on-the-fly performance boosting, contribution score computation, and misbehavior detection (by considering free-riding and poisoning). We find that i) carefully weighting the participants based on the inferred scores smooths the learning curve, ii) the scores could be used as a measure of participant contribution, and iii) the scores are able to reveal misbehaving participants. This latter implies that besides the label correctness rate, QI is also capable of inferring other, more general quality aspects of the data. We are not aware of any work tackling any of the aforementioned issues when *SA* is enabled.

2 The Theoretical Model

In this section we introduce the theoretical model of quality inference and highlight its complexity. We note with n a participant in FL, while N denotes the number of all participants. Similarly, i denotes a round in FL, while I denotes the number of all rounds. The set S_i contains the randomly selected participants for round i, and $b = |S_i|$ captures the number of selected participants. D_n is participant n's dataset consisting of $(x, y) \in D_n$ data-label pairs. We assume D_n is associated with a single scalar u_n, which measures its quality. We use θ_n and v_i to capture the quality of the nth participant's gradient and the quality of the aggregated gradient in the ith round, respectively. A summary of the variables are listed in the Appendix (Table 3).

2.1 Deterministic Case

In this simplified scenario we assume the gradient quality is equal to the dataset quality, i.e., $\theta_n = u_n$. Consequently, the aggregated gradients represent the average quality of the participants' datasets. As a result, the round-wise quality values of aggregated gradients form a linear equation system $Au = v$, where $u = [u_1, \ldots, u_N]^T$, $v = [v_1, \ldots, v_I]^T$, and $a_{i,n} \in A_{I \times N}$ indicates whether participant n is selected for round i. Depending on the dimensions of A, the system can be under- or over-determined. In case of $I < N$ (i.e., no exact solution exists) and if $I > N$ (i.e., many exact solutions exist), the problem itself and the approximate solution are shown in Eq. 1 and 2, respectively.

$$\min_u ||v - Au||_2^2 \quad \Rightarrow \quad u = (A^T A)^{-1} A^T v \tag{1}$$

$$\min_u ||u||_2^2 \text{ s.t. } Au = v \quad \Rightarrow \quad u = A^T (AA^T)^{-1} v \tag{2}$$

2.2 Stochastic Case

The above equations do not take into account any randomness. Given that the training is stochastic, we can treat the quality of participant n's gradient as a random variable θ_n sampled from a distribution with parameter u_n. Moreover, we can represent $\theta_n = u_n + e_n$ where e_n corresponds to a random variable sampled from a distribution with zero mean. We can further assume that e_n and $e_{n'}$ are i.i.d. for $n \neq n'$. As a result, we can express the aggregated gradient $v_i = \sum_n a_{i,n} u_n + E$ where E is sampled from the convolution of the probability density function of e's.

In this case, due to the Gauss–Markov theorem [17], the solution in Eq. 1 is the best linear unbiased estimator, with error $||v - Au||_2^2 = v^T(\mathbf{I} - A(A^T A)^{-1} A^T)v$ (where $\mathbf{I}$ is the identity matrix) with an expected value of $b(\mathbf{I} - N)$. Note, that with more iterations more information is leaking, which should decrease the error. Yet, this is not captured by the theorem as it considers every round as a new constraint.

This problem lies within estimation theory [18], from which we already know that estimating a single random variable with added noise is already hard; moreso, factoring in that in our setting, we have multiple variables forming an equation system. Moreover, these random variables are different per round; a detail we have omitted thus far. Nevertheless, each iteration corresponds to a different expected accuracy improvement level, as with time the iterations improve less-and-less. Consequently, to estimate individual dataset quality we have to know the baseline expected learning curve; in turn, the learning curve depends exactly on those quality values. Being a chicken-egg problem, we focus on empirical observations to break this vicious cycle.

3 Quality Scoring

In this section we devise the three intuitive scoring rules which are the core of QI: they either reward or punish the participants in the FL rounds. The notations used in this section are summarized in the Appendix (Table 4). We define ω_i as the aggregated model's improvement in the ith round and $\varphi_{i,n}$ as the quality score of participant n after round i. Note that in the rest of the paper we slightly abuse the notation by removing index i where it is not relevant.

3.1 Assumptions

We assume a honest-but-curious setting; the aggregator server (and the participants) cannot deviate from the FL protocol. Further restrictions on the attacker include limited computational power and no background knowledge

besides access to an evaluation oracle. For this reason, we neither utilize any contribution score based techniques nor existing inference attacks, as these require either significant computational resources or user-specific relevant background information.

3.2 Scoring Rules

Based on the round-wise improvements ω_i, we created three simple rules to reward or punish the participants. We named them *The Good*, *The Bad*, and *The Ugly* (as in the spaghetti western movie [16]); the first one (G) rewards the participants in the more useful aggregates, the second one (B) punishes in the less useful ones, while the last one (U) punishes when the aggregate does not improve the model at all. Formally, each participant n' contributing in round i that ...

