CHALMERS
UNIVERSITY OF TECHNOLOGY

Detect & Score: Privacy-Preserving Misbehavior Detection and Contribution Evaluation in Federated Learning

Detect & Score: Privacy-Preserving Misbehavior Detection and Contribution Evaluation in Federated Learning

Marvin Xhemrishi
TUM School of Computation, Information and Technology
Technical University of Munich
München, Bayern, Germany
marvin.xhemrishi@tum.de

Alexandre Graell i Amat
Department of Electrical Engineering
Chalmers University of Technology
Gothenburg, Västra Götalands län
Sweden
alexandre.graell@chalmers.se

Balazs Pejo
CrySyS Lab
BME
Budapest, Hungary
HUN-REN-BME Information Systems Research Group
Budapest, Hungary
pejo@crysys.hu

Abstract

Federated learning with secure aggregation enables private and collaborative learning from decentralized data without leaking sensitive client information. However, secure aggregation also complicates the detection of malicious client behavior and the evaluation of individual client contributions to the learning. To address these challenges, QI (Pejo *et al.*) and FedGT (Xhemrishi *et al.*) were proposed for contribution evaluation (CE) and misbehavior detection (MD), respectively. QI, however, lacks adequate MD accuracy due to its reliance on the random selection of clients in each training round, while FedGT lacks the CE ability.

In this work, we combine the strengths of QI and FedGT to achieve both robust MD and accurate CE. Our experiments demonstrate superior performance compared to using either method independently.

Keywords

Contribution evaluation, federated learning, misbehavior detection.

ACM Reference Format:

Marvin Xhemrishi, Alexandre Graell i Amat, and Balazs Pejo. 2025. Detect & Score: Privacy-Preserving Misbehavior Detection and Contribution Evaluation in Federated Learning. In *International Workshop on Secure and Efficient Federated Learning (FL-AsiaCCS '25), August 25–29, 2025, Hanoi, Vietnam.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3709023.3737692

1 Introduction

Federated learning (FL) has emerged as a promising paradigm for privacy-preserving decentralized learning. Unlike centralized learning, which requires aggregating data from all participants into a central server, FL enables multiple clients to train a shared model locally on their private data, exchanging only model updates rather than raw data. However, despite this decentralized design, research has exposed privacy vulnerabilities—revealing that sensitive information about the underlying datasets can still be inferred from the shared model updates. Notable attacks include model inversion attacks [9], membership inference attacks [24], reconstruction attack [40], (hyper)parameter inference [28], and property inference [17].

To mitigate these risks, several privacy-preserving techniques have been proposed, most notably differential privacy (DP) [5] and secure aggregation (SA) [16]. DP offers formal privacy guarantees, but this often comes at the cost of reduced model utility. In contrast, SA obscures individual updates without degrading model performance, making it an attractive solution for many applications. In essence, SA hides the individual model updates by cryptographically aggregating them, ensuring that only the final aggregated model is visible to the server.

While SA effectively protects privacy by concealing individual client updates, it also introduces a significant limitation: the server can no longer inspect individual contributions. This makes it considerably more difficult to detect whether a client has performed a poisoning [27] or backdoor[2] attack, or to evaluate the relative importance of clients with respect to one other and the learning task [12]. As a result, traditional misbehavior detection (MD) methods (e.g., KRUM [4]), contribution evaluation (CE) techniques (e.g., LOO [31]), and joint approaches (e.g., FedSV [18]) cannot be applied under secure aggregation, as they rely on access to individual model updates. Likewise advanced CE methods, such as Zeno [37] and Shapley-value based approaches [23], are not compatible with SA.

To address this challenge, Xhemrishi *et al.* proposed FedGT [36], a scheme for client MD under FL with SA tailored to the cross-silo setting. In parallel, and Pejo *et al.* introduced QI [19], a CE framework designed for the cross-device setting under SA.

Contribution. In this work, we extend the schemes in [36] and [19] and integrate their core ideas to develop a novel approach that supports both MD and CE in a cross-silo FL setting under SA. The proposed scheme leverages spatial and temporal information to enhance both tasks, and empirical results show that it outperforms FedGT, QI, and other relevant baselines in terms of both MD and CE performance.

2 Background & Related Works

2.1 Privacy-Preserving Misbehavior Detection

FL introduces unique challenges in handling malicious behavior due to its distributed nature. Byzantine attackers can disrupt training through two primary attack vectors: data poisoning and model poisoning. Data poisoning [8] involves modifying local training data to mislead the global model, while model poisoning [32] directly

FL-AsiaCCS '25, Hanoi, Vietnam

ACM ISBN 979-8-4007-1420-7/25/08
https://doi.org/10.1145/3709023.3737692

manipulates local model parameters to corrupt it. Both attack types can be targeted (affecting specific classes or features) or untargeted (degrading model performance globally).

