Securing Secure Aggregation: Mitigating Multi-Round Privacy Leakage in Federated Learning

Abstract

Secure aggregation is a critical component in federated learning (FL), which enables the server to learn the aggregate model of the users without observing their local models. Conventionally, secure aggregation algorithms focus only on ensuring the privacy of individual users in a *single* training round. We contend that such designs can lead to significant privacy leakages over *multiple* training rounds, due to partial user selection/participation at each round of FL. In fact, we show that the conventional random user selection strategies in FL lead to leaking users' individual models within number of rounds that is linear in the number of users. To address this challenge, we introduce a secure aggregation framework, Multi-RoundSecAgg, with multi-round privacy guarantees. In particular, we introduce a new metric to quantify the privacy guarantees of FL over multiple training rounds, and develop a structured user selection strategy that guarantees the long-term privacy of each user (over any number of training rounds). Our framework also carefully accounts for the fairness and the average number of participating users at each round. Our experiments on MNIST and CIFAR-10 datasets in the IID and the non-IID settings demonstrate the performance improvement over the baselines, both in terms of privacy protection and test accuracy.

1 Introduction

Federated learning (FL) enables collaborative training of machine learning models over the data collected and stored locally by multiple data-owners (users). Training in FL is coordinated by a central server who maintains a global model that is updated locally by the users. The local updates are then aggregated by the server to update the global model. Throughout the training process, users never share their data with the server, i.e., the data is always kept on device, rather, they only share their local models. However, as has been shown recently, local models may still reveal substantial information about the local datasets, and the private training data can be reconstructed from the local models through inference or inversion attacks (see e.g., (Fredrikson, Jha, and Ristenpart 2015; Nasr, Shokri, and Houmansadr 2019; Zhu and Han 2020; Geiping et al. 2020)).

To prevent such information leakage, *secure aggregation* protocols are proposed (e.g., (Bonawitz et al. 2017; So, Güler, and Avestimehr 2021; Kadhe et al. 2020; Zhao and Sun 2021; Bell et al. 2020)) to protect the privacy of the local models, both from the server and the other users, while still allowing the server to learn their aggregate. More specifically, secure aggregation protocols ensure that, at any given round, the server can only learn the aggregate model of the users, and beyond that no further information is revealed about the local model of a particular user. Secure aggregation protocols, however, only ensure the privacy of the individual users in a *single training round*, and do not consider their privacy over multiple training rounds. On the other hand, due to partial user selection (Cho, Wang, and Joshi 2020; Chen, Horvath, and Richtarik 2020; Cho et al. 2020; Ribero and Vikalo 2020), the server may be able to reconstruct the individual models of some users using the aggregated models from the previous rounds. In fact, we show that after a sufficient number of rounds, all local models can be recovered with a high accuracy if the server chooses a random subset of the users to participate at every round. As a simple illustrative experiment, we provide an experiment with 40 users as shown in Fig. 1 in which all local updates can be recovered with small error.

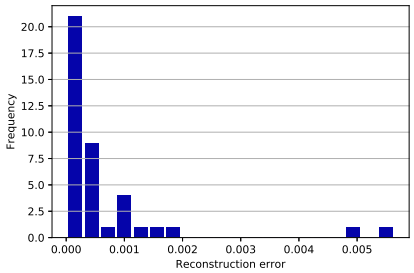


Figure 1: A FL non-IID setting is considered with 40 users, where 8 users are selected at random (if possible) at each round to train the neural network in (McMahan et al. 2017) on the MNIST dataset. Given the aggregate models of the users at each round, the server can approximate the individual gradients of all users with a very small error through the least-squares method. The reconstruction error is measured as the L_2 distance between the true gradients and reconstructed gradients and the histogram of the reconstruction error is shown. This setting is described in detail in Appendix G.

As such motivated, we study long-term user privacy in FL. Specifically, our contributions are as follows.

1. We introduce a new metric to capture long-term privacy guarantees in FL for the first time. This long-term privacy

condition requires that the server cannot reconstruct any individual model using the aggregated models from any number of training rounds. Using this metric, we show that the conventional random user selection schemes can result in leaking the individual user models after a sufficient number of rounds, even if secure aggregation is employed at each round.