- G ... improves the model more than the previous round (i.e., $\omega_i > \omega_{i-1}$) receives $+1$, i.e., $\varphi_{i,n'} \leftarrow \varphi_{i-1,n'} + 1$.
- B ... improves the model less than the following round (i.e., $\omega_i < \omega_{i+1}$) receives -1, i.e., $\varphi_{i+1,n'} \leftarrow \varphi_{i,n'} - 1$.
- U ... does not improve the model at all (i.e., $\omega_i < 0$) receives -1, i.e., $\varphi_{i,n'} \leftarrow \varphi_{i-1,n'} - 1$.

Note, that the quality score in round i is not updated for participant $\hat{n}$ who has not contributed in that round, i.e., $\varphi_{i,\hat{n}} \leftarrow \varphi_{i-1,\hat{n}}$.

It is reasonable to expect that the improvements in consecutive rounds are decreasing (i.e., $\omega_i < \omega_{i-1}$): first the model improves rapidly, while improvement slows down considerably in later rounds. The first two scoring rules (*The Good* and *The Bad*) capture the deviation from this pattern: we can postulate that i) high dataset quality increases the improvement more than in the previous round, and ii) low dataset quality decreases the improvement, which would be compensated in the following round. These phenomena were also shown in [19]. While these rules are relative, the last one (*The Ugly*) is absolute: it builds on the premise that if a particular round does not improve the model, there is a higher chance that some of the corresponding participants have supplied low quality data.

Independently of the participants' dataset qualities, round-wise improvements could deviate from this pattern owing to the stochastic nature of learning. We postulate that this affects all participants evenly, independently of their dataset quality; thus, the relation/ordering among the individual scores are not significantly affected by this "noise". Participant selection also introduces a similar effect; however, we assume that participants are selected uniformly, hence, its effect should also be similar across participants.

3.3 Quantifying QI

The quality scores of the participants are unlikely to converge; hence, we focus on their relation. We denote with $q_{i,n}$ the inferred quality-wise rank of participant n after round i, and we measure the accuracy of the inferred qualities by comparing $q_{i,n}$ for each participant to the baseline quality-wise ordering. For this purpose, we use the Spearman correlation coefficient r_s [20], which is based on the Spearman distance d_s [21] (as seen in Eq. 3). Spearman distance measures the absolute difference of this inferred and the actual position, while the Spearman correlation coefficient assesses monotonic relationships on the scale $[-1, 1]$; 1 corresponds to perfect correlation, while any positive value signals positive correlation between the actual and the inferred quality ordering. E.g., if the inferred quality order (via the three rules) expressed with participant IDs is 5-3-2-4-1, while the actual quality order is 5-4-3-2-1, then the Spearman distances are 0-2-1-1-0, and the Spearman correlation is 0.7, suggesting that the inferred quality order is very close to the original one. Note, that the Spearman distance (and consequently the coefficient) handles any misalignment equally, irrespective of the position.

$$d_s(i,n) = |n - q_{i,n}| \quad r_s(i) = 1 - \frac{6 \cdot \sum_{n=1}^{N} d_s(i,n)^2}{N \cdot (N^2 - 1)} \tag{3}$$

4 Experiments for QI

In this section, we describe our experiments, including quality simulation and the utilized datasets and model architectures, and present the corresponding results.

4.1 Simulating Data Quality

Data quality can only be considered in terms of the proposed use and in relation to other data samples, i.e., participants with different data distributions could have different views of the same dataset. To tackle this issue, we consider only the IID case in our experiments. Besides, data quality entails multiple aspects such as accuracy, completeness, redundancy, readability, accessibility, consistency, usefulness, and trust, with several having their own subcategories [14]. In this paper, we focus on image recognition tasks as it is a key ML task with standard datasets available. Still, we have to consider several of these aspects in relation to image data.

Unfortunately, we are not aware of any public datasets encompassing data from several well-categorized quality classes. Since visual perception is a complex process, to avoid serious pitfalls, we do not manipulate the images themselves, but simulate different qualities similarly to [15]: we modify the label y corresponding to a specific image x. To have a clear quality-wise ordering between the datasets (i.e., the ground truth), we perturbed the labels of the participants according to Eq. 4, where ψ_k is drawn uniformly at random over all available labels. Putting it differently, the labels of the participants' datasets are randomized before training with a linearly decreasing probability, e.g., in case of five participants with IDs [1,2,3,4,5], the ratio of assigned random labels are 100%, 75%, 50%, 25%, and 0%, respectively.

$$\Pr(y_k = \psi_k | (x_k, y_k) \in D_n) = \frac{N - n}{N - 1} \tag{4}$$

4.2 Datasets, ML Models and Experiment Setup

For our experiments, we used the MNIST [22] and the CIFAR10 [23] datasets. MNIST corresponds to the simple task of digit recognition. It contains $70,000$ hand-written digits in the form of 28×28 gray-scale images. CIFAR10 is more involved, as it consists of $60,000$ 32×32 color images of various objects. For MLP, we used a three-layered structure with hidden layer size 64, while for CNN, we used two convolutional layers with 10 and 20 kernels of

size 5×5, followed by two fully-connected hidden layers of sizes 120 and 84. For the optimizer, we used SGD with learning rate 0.01 and dropout rate 0.5. The combination of the two datasets and the two neural network models yield four use-cases. In the rest of the paper, we will refer to these as MM for MLP-MNIST, MC for MLP-CIFAR10, CM for CNN-MNIST, and CC for CNN-CIFAR10.