Broadly, defense mechanisms against adversarial clients in FL fall into two main categories: mitigation and detection. Mitigation techniques aim to reduce the impact of malicious clients by modifying the aggregation process itself. Common approaches include robust statistical methods such as metric-based comparisons using the Euclidean distance [4] or cosine similarity [10]. These methods assess local updates relative to each other or against a baseline (i.e., the global model). The works [22, 25] propose robust aggregation techniques that mitigate the effect of poisoned models on the global model utility [22, 25] without requiring access to individual local updates.

While mitigation strategies seek to minimize the influence of malicious clients, identifying which clients are malicious remains a crucial real-world problem [14]. This is tackled by detection methods, which classify clients as honest or malicious, allowing the server to selectively exclude suspicious participants from training.

Privacy preservation and MD impose a natural trade-off because the former tries to hide the client's attributes (data, model updates, etc.), while the latter requires that the server learns more about the clients, such that a comparison is performed in an informative manner. However, both aspects of FL are important, and hence, it is crucial to balance this trade-off.

FedGT. Identification of malicious clients in a privacy-preserving manner was addressed in [36], where the authors proposed FedGT. Inspired by group testing, FedGT [36] enables MD under SA. The key idea is to group clients into overlapping groups, with each group performing SA. The server only receives the aggregated model updates from each group. By carefully designing the group structure and leveraging a decoding algorithm, the server can identify malicious clients based on the aggregated group outputs. As illustrated in Fig. 1c, FedGT was proposed for single-round testing in a cross-silo FL setting (in the illustrated example, only in round 2). Clients are grouped into overlapping groups according to an assignment matrix $\boldsymbol{A}$. After the group aggregates are collected, the server applies a testing algorithm followed by a decoding step to determine which clients are likely to be malicious.

The primary goal of FedGT is not to identify malicious clients, but to do so in order to achieve high good model utility even in the presence of malicious clients. In extensive experiments, FedGT outperformed state-of-the-art private robust aggregation methods [22, 25]—which do not support the ability to identify malicious clients—in terms of global model utility and communication efficiency. As noted in [36], FedGT's MD performance could be further improved by extending it to a multi-round testing scheme, which we pursue in this work.

Baseline. As a baseline for MD, we use *cosine similarity* (COS). COS is a versatile tool in machine learning, applied to both accelerate convergence [35] and defend against adversarial attacks [41]. Here, we assign scores to clients based on the similarity between their local models and the global aggregated gradient. While this approach is simplistic, it is highly efficient, requiring only the computation of an inner (dot) product. However, in its standard form, COS is not privacy-preserving since it depends on client-specific models. Nevertheless, its computational simplicity allows it to be implemented using homomorphic encryption (HE) [1].

2.2 Privacy-Preserving Contribution Evaluation

CE can be broadly categorized into three disciplines [23]: Explainability [11], which assigns importance scores to individual data features, data evaluation [30], which assesses the contribution of individual data points, and contribution scoring [29], which evaluates the impact of entire datasets corresponding to FL clients. Within this work, we will focus on the latter.

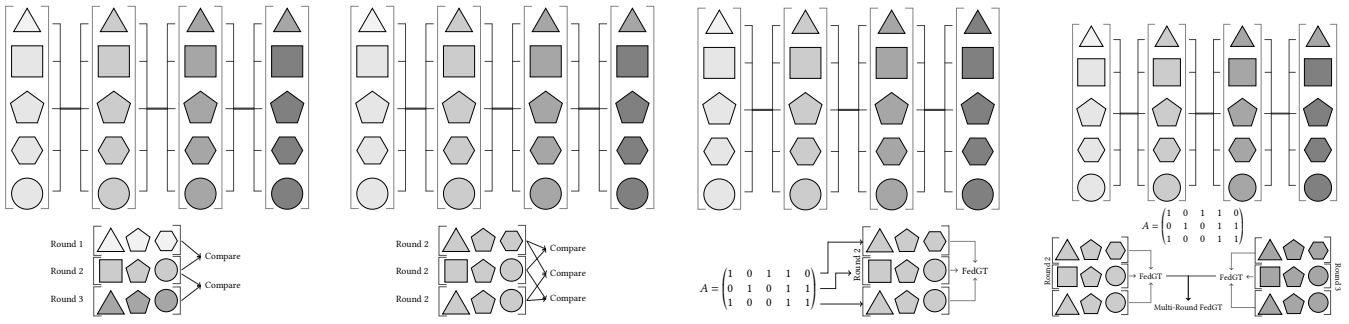
Another distinctive aspect of CE is the reliance on an external test dataset. Some schemes, such as gradient similarity-based approaches [38], avoid test sets and instead compute distance metrics between local and global models or gradients. The assumption is that clients whose updates closely align with the global model contribute more. In this paper, we assume the availability of an external test dataset.