2. We propose Multi-RoundSecAgg, a privacy-preserving structured user selection strategy that ensures the long-term privacy of the individual users over any number of training rounds. This strategy also takes into account the fairness of the selection process and the average number of participating users at each round.
3. We demonstrate that Multi-RoundSecAgg creates a trade-off between the long-term privacy guarantee and the average number of participating users. In particular, as the average number of participating users increases, the long-term privacy guarantee becomes weaker.
4. We provide the convergence analysis of Multi-RoundSecAgg, which shows that the long-term privacy guarantee and the average number of participating users control the convergence rate. The convergence rate is maximized when the average number of participating users is maximized. As we require stronger long-term privacy guarantees, the average number of participating users decreases and a larger number of training rounds is required to achieve the same level of accuracy as the random selection strategy.
5. Finally, our experiments in both IID and non-IID settings on MNIST and CIFAR-10 datasets demonstrate that Multi-RoundSecAgg achieves almost the same test accuracy compared to the random selection scheme while providing better long-term privacy guarantees.

2 Related Work

The underlying principle of the secure aggregation protocol in (Bonawitz et al. 2017) is that each pair of users exchange a pairwise secret key which they can use to mask their local models before sharing them with the server. The pairwise masks cancel out when the server aggregates the masked models, allowing the server to learn the aggregate of the local models. These masks also ensure that the local models are kept private, i.e., no further information is revealed beyond the aggregate of the local models. This protocol incurs a significant communication cost due to exchanging and reconstructing the pairwise keys. Recently, several works have developed communication-efficient protocols (So, Güler, and Avestimehr 2021; Kadhe et al. 2020; Bell et al. 2020; Tang et al. 2021; Choi et al. 2020; Elkordy and Avestimehr 2020), which are complementary to and can be combined with our work. Another line of work focused on designing partial user selection strategies to overcome the communication bottleneck in FL while speeding up the convergence by selecting the users based on their local loss (Cho, Wang, and Joshi 2020; Chen, Horvath, and Richtarik 2020; Cho et al. 2020; Ribero and Vikalo 2020).

Previous works, either on secure aggregation or on partial user selection, however, do not consider mitigating the potential privacy leakage as a result of partial user participation and the server observing the aggregated models across multiple training rounds. While (Pejó and Biczók 2020) pointed out to the privacy leakage of secure aggregation, mitigating this leakage has not been considered and our work is the first to address this challenge.

3 System Model

In this section, we first describe the basic federated learning model in Section 3.1. Next, we introduce the multi-round secure aggregation problem for federated learning and define the key metrics to evaluate the performance of a multi-round secure aggregation protocol in Section 3.2.

3.1 Basic Federated Learning Model

We consider a cross-device federated learning setup consisting of a server and N devices (users). User $i \in [N]$ has a local dataset $\mathcal{D}_i$ consisting of $m_i = |\mathcal{D}_i|$ data samples. The users are connected to each other through the server (McMahan et al. 2017; Bonawitz et al. 2017; Kairouz et al. 2019). The goal is to collaboratively learn a global model $\boldsymbol{x}$ with dimension d, using the local datasets that are generated, stored, and processed locally by the users. The training task can be represented by minimizing a global loss function,

$$\min_{\boldsymbol{x}} L(\boldsymbol{x}) \text{ s.t. } L(\boldsymbol{x}) = \sum_{i=1}^{N} w_i L_i(\boldsymbol{x}), \tag{1}$$

where L_i is the loss function of user i and $w_i \geq 0$ is a weight parameter assigned to user i to specify the relative impact of the user, where $\sum_i w_i = 1$. A common choice for the weight parameters is $w_i = \frac{m_i}{m}$, where $m = \sum_{i=1}^{N} m_i$ (Kairouz et al. 2019). We define the optimal model parameters $\boldsymbol{x}^*$ and $\boldsymbol{x}_i^*$ as $\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} L(\boldsymbol{x})$ and $\boldsymbol{x}_i^* = \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} L_i(\boldsymbol{x})$.