We ran all the experiments for 100 rounds and with three different FL settings, corresponding to 5, 25, and 100 participants where 2, 5, and 10 of them are selected in each round, respectively. The three FL settings combined with the four use-cases result in twelve evaluation scenarios. We ran every experiment 10-fold, with randomly selected participants.

4.3 Empirical Quality Scores

We present the pseudo-code of the whole process in Algorithm 1. We split the dataset randomly into $N+1$ parts (line 1), representing the N datasets of the participants and the test set D_{N+1}, to determine the quality of the aggregated updates. As highlighted earlier, the splitting is done in a way that the resulting sub-datasets are IID; otherwise, the splitting itself would introduce some quality difference between the participants.

Concerning D_{N+1}, having access to a dataset is standard practice both in the field of privacy attacks and contribution score computation, and our work is in the intersection of these. Shadow datasets is a widespread technique to mimic the training dataset, and having access to an evaluation oracle (via an IID test set) is a fundamental assumption for contribution score computation methods. Although we foresee multiple options how D_{N+1} could be obtained, this is orthogonal to our main contribution; we leave it as a future work.

Next, we artificially create the baseline dataset qualities using Eq. 4 (line 3): each participant's labels are randomized with a different ratio. This is followed by FL (line 5-9). Round-wise improvements are captured by ω (declared in line 11 using the accuracy difference of the current and previous models). Quality scores $(\varphi_1, \ldots, \varphi_N)$ are updated in the ith round with ± 1 each time one of the three scoring rules is invoked (line 12, 13, and 15 for *The Good*, *The Bad*, and *The Ugly*, respectively).

Algorithm 1 QI in FL with SA

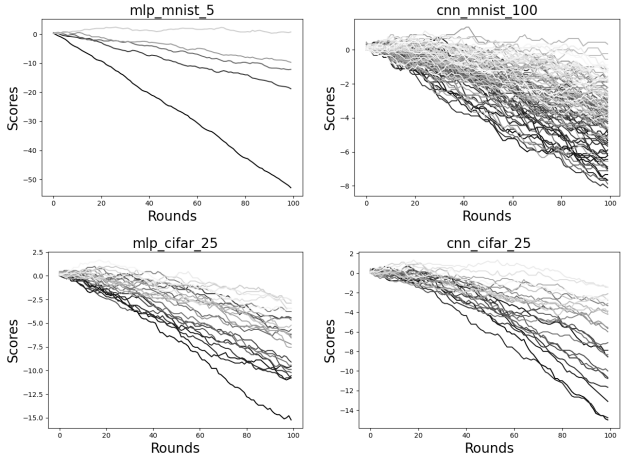
Input: data D; participants N; rounds I

1: $\text{Split}(D, N) \rightarrow \{D_1, \ldots, D_N, D_{N+1}\}$
2: **for** $n \in [1, \ldots, N]$ **do**
3: $\quad \forall\ (x_k, y_k) \in D_n : y_k \sim$ Eq. 4
4: $\varphi = [0, \ldots, 0]$; $M_0 \leftarrow \text{Rand}()$
5: **for** $i \in [1, \ldots, I]$ **do**
6: $\quad \text{RandSelect}([1, \ldots, N], b) \rightarrow S_i$
7: $\quad$ **for** $n \in S_i$ **do**
8: $\quad\quad \text{Train}(M_{i-1}, D_n) = M_i^{(n)}$
9: $\quad M_i = \frac{1}{b} \sum_{n \in S_i} M_i^{(n)}$
10: $\quad \omega_i = \text{Acc}(M_i, D_{N+1}) - \text{Acc}(M_{i-1}, D_{N+1})$
11: $\quad$ **if** $i > 1$ and $\omega_i > \omega_{i-1}$ **then**
12: $\quad\quad$ **for** $n \in S_i$ **do** $\varphi_n \leftarrow \varphi_n + 1$
13: $\quad\quad$ **for** $n \in S_{i-1}$ **do** $\varphi_n \leftarrow \varphi_n - 1$
14: $\quad$ **if** $\omega_i < 0$ **then**
15: $\quad\quad$ **for** $n \in S_i$ **do** $\varphi_n \leftarrow \varphi_n - 1$