A flagship technique for CE is the Shapley value [34], derived from cooperative game theory. It uniquely satisfies four key axioms, making it the only theoretically fair method for distributing rewards among players. The score of a client is computed by averaging its marginal contributions across all possible subsets of other clients. Yet, this computation is infeasible in real-world settings due to its exponential complexity. Consequently, several model-agnostic approximations have been developed, but most rely on individual-level information (such as the gradients and model updates). This exposes client data to potential privacy risks, an aspect largely overlooked in existing literature.

There are only a handful of work dealing with privacy-preserving CE. In [39], the authors proposed a multi-server solution relying on encryption, while in [33] they utilized DP, and [15] built a solution on the blockchain technology. Regarding marginal contributions, in [20, 21], the authors proposed new techniques where the clients either conduct a self-evaluation or they evaluate everybody else, respectively. This work improves on [19] where the authors proposed *Quality Inference* (QI), a group comparison solution that takes advantage of the client selection process in cross-device FL.

Quality Inference. QI operates within an honest-but-curious cross-device setting, where both the aggregator server and participants strictly adhere to the FL protocol. In an ideal world, during learning, it is expected that 1) the model improves every round, but 2) the rate of improvement decreases every round. QI captures deviations from these patterns by scoring the participating clients. In its core, as illustrated by Fig. 1a, QI relies on three scoring rules: *The good*, *the bad*, and *the ugly*. The authors of [19] experimented with various combinations of these rules as well as scaling the scores, but these had a limited effect. Here, we will consider the basic setting where equal weight is assigned to these unit scores. In more detail, the clients are scored according to the good and the bad (if the current round improved the model more than the previous round, the selected clients for this and the previous round gain $+1$ and -1, respectively) and the ugly (if the current round does not improves the model, the selected clients gain -1) rules.

Baseline. Within this work we will use the well-known Leave-One-Out (LOO) as our baseline. LOO is widely used in machine learning,



(a) Cross-Device QI. **(b) Cross-Silo QI.** **(c) Cross-Silo FedGT.** **(d) Cross-Silo Multi-Round FedGT.**

Figure 1: Illustration of the envisioned settings where the shapes represent clients and the grayness represents the rounds.

Table 1: Notations used in the paper for FedGT and QI.

	Symbol	Description
FL	N & T	Number of clients & Number of epochs
	$\boldsymbol{M}_n^t$	Client n's locally trained model in round t
	$\boldsymbol{M}^t$	Aggregated model of clients in round t
Dev.	$\boldsymbol{S}_{N\times T}$	Binary client selection matrix for each round
	K	Number of clients participating in a round
Silo	τ	Training round where the FedGT tests are performed
	L & k	Number of groups and their size for testing
	$\boldsymbol{A}_{N\times L}$	Assignment matrix for in-round testing
	$\hat{\boldsymbol{M}}_l$	Aggregated model of clients in test group l

including stability [7] and fairness [3]. LOO measures the marginal difference between the global model's performance with and without a single client's update. Hence, similarly to COS, LOO is also not inherently privacy-preserving, but its complexity is linear, which makes its computation feasible with HE.

3 FedGT and QI for Misbehavior Detection and Contribution Evaluation

In this section, we cross-validate FedGT and QI on MD and CE. The solutions for MD (aka detecting malicious clients) and CE (aka scoring participants) overlap and are somewhat interchangeable. While the original goal for FedGT was the former, QI aimed at the latter, so it is ambiguous in which domain they should be evaluated or measured. We envision several comparisons and combinations, and for brevity reasons, Fig. 1 illustrates a few of them.

Notations. In Table 1, we summarize the basic notations for FedGT and QI, such as $\boldsymbol{M}_n^t$ (model trained locally by client n in round t), $\boldsymbol{M}^t$ (aggregated model in round t determined via selection matrix $\boldsymbol{S}$ for cross-device FL), and $\hat{\boldsymbol{M}}_l$ (aggregated model of group l determined via assignment matrix $\boldsymbol{A}$). K and k corresponds to the number of clients affiliated with $\boldsymbol{M}^t$ and $\hat{\boldsymbol{M}}_l$, and as usual, N and T stands for the overall number of clients and iterations, respectively. τ is the round in which the testing is utilized, and L is the number of groups for the testing (e.g., for FedGT).

3.1 Cross-Silo Federated Learning

3.1.1 Adopting QI for Cross-Silo FL. As illustrated in Fig. 1a, QI performs comparisons between groups defined by the random selection matrix $\boldsymbol{S}$, where each group has size K. However, in a cross-silo setting, all clients participate in every round, making the original QI approach inapplicable. FedGT resolves this by creating subgroups of clients based on the assignment matrix $\boldsymbol{A}$ (as shown in Fig. 1c), which is designed based on error-correcting codes. Building on this idea, QI can also be adapted to use these structured groups, and perform pairwise comparisons between them, as shown in Fig. 1b. Rather than comparing randomly-formed groups across different rounds, comparisons are now made between carefully designed groups within the same round. This adaptation shifts from performing $2\cdot(T-1)$ temporal comparisons to $L\cdot(L-1)$ spatial comparisons, leveraging the group structure for more effective evaluation in the cross-silo setting.