Federated Averaging with Partial User Participation. To solve (1), the most common algorithm is the *FedAvg* (federated averaging) algorithm (McMahan et al. 2017). *FedAvg* is an iterative algorithm, where model training is done by repeatedly iterating over individual local updates. At the beginning of training round t, the server sends the current state of the global model, denoted by $\boldsymbol{x}^{(t)}$, to the users. Each round consists of two phases, local training and aggregation. In the local training phase, user $i \in [N]$ updates the global model by carrying out E (≥ 1) local stochastic gradient descent (SGD) steps and sends the updated local model $\boldsymbol{x}_i^{(t)}$ to the server. One of key features of cross-device federated learning is partial device participation. Due to various reasons such as unreliable wireless connectivity, or battery issues, at any given round, only a fraction of the users are available to participate in the protocol. We refer to such users as *available* users throughout the paper. In the aggregation phase, the server selects $K \leq N$ users among the available users and aggregates the local models of the selected users. After receiving the local updates, the server updates the global model as follows

$$\boldsymbol{x}^{(t+1)} = \sum_{i\in\mathcal{S}^{(t)}} \boldsymbol{x}_i^{(t)} = {\mathbf{X}^{(t)}}^{\top} \boldsymbol{p}^{(t)}, \tag{2}$$

where $\mathcal{S}^{(t)}$ is the set of participating users at round t and $\boldsymbol{p}^{(t)} \in \{0,1\}^N$ is the corresponding characteristic vector. That is, $\boldsymbol{p}^{(t)}$ denotes a participation vector at round t whose i-th entry is 0 when user i is not selected and 1 otherwise. $\mathbf{X}^{(t)}$ denotes the concatenation of the weighted local models at round t, i.e., $\mathbf{X}^{(t)} = \left[w_1\boldsymbol{x}_1^{(t)}, \ldots, w_N\boldsymbol{x}_N^{(t)}\right]^\top \in \mathbb{R}^{N\times d}$. Finally, the server broadcasts the updated global model $\boldsymbol{x}^{(t+1)}$ to the users for the next round.

3.2 Multi-round Secure Aggregation

Conventional secure aggregation protocols only consider the privacy guarantees over a single training round. While secure aggregation protocols have provable privacy guarantees at any single round, in the sense that no information is leaked beyond the aggregate model at each round, the privacy guarantees do not extend to attacks *that span multiple training rounds*. Specifically, by using the aggregate models and participation information across multiple rounds, an individual model may be reconstructed. For instance, consider the following user participation strategy across three training rounds, $\boldsymbol{p}^{(1)} = [1,1,0]^\top$, $\boldsymbol{p}^{(2)} = [0,1,1]^\top$, and $\boldsymbol{p}^{(3)} = [1,0,1]^\top$. Assume a scenario where the local updates do not change significantly over time (e.g., models start to converge, or the server fixes the global model over consecutive rounds), i.e., $\boldsymbol{x}_i = \boldsymbol{x}_i^{(t)}$ for all $i \in [3]$ and $t \in \{0,1,2\}$. Then, the server can single out individual model, e.g., $\boldsymbol{x}_1 = \frac{\boldsymbol{x}^{(0)}+\boldsymbol{x}^{(2)}-\boldsymbol{x}^{(3)}}{2}$. Similarly, the server can single out all individual models $\boldsymbol{x}_i$, $i \in [3]$, even if a secure aggregation protocol is employed at each round.

In this paper, we study secure aggregation protocols with long-term privacy guarantees (which we term *multi-round secure aggregation*) for the cross-device FL setup which has not been studied before. We assume that user $i \in [N]$ drops from the protocol at each round with probability p_i. $\mathcal{U}^{(t)}$ denotes the index set of available users at round t and $\boldsymbol{u}^{(t)} \in \{0,1\}^N$ is a vector indicating the available users, such that $\{\boldsymbol{u}^{(t)}\}_j = \mathbb{1}\{j \in \mathcal{U}^{(t)}\}$ where $\{\boldsymbol{u}\}_j$ is j-th entry of $\boldsymbol{u}$ and $\mathbb{1}\{\cdot\}$ is the indicator function. The server selects K users from $\mathcal{U}^{(t)}$, if $|\mathcal{U}^{(t)}| \geq K$, based on the history of selected users in previous rounds. The local models of the selected users are then aggregated via a secure aggregation protocol (i.e., by communicating masked models), at the end of which the server learns the aggregate of the local models of the selected users. Our goal is to design a user selection algorithm $\mathcal{A}^{(t)} : \{0,1\}^{t\times N} \times \{0,1\}^N \to \{0,1\}^N$,