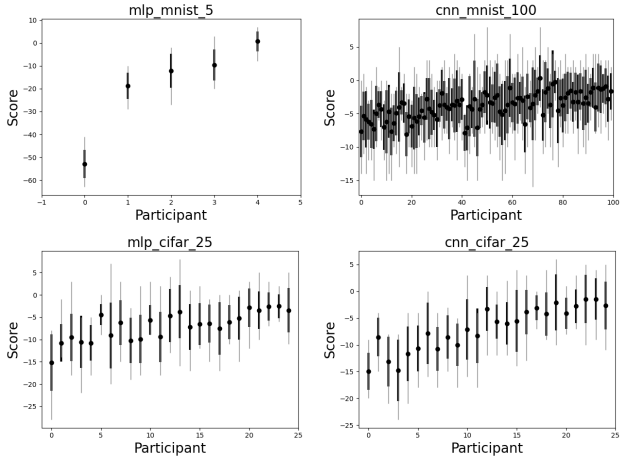
QI Results

The quality scores based on the three scoring rules for a handful of selected scenarios are presented in Fig. 1; the rest of the studied cases are shown in the Appendix (Fig. 6 and Fig. 7). In Fig. 1a we visualize the round-wise evolution of scores for each participant where the corresponding gray-ness level depends on the participant ID. More precisely, the lighter shades correspond to participants with higher IDs (i.e., less noisy labels according to Eq. 4), while the darker shades mark low ID participants (i.e., higher ratio of random labels). It is visible that the more rounds have passed, the better our scoring rules correctly differentiate the participants.

In Fig. 1b we show the mean (dot), the variance (black line), the minimum, and maximum values (gray line) of the inferred quality scores for each participant. One can see an increasing trend of the quality scores following the participant IDs. This is in line with the ground truth based



(a) The average round-wise change of the participants' scores. The lighter the better (the darker the worse) corresponding dataset quality.



(b) Score of the participants after 100 training rounds. IDs shown on x axis where lower number correspond to lower dataset quality.

Fig. 1: Quality scores of the participants. Left - MLP, right - CNN, top - MNIST with 5 and 100 participants, bottom - CIFAR with 25 participants.

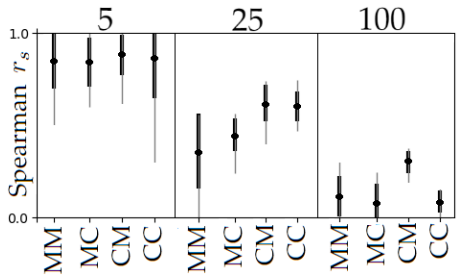


Fig. 2: Spearman coefficient for the 12 scenarios.

on Eq. 4. Note, that even for the participant with the perfect label quality (i.e., the highest ID or the lightest curve), the quality score is rather negative, and keeps decreasing with more rounds. This is an expected characteristic of the scoring rules: there is only one rule increasing the score (*The Good*), while two decreasing it (*The Bad* and *The Ugly*). Applied jointly, these three heuristic scoring rules approximate the ground truth label quality ordering remarkably well *exclusively from the aggregates*.

Finally, we utilize the Spearman coefficient r_s introduced in Eq. 3 to measure the accuracy of the inferred qualities; the 12 studied scenarios are presented in Fig. 2. Note, that $r_s \in [-1, 1]$, and any positive value indicates correlation. Thus, the value of the baseline (i.e., randomly guessed ordering) is zero. Consequently, the three simple rules significantly improve on the baseline, as the coefficients for all scenarios are positive. Moreover, as suggested by 1a, this value keep increasing with more rounds, as shown in the Appendix (Fig. 8).

Fine-tuning

We consider four ways of improving the accuracy of QI.

- *Rule combination*: we apply all possible combinations of scoring rules in order to remove redundancies and to find which setup obtains the highest accuracy.
- *Thresholding*: we consider using a threshold for the scoring rules, i.e., *The Ugly* only applies when the improvement is below some value, while *The Good*/*The Bad* applies if the improvement difference is above/below such a threshold, respectively.
- *Actual values*: we consider using improvement differences instead of ± 1 to account for a more precise differentiation.
- *Round skipping*: In the early rounds the model does improve almost independently of the dataset qualities, therefore, we consider discarding the information from the first few rounds to decrease noise.

Although we performed an exhaustive grid search (e.g., $\{0, 2^0, \ldots, 2^8\}/100$ for thresholding and $[0, 1, \ldots, 10]$ for round skipping), the overall improvements obtained were minor. The corresponding results are presented in the Appendix (Fig. 4). This implies that the original rules are quite efficient, and the heuristic thumbs-up/thumbs-down rules (e.g., using ± 1 to update the scores) could be interpreted as a normalizer across the different improvement levels of the rounds. Therefore, in the following applications, we use the original rules without any fine-tuning.

4.4 Mitigation

Note that the demonstrated quality information leakage is not by design; this is a bug, rather than a feature in FL. The simplest and most straightforward way to mitigate this vulnerability is to use a protocol where every participant contributes in each round (incurring a sizable communication overhead). Another approach is to hide the participants' IDs (e.g., via mixnets [24]), so no-one knows which participant contributed in which round except for the participants themselves. Finally, the aggregation itself could be done in a differentially private manner as well, where a carefully calculated noise is added to the updates in each round. Client-level DP [25] would by default hide the dataset quality of the participants, although at the price of requiring large volumes of noise, and therefore, having low utility.