3.1.2 Using FedGT for CE. FedGT is originally designed to flag attackers using an optimal soft-decoding algorithm. This decoder utilizes likelihood ratios, which are soft labels for a binary outcome (malicious or not) and therefore, might be suitable for CE: the smaller or bigger the likelihood, presumably the higher or lower the corresponding client's data quality.

3.1.3 Multi-Round. Compared to QI, FedGT relies on a single round. Extending it to multiple rounds (as illustrated in Fig. 1d) would allow for more information regarding clients behavior in comparison with the single-round testing. Thus, it is expected that the MD performance would increase when the likelihoods are aggregated across test rounds. This extension implies that τ is a set (rather than a number) where the group testing commences. We envision three strategies for forming the assignment matrix across rounds:

- **Same**: Use the same test groups across all rounds, i.e., $\boldsymbol{A}^{\tau_i} = \boldsymbol{A}^{\tau_j}$ for all $\tau_i, \tau_j \in \tau$.
- **Prefixed**: Use different but predetermined test groups for each round, where $\boldsymbol{A}^{\tau_i}$ depends only on the round index i for all $\tau_i \in \tau$.
- **Adaptive**: Use adaptive test groups for each round, where $\boldsymbol{A}^{\tau_i}$ depends on $\boldsymbol{A}^{\tau_j}$ and on $\mathcal{G}^{\tau_j}$ for all $\tau_j < \tau_i$.

In this work, we focus on the Prefixed strategy, which we refer to as multi-round FedGT (MR-FedGT). Experimental results showed that the Same strategy provided no improvement, while exploration of the Adaptive strategy is left for future work.

Alternatively, QI can also be extended to the multi-round setting. In this case, a single group defined by the assignment matrix

$\boldsymbol{A}^{\tau_i}$ is compared with the other $L-1$ groups from the same round (within-round comparison, as shown in Fig. 1b) as well as with the $2 \cdot L$ subgroups defined by $\boldsymbol{A}^{\tau_{i-1}}$ and $\boldsymbol{A}^{\tau_{i+1}}$ (across-round comparison, as illustrated in Fig. 1a). Since this scheme considers a larger set of group comparisons—unlike FedGT, which operates only within rounds—it is expected to achieve superior performance. We refer to this extension as multi-round QI (MR-QI). Similarly to MR-FedGT, MR-QI adopts the Prefixed strategy for constructing multiple assignment matrices across rounds.

The assignment matrix $\boldsymbol{A}$ is designed to ensure that no client's raw model $\boldsymbol{M}_n^t, \forall n \in [N]$ can be inferred from any linear combination of the tested groups within a single round. However, as the learning process progresses toward convergence, models from consecutive rounds become increasingly similar, i.e., models $\boldsymbol{M}_n^t$ and $\boldsymbol{M}_n^{t+1}$ are nearly identical when t is large. Therefore, when considering both $\boldsymbol{A}^{\tau_i}$ and $\boldsymbol{A}^{\tau_{i+1}}$ simultaneously, if τ_i is sufficiently large and τ_{i+1} is close to it, the privacy guarantees may deteriorate. Designing assignment matrices that account for such subtle correlations—whether under Prefixed or Adaptive strategies—remains an open problem and is left for future work.

3.2 Cross-Device Federated Learning

As illustrated on Fig. 1c, FedGT utilizes the groups defined by the assignment matrix $\boldsymbol{A}$, where the size of each set is k. In contrast, in a cross-device setting, K clients participate in each round; hence, the vanilla FedGT is not applicable. QI resolves this by comparing the random groups created by the client selection mechanism (as shown in Fig. 1a). Hence, FedGT can also use the groups defined by the round: the clients are grouped as per the rows of the assignment matrix in different rounds and only later are the groups tested, after the last sampling round has been collected. This strategy requires that the client sampling is performed not randomly, but in a structured manner following $\boldsymbol{A}$, which may introduce bias in the model. By chance, the random selection process might define the groups as $\boldsymbol{A}$, but they could correspond to very different iterations; thus, comparing them carries little meaning. Thus, a trade-off between the model's performance and the accuracy of MD and CE exists.

4 Experiments

This section details our experimental setup and key findings, i.e., we verify our adaptation of QI for the cross silo setting, experiment with FedGT for CE, and measure the performance improvement of multi-round techniques.