$$\mathcal{A}^{(t)}\left(\mathbf{P}^{(t)}, \boldsymbol{u}^{(t)}\right) = \boldsymbol{p}^{(t)} \text{ such that } \|\boldsymbol{p}^{(t)}\|_0 \in \{0, K\}, \quad (3)$$

to prevent the potential information leakage over multiple rounds, where $\boldsymbol{p}^{(t)} \in \{0,1\}^N$ is the participation vector defined in (2), $\|\boldsymbol{x}\|_0$ denotes the L_0-"norm" of a vector $\boldsymbol{x}$ and K denotes the maximum number of selected users. We note that $\mathcal{A}^{(t)}$ can be a random function. $\mathbf{P}^{(t)}$ is a matrix representing the user participation information up to round t, and is termed the *participation matrix*, given by

$$\mathbf{P}^{(t)} = \left[\boldsymbol{p}^{(0)}, \boldsymbol{p}^{(1)}, \ldots, \boldsymbol{p}^{(t-1)}\right]^\top \in \{0,1\}^{t\times N}. \quad (4)$$

If $|\mathcal{U}^{(t)}| < K$, the server skips the aggregation phase.

Key Metrics. A multi-round secure aggregation protocol can be represented by $\mathcal{A} = \{\mathcal{A}^{(t)}\}_{t\in[J]}$, where $\mathcal{A}^{(t)}$ is the user selection algorithm at round t defined in (3) and J is the total number of rounds. The inputs of $\mathcal{A}^{(t)}$ are a random vector $\boldsymbol{u}^{(t)}$, which indicates the available users at round t, and the participation matrix $\mathbf{P}^{(t)}$ defined in (4) which can be a random matrix. Given the participation matrix $\mathbf{P}^{(J)}$, we evaluate the performance of the corresponding multi-round secure aggregation protocol through the following metrics.

1. **Multi-round Privacy Guarantee.** We consider a security model where the server is honest-but-curious. The secure aggregation protocols ensure that the server can only learn the sum of the local models of some users in each single round, but they do not consider what the server can learn over the long run. Our multi-round privacy definition extends the guarantees of the secure aggregation protocols from one round to all training rounds by requiring that the server can only learn a sum of the local models even if the server exploits the aggregate models of all rounds. That is, our multi-round privacy guarantee is a natural extension of the privacy guarantee provided by the secure aggregation protocols considering a single training round.

   Specifically, a multi-round privacy guarantee T requires that any non-zero partial sum of the local models that the server can reconstruct, through any linear combination $\mathbf{X}^\top{\mathbf{P}^{(J)}}^\top\boldsymbol{z}$, where $\boldsymbol{z} \in \mathbb{R}^J \setminus \{\mathbf{0}\}$, must be of the form[1]

$$\begin{aligned}\mathbf{X}^\top{\mathbf{P}^{(J)}}^\top\boldsymbol{z} &= \sum_{i\in[n]} a_i \sum_{j\in\mathcal{S}_i} \boldsymbol{x}_j \\ &= a_1\sum_{j\in\mathcal{S}_1}\boldsymbol{x}_j + a_2\sum_{j\in\mathcal{S}_2}\boldsymbol{x}_j + \cdots + a_n\sum_{j\in\mathcal{S}_n}\boldsymbol{x}_j, \quad (5)\end{aligned}$$

   where $|\mathcal{S}_i| \geq T, a_i \neq 0, \forall i \in [n]$ and $n \in \mathbb{Z}^+$. Here all the sets $\mathcal{S}_i$, the number of sets n, and each a_i could all depend on $\boldsymbol{z}$. In equation (5), we consider the worst-case scenario, where the local models do not change over the rounds. That is, $\mathbf{X}^{(t)} = \mathbf{X}$, $\forall t \in [J]$. Intuitively, this guarantee ensures that the best that the server can do is to reconstruct a partial sum of T local models which corresponds to the case where $n = 1$. When $T \geq 2$, this condition implies that the server cannot get any user model from the aggregate models of all training rounds (the best it can obtain is the sum of two local models).