5 Application of QI

Even though QI is not a mechanism purposefully engineered into FL (with *SA*), it does enable beneficial applications such as training accuracy stabilization, contribution score computation, and misbehaviour detection. Our results are shown in Fig. 3. Note that while there are a handful of existing mechanisms for these tasks within FL, they do not work under *SA*; hence, we do not compare our results quantitatively to the SotA methods.

5.1 Enhancing the Training

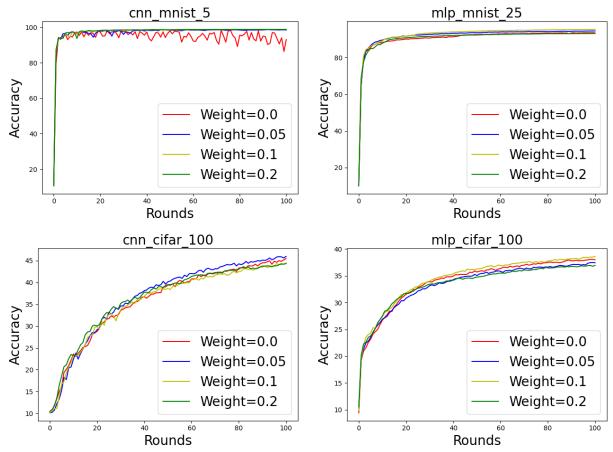
It is expected that both training speed and obtained accuracy could be improved by weighting the participants according to their data qualities. Hence, a potential use case for QI is to adopt the inferred scores as weights during training. For weighting we used the multiplicative weight update approach [26], which multiplies the weights with a fixed rate κ, i.e., each time during training one of the three scoring rules is invoked in Algorithm 1, the weights (initialized as $[1, \ldots, 1]$) are updated in the ith round with $\times(1 \pm \kappa)$ for the appropriate participants.

Note that without access to individual gradients (owing to *SA*), only the aggregates can be scaled by the server. Consequently, in each round only the aggregate is scaled with the arithmetic mean of the selected participants' weights. For our experiments, we set $\kappa = \{0.00, 0.05, 0.10, 0.20\}$, where the first value corresponds to the baseline without participant weighting. We highlight some of our results in Fig. 3a; the rest can be found in the Appendix (Fig. 9). It is conclusive that using weights based on our scoring rules enhances the training as the training curves are smoother and the final accuracies are higher.

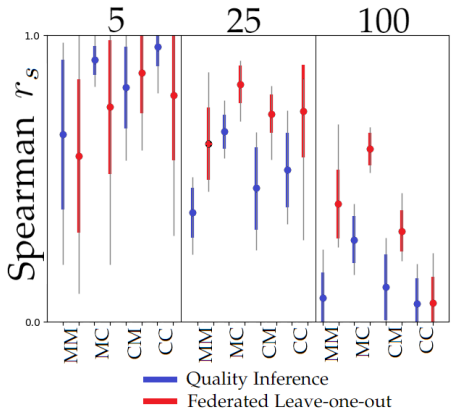
5.2 Contribution Score Computation

The second use case we envisioned for QI is contribution score computation. The holy grail of this sub-discipline is the Shapley value [27], which is exponentially hard to compute, as besides the individual information, it requires information about all potential coalitions of participants. Thus, many approximation methods exist (e.g., [28], [15]. Yet, all methods assume explicit access to the individual datasets or the corresponding gradients, which is not possible with *SA*. Consequently, there exists no contribution scoring mechanism which could be considered as a relevant baseline for QI.

This article has been accepted for publication in IEEE Transactions on Big Data. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2023.3280406



(a) The round-wise accuracy of the trained models with various weights. Left - CNN, right - MLP, top - MNIST with 5 and 25 participants, bottom - CIFAR with 100 participants.



(b) Spearman coefficient of QI and the leave-one-out method for the 12 scenarios.

Fig. 3: QI application scenarios.

According to [29], payment distribution based on the Shapley value is optimal for our IID setting. Moreover, the federated leave-one-out method (LO) method approximates the Federated Shapley value well in this case [15]. Although LO does need individual information (hence, not applicable with *SA*), we compare our method to it, as it only utilizes each individual gradient once (to obtain the grand coalition minus that participant).

The Spearman coefficients of the ordering based on QI and LO are presented in Figure 3b. As expected, LO is superior to QI, as it operates on individual information, which is by-design avoided by QI. What is somewhat surprising is that LO (benefiting from individual gradients) also struggles with reconstructing the quality-wise ordering perfectly. This suggest that separating participants with different label qualities is indeed a challenging task; given the restricted information setting, QI performs reasonably well.

5.3 Misbehaviour Detection

Another potential application of QI is misbehavior detection. It is a notoriously hard task even without *SA* [30]. At the time of writing we are not aware of any work tackling this problem in the *SA* setting.