Setup. We conduct experiments for a classification problem over image datasets, namely the well-known CIFAR-10 and ISIC2019 datasets. The latter contains skin lesions and is tailored for cross-silo FL [6, 36]. We train Resnet-18 (with 0.05 as learning rate, 0.9 as momentum, and 0.001 as weight decay) and EfficientNet-B0 (pre-trained on ImageNet dataset, with 0.0005 as learning rate, 0.9 as momentum, and 0.0001 as weight decay) for CIFAR-10 and ISIC2019, respectively. We simulate FL using 15 clients; ISIC2019 corresponds to non-i.i.d. scenario (as it is inherited real-world heterogeneity) while we utilize two distribution scenarios for CIFAR-10: i.i.d and non-i.i.d using Dirichlet distribution with a parameter of 0.5. Measuring ground-truth contribution scores is difficult (as the Shapley value is computationally infeasible), so following [19], we introduce noise linearly to the labels to create variation: for client n each label is changed with probability $\frac{n}{N+1}$. To ensure statistically significant results, we fix the data splits and the injected noise and repeat each training process ten times.

Table 2: Detection performance (F1-score) of QI and FedGT where 1, 3, 5, and 7 clients are attackers (out of 15). The testing is performed in every round and the measurement per round is aggregated. The tabulated results are measured at the last training round. 1R denotes the single round approach (the maximum throughout all the rounds), while MR refers to multi-round approach.

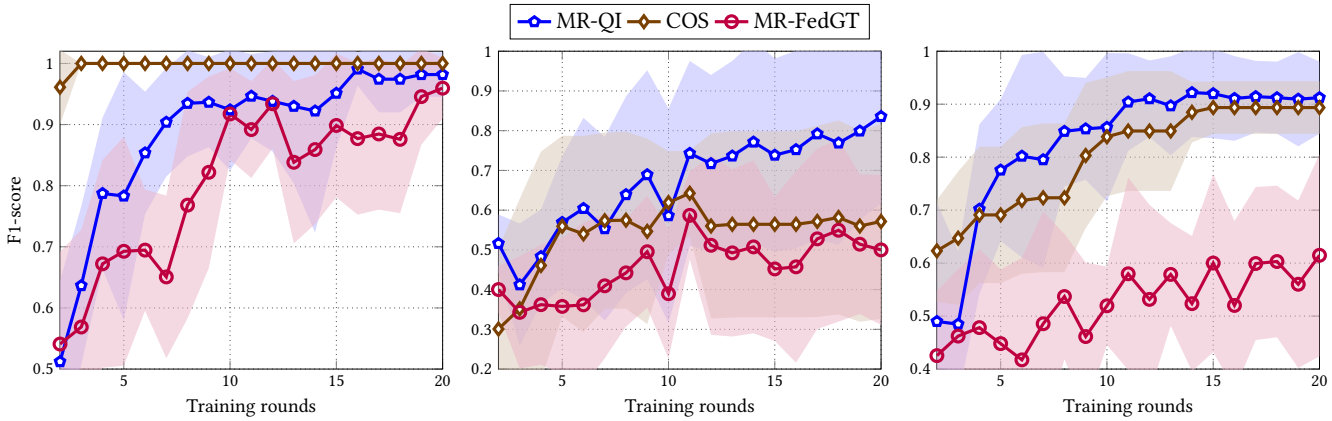
		Cross-Silo	MR-QI	MR-FedGT	COS	1R-QI	1R-FedGT
CIFAR-10	IID	1	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.70 ± 0.46	1.00 ± 0.00
		3	0.97 ± 0.06	1.00 ± 0.00	1.00 ± 0.00	0.67 ± 0.26	0.77 ± 0.21
		5	0.98 ± 0.04	0.96 ± 0.05	1.00 ± 0.00	0.74 ± 0.13	0.58 ± 0.14
		7	0.98 ± 0.05	0.98 ± 0.05	1.00 ± 0.00	0.79 ± 0.07	0.74 ± 0.09
	non-IID	1	0.72 ± 0.36	0.42 ± 0.35	0.61 ± 0.34	0.50 ± 0.50	0.30 ± 0.46
		3	0.87 ± 0.19	0.54 ± 0.21	0.74 ± 0.18	0.67 ± 0.26	0.37 ± 0.28
		5	0.84 ± 0.22	0.50 ± 0.19	0.57 ± 0.25	0.70 ± 0.18	0.44 ± 0.17
		7	0.88 ± 0.17	0.70 ± 0.20	0.38 ± 0.29	0.73 ± 0.08	0.59 ± 0.13
ISIC19	non-IID	1	0.88 ± 0.26	0.90 ± 0.30	1.00 ± 0.00	0.60 ± 0.49	0.40 ± 0.49
		3	0.80 ± 0.17	0.60 ± 0.19	0.96 ± 0.08	0.57 ± 0.21	0.27 ± 0.25
		5	0.91 ± 0.07	0.61 ± 0.19	0.89 ± 0.05	0.66 ± 0.13	0.44 ± 0.25
		7	0.91 ± 0.06	0.59 ± 0.11	0.68 ± 0.08	0.76 ± 0.07	0.56 ± 0.23

The experiments consist of 20 federated rounds where the clients perform 5 (for CIFAR-10) or 1 local epochs (for ISIC) using stochastic gradient descent optimization with 128 (for CIFAR-10) and 64 (for ISIC) as batch sizes. For experiments over CIFAR-10, we use the cross entropy loss, while for ISIC, we use the focal loss since it is shown to perform better over unbalanced datasets. Regarding the attack scenarios, we employ an untargeted label-flipping data-poisoning attack, offsetting the labels of the training data by shifting them by one. The components (testing algorithm, validation dataset size, decoder etc) of FedGT and the hyper-parameters are taken from the original FedGT paper [36].