   Remark 1. (Weaker Privacy Notion). It is worth noting that, a weaker privacy notion would require that $\|{\mathbf{P}^{(J)}}^\top\boldsymbol{z}\|_0 \geq T$ when ${\mathbf{P}^{(J)}}^\top\boldsymbol{z} \neq \mathbf{0}$. When $T = 2$, this definition requires that the server cannot reconstruct any individual model (the best it can do is to obtain a linear combination of two local models). This definition, however, allows constructions in the form of $a\boldsymbol{x}_i + b\boldsymbol{x}_j$ for any $a, b \in \mathbb{R}\setminus\{0\}$. When $a \gg b$, however, this is almost the same as recovering $\boldsymbol{x}_i$ perfectly, hence this privacy criterion is weaker than that of (5).

[1] We assume that $w_i = \frac{1}{N}, \forall i \in [N]$ in this paper.

Remark 2. (Multi-round Privacy of Random Selection). In Sec. 6, we empirically show that a random selection strategy in which K available users are selected at random at each round does not ensure multi-round privacy even with respect to the weaker definition of Remark 1. Specifically, the local models can be reconstructed within a number of rounds that is linear in N. We also show this theoretically in Appendix G.

Remark 3. (Worst-Case Assumption). In (5), we considered the worst-case assumption where the models do not change over time. When the local models change over rounds, the multi-round privacy guarantee becomes even stronger as the number of unknowns increases.

2. **Aggregation Fairness Gap.** The average aggregation fairness gap quantifies the largest gap between any two users in terms of the expected relative number of rounds each user has participated in training. Formally, the average aggregation fairness gap is defined as follows

$$F = \max_{i\in[N]} \limsup_{J\to\infty} \frac{1}{J}\mathbb{E}\Big[\sum_{t=0}^{J-1} \mathbb{1}\big\{\{\boldsymbol{p}^{(t)}\}_i = 1\big\}\Big] - \min_{i\in[N]} \liminf_{J\to\infty} \frac{1}{J}\mathbb{E}\Big[\sum_{t=0}^{J-1} \mathbb{1}\big\{\{\boldsymbol{p}^{(t)}\}_i = 1\big\}\Big], \quad (6)$$

where $\mathbb{1}\{\cdot\}$ is an indicator function, $\{\boldsymbol{p}^{(t)}\}_i$ is i-th entry of the vector $\boldsymbol{p}^{(t)}$, and the expectation is over the randomness of the user selection algorithm $\mathcal{A}$ and the user availability. The main intuition behind this definition is that when $F = 0$, all users participate on average on the same number of rounds. This is important to take the different users into consideration equally and our experiments show that the accuracy of the schemes with small F are much higher than the schemes with high F.

3. **Average Aggregation Cardinality.** The aggregation cardinality quantifies the expected number of models to be aggregated per round. Formally, it is defined as

$$C = \liminf_{J\to\infty} \frac{\mathbb{E}\big[\sum_{t=0}^{J-1} \|\boldsymbol{p}^{(t)}\|_0\big]}{J}, \quad (7)$$

where the expectation is over the randomness in $\mathcal{A}$ and the user availability. Intuitively, less number of rounds are needed to converge as more users participate in the training. In fact, as we show in Sec. 5.2, C directly controls the convergence rate.

3.3 Baseline Schemes

In this subsection, we introduce three baseline schemes for multi-round secure aggregation.

Random Selection. In this scheme, at each round, the server selects K users at random from the set of available users. If less than K users are available, the server skips this round.

Random Weighted Selection. This scheme is a modified version of random selection to reduce F when the dropout probabilities of the users are not equal. Specifically, K users are selected at random from the available users with the minimum frequency of participation in the previous rounds. If less than K users are available, the server skips this round.