Here we consider both malicious attackers and free-riders. Their goal is either to decrease the accuracy of the aggregated model, or to benefit from the aggregated model without contributing, respectively. We do not scramble the labels of honest participants, and simulate attackers by computing the additive inverse of the correct gradients, while we use zero as the gradient for free-riders. These are naive but stealthy strategies owing to *SA*. With this use case, our goal is not to propose a defense against SotA attackers, but rather to demonstrate the usability of QI besides label quality inference. Note that QI also shows promise for being applicable to determine other quality disparities among participants.

Setup	Attacker	MM	MC	CM	CC
5/2/1	p-value	2.0e-20	3.5e-37	8.3e-58	5.4e-84
T-Test	Stat	16.4	17.7	21.9	27.0
25/5/2	p-value	6.8e-06	1.9e-10	1.1e-14	6.0e-27
χ^2-Test	Stat	38.3	62.0	83.3	146.0
100/10/5	p-value	8.6e-03	1.2e-03	1.7e-07	1.1e-08
KS-Test	Stat	0.25	0.17	0.19	0.18

Setup	Free-Rider	MM	MC	CM	CC
5/2/1	p-value	3.7e-21	1.6e-42	4.3e-69	5.1e-98
T-Test	Stat	12.3	18.1	24.0	29.5
25/5/2	p-value	7.7e-03	7.5e-12	7.0e-17	6.1e-39
χ^2-Test	Stat	20.8	69.1	96.8	203.5
100/10/5	p-value	9.0e-02	2.0e-05	3.6e-07	4.8e-14
KS-Test	Stat	0.13	0.18	0.17	0.21

TABLE 1: Statistics and p-values of the selected scenarios: X/Y/Z mean number of participants, number of round-wise selected participants, and number of cheaters, respectively. It is clear, that the hypothesis "the scores of honest and cheating participants are similar" (e.g., having the same mean for the Student T-Test, having the same frequencies for the χ^2-Test, and coming from the same distribution for the Kolmogorov-Smirnov Test) is rejected with high confidence.

We studied the score of the honest and malicious participants; the average values for the selected scenarios are presented in Table 1; the rest can be found in the Appendix (Table 2). We also run various statistical tests to determine whether there is any difference between the honest and malicious participant's scores. Table 1 contains highlighted results, while the rest is presented in the Appendix (Table 5, 6, 7, 8, and 9). The tests concluded unanimously that the two score distributions are different, thus, QI is capable of correctly flagging dishonest participants. Besides the score differences we also studied the inferred position of a single cheater, which is always in the bottom half (see Fig. 5 in the Appendix).

6 Related Work

In this section, we briefly present related research efforts, including but not limited to data quality scoring mechanisms and well-known privacy attacks against machine learning. The theoretical analysis of QI does relate to [31] as attempting to reconstruct the dataset quality order is similar to reconstructing the entire dataset based on query outputs.

6.1 Participant Scoring

Simple but effective scoring rules are prevalent in complex ICT-based systems, especially characterizing quality. For instance, binary or counting signals can be utilized to i) steer peer-to-peer systems measuring the trustworthiness of peers [32], ii) assess and promote content in social media [33], iii) ensure the proper natural selection of products in

online marketplaces [34], and iv) select trustworthy clients via simple credit scoring mechanisms [35].

There exist free-rider detection mechanisms for collaborative learning [36], [37]. In contrast, [38] proposes an online evaluation method that defines each participant's impact based on the current and the previous rounds. Although their goal is similar to ours, we consider *SA* being utilized, while neither of the above mechanisms is applicable in such a case. A disaggregation technique is presented in [39], which reconstructs the participation matrix by simulating the same round several times with different participants. Instead, we assume such participation information to be available, and emulate the training rounds by properly updating the model.

Accuracy boosting by participant weighting is considered in [40] where the weights are determined by the underlying data quality calculated via the cross-entropy of the local model predictions. These experiments consider only five participants and two quality classes (fully correct or incorrect); we study fine-grained quality levels with larger sets of participants. A similar method was utilized in an *SA* setting in [41] using homomorphic encryption. In contrast, our method does not require any cryptographic primitive and can be utilized on top of any federated learning protocol.

We naively assume that data quality is directly related to the noise present in the labels. Naturally, this is a simplification: there is an entire computer science discipline devoted to data quality [14].

Authors of [42] listed several incentive mechanisms for contribution computation in FL (which can be interpreted as data quality). A pertinent notion is the Shapley value [27], which was designed to allocate goods to players proportionally to their contributions. A high-level summary of the role of the Shapley value within ML is presented in [43]. The main drawback of the Shapley value is its exponential computational requirement, which makes it unfeasible in most scenarios. Several approximation methods were proposed in the literature using sampling [44], gradients [28] and influence functions [45]. Although some are promising (e.g., the conceptual idea in [46]), all previous methods assume explicit access to either the datasets or the corresponding gradients. Consequently, these methods are not applicable when *SA* is enabled during FL. QI can be considered as the first step towards a contribution score when *no information on individual datasets is available*.