As a baseline for MD, we use COS. While more advanced MD strategies exist, most are incompatible with privacy-preserving settings. Based on the obtained detection scores, we utilized a clustering approach to separate the suspected attackers and the anticipated benign clients based on agglomerative Clustering technique. For CE, we use LOO as our baseline. Note, that the scores are expected to be on different scales: FedGT scores are unbounded (as they rely on likelihood ratios), QI is mostly negative (as the majority of scoring rules are punishing the participants), and for LOO some scores are above while some are below zero (similarly to the Shapley Value). For fair comparison, we transform them via minimum offset correction (shifts all scores by subtracting the minimum score from each) and normalization (the scores are divided by the sum).

4.1 Adopting QI for Cross Silo

While the paper proposing QI [19] did consider MD, it only did so in the cross-device setting for IID clients, while FedGT [36], only considered the cross-silo setting. This paper investigates both the IID and non-IID settings and compares the schemes with COS. Our results are summarized on the right side of Table 2 for various numbers of attacks. These values suggest a correlation between the number of attackers and the detection performance, as more



(a) **CIFAR-10, IID.** (b) **CIFAR-10, non-IID.** (c) **ISIC19, non-IID.**

Figure 2: Detection performance (F1-score) of QI, FedGT and COS where 5 clients are malicious (out of 15) versus the communication rounds. The highlighted area represents the standard deviation obtained in our experiments.

Table 3: Scoring performance (the distance of the score vectors L_2 and their ordering differences ϕ) of QI, FedGT, and LOO. The testing is performed in every round and the tabulated results are measured at the last training round.

Cross-Silo			MR-QI	MR-FedGT	LOO
CIFAR	IID	L_2	0.015 ± 0.002	0.029 ± 0.007	0.036 ± 0.012
		ϕ	0.944 ± 0.027	0.761 ± 0.094	0.996 ± 0.004
	non-IID	L_2	0.025 ± 0.009	0.041 ± 0.007	0.034 ± 0.007
		ϕ	0.791 ± 0.146	0.429 ± 0.158	0.606 ± 0.246
ISIC	non-IID	L_2	0.036 ± 0.007	0.045 ± 0.009	0.037 ± 0.005
		ϕ	0.675 ± 0.097	0.299 ± 0.299	0.582 ± 0.111

attackers correspond to better detection (except for FedGT with IID). Comparing the last two columns, we can conclude that the simplistic QI rules outperform the more involved FedGT scheme when the data distribution is non-IID. At the same time, the comparison regarding the IID setting is inconclusive.

4.2 Multi-round MD

In multi-round MD, the testing is performed at each training round (except the first). We use QI with both across- and within-round comparisons, while for FedGT, we restrict ourselves to in-round comparisons as input for the testing algorithm. We tried to adopt FedGT for the Cross-Device setting (to enable across round testing too) by selecting one prefixed groups for each training round, but even with a Taylor-based interpolation (to normalize the improvement from different rounds to the same expected scale) FedGT is failing to provide meaningful results, so we discarded this option.

We present our results in the left side of Table 2. It is clear that performing the tests in multiple round significantly enhances both schemes, as the values on the left (multi-round) are consistently larger than on the right (single-round). However, the trend (more attackers imply better detection) visible for 1R is not holding for MR. Regarding the baseline COS mechanism (which we also applied in multiple round by accumulating the computed scores), it does slightly outperform the two privacy-preserving schemes in the IID setting. On the other hand, the more diverse the client's data distributions (ISIC can be considered low imbalance, CIFAR with Dirichlet(0.5) can be considered high imbalance), the worst COS performs. In contrast, QI seems to be robust against such changes and clearly outperforms both COS and FedGT in that setting (middle lines of the table). These results suggests, that CE schemes (such as QI) could be appropriate for MD as well; the fine-grained scoring can be turned into [0,1] (as benign and malicious) accurately.