6.2 Privacy Attacks

There are several indirect threats against FL models. These could be categorized into model inference [5], membership inference [6], parameter inference [8], and property inference [9]. QI could be considered as an instance of the last. Source inference [47] is also such an attack, which could tie the extracted information to specific participants of FL. However, it does not work with *SA*. Another property inference attack is the quantity composition attack [48], which aims at inferring the proportion of training labels among the participants in FL. This attack is successful even under *SA* protocols or DP. In contrast to our work, the paper focuses on inferring the distributions of the non-IID datasets while we aim to recover the relative quality information on IID datasets. Finally, [49] also attempts to explore user-level privacy leakage within FL. Similarly to our work, the attack defines client-dependent properties, which then can be used to distinguish the clients from one another. The authors assume an active malicious server utilizing a computationally heavy GAN for the attack, which is the exact opposite of our honest-but-curious setup with limited computational power.

6.3 Privacy Defenses

QI can be considered as a property inference attack; hence, naturally, it can be "mitigated" via client-level DP [25]. Moreover, as we simulate different dataset qualities with the amount of added noise, we want to prevent the leakage of the added noise volume. Consequently, this problem relates to private privacy parameter selection, as label perturbation [50] (which we use to mimic different dataset quality levels) is one technique for achieving DP [10]. Although some works set the privacy parameter using economic incentives [1], we are not aware of any research considering defining the privacy parameter itself also privately.

7 Conclusion

Federated learning is the most popular collaborative learning framework, wherein each round only a subset of participants updates a joint machine learning model. Fortified with secure aggregation, only aggregated information is learned both by the participants and the server. Yet, in this paper, we devised a simple set of quality scoring rules that successfully recover the relative ordering of the participant's dataset qualities (measured by perturbed label ratio). Besides a small representative dataset to evaluate the improvement of the model after each aggregation, our method neither requires any computational power nor background information.

Through a series of image recognition experiments, we showed that it is possible to restore the relative ordering based on label quality with reasonably high accuracy. Our experiments also revealed a connection between the accuracy of the quality inference and both the complexity of the task and the used architecture. Moreover, we performed an ablation study suggesting that the original rules are near optimal. Lastly, we demonstrated how quality inference could i) boost training efficiency by weighting the participants, ii) yield an operational contribution metric, and iii) detect misbehaving participants based on their quality scores.

Limitations and Future Work

This paper has barely scratched the surface of quality inference based only on aggregated updates. We foresee multiple avenues towards improving and extending this work, e.g., using machine learning techniques to replace our naive rules by relaxing the attacker constraints concerning computational power and background knowledge. In the early rounds, selecting the participants in a non-random manner similar to [51] could also be beneficial.

For clarity, we have restricted our experiments to visual recognition tasks with noisy labels as the measure of data quality. Although we expect our results to generalize well to other domains, we leave further experiments as future work. Finally, the personal data protection implications of

the information leakage caused by quality inference is also of interest: should such quality information be considered private, and, consequently, should it fall under data protection regulations such as the GDPR? This issue has significant practical relevance to federated learning platforms already in operation.

REFERENCES

[1] B. Pejo, Q. Tang, and G. Biczok, "Together or alone: The price of privacy in collaborative learning," *Proceedings on Privacy Enhancing Technologies*, 2019.

[2] R. Cramer, I. B. Damgård *et al.*, *Secure multiparty computation*. Cambridge University Press, 2015.

[3] C. Gentry *et al.*, *A fully homomorphic encryption scheme*. Stanford university Stanford, 2009.

[4] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated Learning: Strategies for Improving Communication Efficiency," *arXiv:1610.05492 [cs]*, 2016.

[5] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015.

[6] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017.

[7] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Advances in Neural Information Processing Systems*, 2019.

[8] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction apis," in *25th USENIX Security Symposium (USENIX Security 16)*, 2016.

[9] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019.

[10] D. Desfontaines and B. Pejó, "Sok: Differential privacies," *Proceedings on Privacy Enhancing Technologies*, 2020.

[11] H. B. McMahan, E. Moore, D. Ramage, S. Hampson *et al.*, "Communication-efficient learning of deep networks from decentralized data," *arXiv preprint arXiv:1602.05629*, 2016.

[12] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002.

[13] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020.

[14] C. Batini, M. Scannapieco *et al.*, "Data and information quality," *Cham, Switzerland: Springer International Publishing. Google Scholar*, 2016.

[15] T. Wang, J. Rausch, C. Zhang, R. Jia, and D. Song, "A principled approach to data valuation for federated learning," in *Federated Learning*. Springer, 2020.

[16] IMDB, "The Good, the Bad and the Ugly," 1966, https://www.imdb.com/title/tt0060196/.

[17] D. Harville, "Extension of the gauss-markov theorem to include the estimation of random effects," *The Annals of Statistics*, 1976.

[18] L. C. Ludeman, *Random processes: filtering, estimation, and detection*. John Wiley & Sons, Inc., 2003.

[19] R. Kerkouche, G. Ács, and C. Castelluccia, "Federated learning in adversarial settings," *arXiv preprint arXiv:2010.07808*, 2020.

[20] J. H. Zar, "Spearman rank correlation," *Encyclopedia of Biostatistics*, 2005.

[21] P. Diaconis and R. L. Graham, "Spearman's footrule as a measure of disarray," *Journal of the Royal Statistical Society: Series B (Methodological)*, 1977.

[22] L. Deng, "The mnist database of handwritten digit images for machine learning research [best of the web]," *IEEE Signal Processing Magazine*, 2012.

[23] A. Krizhevsky, V. Nair, and G. Hinton, "The cifar-10 dataset," *online: http://www. cs. toronto. edu/kriz/cifar. html*, 2014.

[24] D. L. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," *Communications of the ACM*, 1981.

[25] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," *arXiv preprint arXiv:1712.07557*, 2017.

[26] S. Arora, E. Hazan, and S. Kale, "The multiplicative weights update method: a meta-algorithm and applications," *Theory of Computing*, 2012.

[27] L. S. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, 1953.

[28] A. Ghorbani and J. Zou, "Data shapley: Equitable valuation of data for machine learning," *arXiv preprint arXiv:1904.02868*, 2019.

[29] J. Huang, C. Hong, L. Y. Chen, and S. Roos, "Is shapley value fair? improving client selection for mavericks in federated learning," *arXiv preprint arXiv:2106.10734*, 2021.

[30] C. Fung, C. J. Yoon, and I. Beschastnikh, "Mitigating sybils in federated learning poisoning," *arXiv preprint arXiv:1808.04866*, 2018.

[31] I. Dinur and K. Nissim, "Revealing information while preserving privacy," in *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2003.

[32] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina, "Incentives for combatting freeriding on p2p networks," in *European Conference on Parallel Processing*. Springer, 2003.

[33] P. Van Mieghem, "Human psychology of common appraisal: The reddit score," *IEEE Transactions on Multimedia*, 2011.

[34] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, "Detecting product review spammers using rating behaviors," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010.

[35] L. Thomas, J. Crook, and D. Edelman, *Credit scoring and its applications*. SIAM, 2017.

[36] J. Lin, M. Du, and J. Liu, "Free-riders in federated learning: Attacks and defenses," *arXiv preprint arXiv:1911.12560*, 2019.

[37] Y. Fraboni, R. Vidal, and M. Lorenzi, "Free-rider attacks on model aggregation in federated learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021.

[38] B. Liu, B. Yan, Y. Zhou, J. Wang, L. Liu, Y. Zhang, and X. Nie, "Fedcm: A real-time contribution measurement method for participants in federated learning," *arXiv preprint arXiv:2009.03510*, 2021.

[39] J. So, R. E. Ali, B. Guler, J. Jiao, and S. Avestimehr, "Securing secure aggregation: Mitigating multi-round privacy leakage in federated learning," *arXiv preprint arXiv:2106.03328*, 2021.

[40] Y. Chen, X. Yang, X. Qin, H. Yu, B. Chen, and Z. Shen, "Focus: Dealing with label quality disparity in federated learning," *arXiv preprint arXiv:2001.11359*, 2020.

[41] J. Guo, Z. Liu, K.-Y. Lam, J. Zhao, Y. Chen, and C. Xing, "Secure weighted aggregation for federated learning," *arXiv preprint arXiv:2010.08730*, 2020.

[42] J. Huang, R. Talbi, Z. Zhao, S. Boucchenak, L. Y. Chen, and S. Roos, "An exploratory analysis on users' contributions in federated learning," in *2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. IEEE, 2020.

[43] B. Rozemberczki, L. Watson, P. Bayer, H.-T. Yang, O. Kiss, S. Nilsson, and R. Sarkar, "The shapley value in machine learning," *arXiv preprint arXiv:2202.05594*, 2022.

[44] J. Castro, D. Gómez, and J. Tejada, "Polynomial calculation of the shapley value based on sampling," *Computers & Operations Research*, 2009.

[45] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," *arXiv preprint arXiv:1703.04730*, 2017.

[46] B. Pejó, G. Biczók, and G. Ács, "Measuring contributions in privacy-preserving federated learning," *ERCIM NEWS*, p. 35, 2021.

[47] H. Hu, Z. Salcic, L. Sun, G. Dobbie, and X. Zhang, "Source inference attacks in federated learning," *arXiv preprint arXiv:2109.05659*, 2021.

[48] L. Wang, S. Xu, X. Wang, and Q. Zhu, "Eavesdrop the composition proportion of training labels in federated learning," *arXiv:1910.06044 [cs, stat]*, 2019.

[49] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi, "Beyond inferring class representatives: User-level privacy leakage from federated learning," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019.

[50] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," *arXiv preprint arXiv:1610.05755*, 2016.

[51] Z. Liu, Y. Chen, H. Yu, Y. Liu, and L. Cui, "Gtg-shapley: Efficient and accurate participant contribution evaluation in federated learning," *arXiv preprint arXiv:2109.02053*, 2021.