Fig. 2 showcases how the detection accuracy behaves in respect to training rounds. Similar to the results in Table 2 the non-private scheme COS achieves almost perfect detection and it achieves it very fast (third round, see Figure 2a). However, the proposed multi-round QI performs well and reaches F1-score at seventh round. For experiments over CIFAR-10, where the client data is distributed according to Dirichlet with parameter 0.5 (Figure 2b), the proposed multi-round QI outperforms COS in almost every round. At the eleventh round, QI reaches a F1-score above 0.7. Due to the heterogeneity, all schemes suffer from a relatively high standard deviation. A similar pattern is observed for experiments over ISIC19 datasets, as one can see from Figure 2c. Except for the first two measures (recall that the measurement starts at the second round), the multi-round QI outperforms all the other schemes. However, it is important to note that the proposed multi-round FedGT improved compared to its original single-round form, but is outperformed by both COS and the proposed multi-round QI.

4.3 Using MR-FedGT for CE

The paper proposing FedGT [36] only considered MD, so here we are testing the hypothesis whether an MD mechanism is appropriate for CE. This is less studied in the literature, while the opposite direction is prevalent, see for instance [26]. Note we are injecting noise to the client labels as described earlier. Our results are summarized in Table 3 where the privacy-preserving CE methods are compare to the utilized noise ratios, that are also transformed similarly to the other scores described previously. The table presents the L_2 errors where smaller is the better and the Spearman correlation coefficient ϕ where larger is the better. The first shows the absolute difference between the transformed scores, while the latter captures

the ordering preservation of the clients: the metric ranges from $[-1, 1]$, where positive values indicate strong agreement and values near zero imply little to no correlation. This is a standard technique in CE [13] to determine the accuracy of inferring the top and bottom performing clients. These results suggest that mechanisms tailored for MD (such as FedGT) are not necessarily applicable to CE as FedGT performs poorly in this task. It is indeed tricky to turn a classification (attacker vs non-attacker) into regression (client-scoring). On the other hand, QI does outperforms LOO both in the score ordering and in the score distances, showing that simple privacy-preserving solutions could outperform naive scoring techniques which rely on individual differences (hence, need to be combined with encryption for any privacy guarantee).

5 Conclusion

We improved and combined two existing schemes, QI and FedGT, into a scheme that we coin multi-round QI. The proposed scheme outperforms their previous versions and baselines for both misbehavior detection and contribution evaluation for a cross-silo FL scenario. The proposed scheme achieves very good performance, especially when the data is non-i.i.d., while still preserving in-round clients privacy, due to compatibility with secure aggregation.

Acknowledgments. The authors are grateful to Gergely Biczók and Johan Östman for their comments on this work. This work was partially supported by the German Research Foundation (DFG) under Grant Agreement No. WA 3907/7-1, by the Swedish Research Council under grants 2020-03687 and 2023-05065, and by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Balázs Pejó was supported 1) by Project no. 145832 (implemented by the Ministry of Innovation and Technology from the NRDI Fund), 2) by the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory, and 3) by the European Union (Grant Agreement Nr. 10109571, SECURED Project).

References

[1] Abbas Acar, Hidayet Aksu, A Selcuk Uluagac, and Mauro Conti. 2018. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys (Csur)* (2018).
[2] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. 2020. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR.
[3] Emily Black and Matt Fredrikson. 2021. Leave-one-out unfairness. In *Proc. of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.
[4] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. 2017. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems* (2017).
[5] Damien Desfontaines and Balázs Pejó. 2020. Sok: Differential privacies. *Proc. on Privacy Enhancing Technologies* (2020).
[6] Jean Ogier du Terrail, Samy-Safwan Ayed, Edwige Cyffers, and et al. 2022. FLamby: datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. In *Proc. 36th Int. Conf. on Neural Infor. Process. Syst.*
[7] Theodoros Evgeniou, Massimiliano Pontil, and André Elisseeff. 2004. Leave one out error, stability, and generalization of voting combinations of classifiers. *Machine learning* (2004).
[8] Jiaxin Fan, Qi Yan, Mohan Li, Guanqun Qu, and Yang Xiao. 2022. A survey on data poisoning attacks and defenses. In *2022 7th IEEE International Conference on Data Science in Cyberspace (DSC)*. IEEE.
[9] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proc. of the 22nd ACM SIGSAC Conference on Computer and Communications Security*.
[10] Clement Fung, Chris JM Yoon, and Ivan Beschastnikh. 2018. Mitigating sybils in federated learning poisoning. *arXiv preprint arXiv:1808.04866* (2018).
[11] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable AI. *Science robotics* (2019).
[12] Jiyue Huang, Rania Talbi, Zilong Zhao, Sara Boucchenak, Lydia Y Chen, and Stefanie Roos. 2020. An exploratory analysis on users' contributions in federated learning. In *2020 IEEE Int. Conf. Trust, Priv. and Sec. in Intel. Syst. and applications*.
[13] Kevin Jiang, Weixin Liang, James Y Zou, and Yongchan Kwon. 2023. Opendataval: a unified benchmark for data valuation. *Adv. in Neural Inf. Processing Syst.* (2023).
[14] Peter Kairouz, H. Brendan McMahan, and et al. 2021. Advances and Open Problems in Federated Learning. (2021). doi:10.1561/2200000083
[15] Shuaicheng Ma, Yang Cao, and Li Xiong. 2021. Transparent contribution evaluation for secure federated learning on blockchain. In *2021 IEEE 37th international conference on data engineering workshops (ICDEW)*. IEEE.
[16] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. 2016. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629* (2016).
[17] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. 2019. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE.
[18] Khaoula Otmani, Rachid El-Azouzi, and Vincent Labatut. 2024. FedSV: Byzantine-Robust Federated Learning via Shapley Value. In *ICC 2024-IEEE International Conference on Communications*. IEEE.
[19] Balázs Pejó and Gergely Biczók. 2023. Quality inference in federated learning with secure aggregation. *IEEE Transactions on Big Data* (2023).
[20] Balázs Pejó, Gergely Biczók, and Gergely Ács. [n. d.]. Measuring Contributions in Privacy-Preserving Federated Learning. *ERCIM NEWS* ([n. d.]).
[21] Balázs Pejó and Delio Jaramillo Velez. 2025. Inferring Contributions in Privacy-Preserving Federated Learning. *ERCIM NEWS* (2025).
[22] Krishna Pillutla, Sham M Kakade, and Zaid Harchaoui. 2022. Robust aggregation for federated learning. *IEEE Trans. on Signal Processing* (2022).
[23] Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Olivér Kiss, Sebastian Nilsson, and Rik Sarkar. 2022. The Shapley Value in Machine Learning. *arXiv preprint arXiv:2202.05594* (2022).
[24] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE.
[25] Jinhyun So, Basak Guler, and Amir Salman Avestimehr. 2020. Byzantine-Resilient Secure Federated Learning. *IEEE Journal on Selected Areas in Communications* (2020). https://api.semanticscholar.org/CorpusID:220686676
[26] Qiheng Sun, Xiang Li, Jiayao Zhang, Li Xiong, Weiran Liu, Jinfei Liu, Zhan Qin, and Kui Ren. 2023. Shapleyfl: Robust federated learning based on shapley value. In *Proc. of the 29th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*.
[27] Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. 2020. Data poisoning attacks against federated learning systems. In *25th European Symposium on Research in Computer Security, ESORICS 2020*. Springer.
[28] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction apis. In *25th USENIX Security Symposium (USENIX Security 16)*.
[29] Guan Wang, Charlie Xiaoqian Dang, and Ziye Zhou. 2019. Measure contribution of participants in federated learning. In *2019 IEEE Int. Conf. on Big Data*. IEEE.
[30] Richard Y Wang and Diane M Strong. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems* (1996).
[31] Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. 2020. A principled approach to data valuation for federated learning. *Federated Learning: Privacy and Incentive* (2020).
[32] Zhilin Wang, Qiao Kang, Xinyi Zhang, and Qin Hu. 2022. Defense strategies toward model poisoning attacks in federated learning: A survey. In *2022 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE.
[33] Lauren Watson, Rayna Andreeva, Hao-Tsung Yang, and Rik Sarkar. 2022. Differentially private Shapley values for data evaluation. *arXiv:2206.00511* (2022).
[34] Eyal Winter. 2002. The shapley value. *Handbook of game theory with economic applications* (2002).
[35] Hongda Wu and Ping Wang. 2021. Fast-convergent federated learning with adaptive weighting. *IEEE Trans. on Cognitive Commun. and Networking* (2021).
[36] Marvin Xhemrishi, Johan Östman, Antonia Wachter-Zeh, and Alexandre Graell i Amat. 2025. FedGT: Identification of Malicious Clients in Federated Learning With Secure Aggregation. *IEEE Trans. on Inf. Forensics and Sec.* (2025).
[37] Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. 2019. Zeno: Distributed Stochastic Gradient Descent with Suspicion-based Fault-tolerance. arXiv:1805.10032
[38] Xinyi Xu, Lingjuan Lyu, Xingjun Ma, Chenglin Miao, Chuan Sheng Foo, and Bryan Kian Hsiang Low. 2021. Gradient driven rewards to guarantee fairness in collaborative machine learning. *Adv. in Neural Inf. Processing Syst.* (2021).
[39] Shuyuan Zheng, Yang Cao, and Masatoshi Yoshikawa. 2022. Secure Shapley Value for Cross-Silo Federated Learning (Technical Report). *arXiv preprint arXiv:2209.04856* (2022).
[40] Ligeng Zhu, Zhijian Liu, and Song Han. 2019. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*.
[41] Tengteng Zhu, Zehua Guo, Chao Yao, Jiaxin Tan, Songshi Dou, Wenrun Wang, and Zhenzhen Han. 2024. Byzantine-robust federated learning via cosine similarity aggregation. *Computer Networks* (2024).