User Partitioning (Grouping). In this scheme, the users are partitioned into $G = N/K$ equal-sized groups denoted as $\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_G$. At each round, the server selects one of the groups if none of the users in this group has dropped out. If multiple groups are available, to reduce the aggregation fairness gap, the server selects a group including a user with the minimum frequency of participation in previous rounds. If no group is available, the server skips this round.

4 Proposed Scheme: Multi-RoundSecAgg

In this section, we present Multi-RoundSecAgg, which has two components as follows.

- The first component designs a family of sets of users that satisfy the multi-round privacy requirement. The inputs of the first component are the number of users N, the number of users desired to be selected at each round K and the desired multi-round privacy guarantee T. The output is a family of sets of K users that satisfy the multi-round privacy guarantee T, termed as a *privacy-preserving family*. This family is represented by a matrix $\mathbf{B}$, where the rows are the characteristic vectors of these user sets.
- The second component selects a set from this designed family to satisfy the fairness guarantee. The inputs to the second component are the privacy-preserving family $\mathbf{B}$, the set of available users at round t, $\mathcal{U}^{(t)}$, and the frequency of participation of each user. The output is the set of users that will participate at round t.

We now describe these two components in detail.

Component 1 (Batch Partitioning (BP) of the users to guarantee multi-round privacy). The first component designs a family of R_{BP} sets, where R_{BP} is given in (8), satisfying the multi-round privacy requirement T. We denote the $R_{\text{BP}} \times N$ binary matrix corresponding to these sets by $\mathbf{B} = [\boldsymbol{b}_1, \cdots, \boldsymbol{b}_{R_{\text{BP}}}]^\top$, where $\|\boldsymbol{b}_i\|_0 = K$ for all $i \in [R_{\text{BP}}]$. That is, the rows of $\mathbf{B}$ are the characteristic vectors of those sets. The main idea of our proposed scheme is to restrict certain sets of users of size T, denoted as batches, to either participate together or not participate at all. This guarantees a multi-round privacy T as we show in Section 5.

To construct a family of sets with this property, the users are first partitioned into N/T batches. At any given round, either all or none of the users of a particular batch participate in training. The server can choose K/T batches to participate in training, provided that all users in any given selected batch are available. Since there are $\binom{N/T}{K/T}$ possible sets with this property, then the size of this privacy-preserving family of sets is given by[2]

$$R_{\text{BP}} \stackrel{\text{def}}{=} \binom{N/T}{K/T}. \quad (8)$$

In the extreme case of $T = 1$, this strategy specializes to the random selection strategy where the server can choose any set from the $\binom{N}{K}$ possible sets of K users. In the other extreme case of $T = K$, this strategy specializes to the

[2]We assume for simplicity that N/T and K/T are integers.

partitioning strategy where there are N/K possible sets. We next provide an example to illustrate the construction of $\mathbf{B}$.

Example 1 ($N = 8, K = 4, T = 2$)**.** *In this example, the users are partitioned into* 4 *batches as* $\mathcal{G}_1 = \{1, 2\}, \mathcal{G}_2 = \{3, 4\}, \mathcal{G}_3 = \{5, 6\}$ *and* $\mathcal{G}_4 = \{7, 8\}$ *as given in (9). The server can choose any two batches out of these* 4 *batches, hence we have* $R_{BP} = \binom{4}{2} = 6$ *possible sets. This ensures a multi-round privacy* $T = 2$.

$$\mathbf{B} = \left[\begin{array}{cc|cc|cc|cc} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{array}\right]. \tag{9}$$

Component 2 (Available batch selection to guarantee fairness). At round t, user $i \in [N]$ is available to participate in the protocol with a probability $1 - p_i \in (0, 1]$. The frequency of participation of user i before round t is denoted by $f_i^{(t)} \stackrel{\text{def}}{=} \sum_{j=0}^{t-1} \mathbb{1}\left\{\{\boldsymbol{p}^{(j)}\}_i = 1\right\}$. Given the set of available users at round t, $\mathcal{U}^{(t)}$, and the frequencies of participation $\boldsymbol{f}^{(t-1)} = (f_1^{(t-1)}, \cdots, f_N^{(t-1)})^\top$, the server selects K users. To do so, the server first finds the submatrix of $\mathbf{B}$ denoted by $\mathbf{B}^{(t)}$ corresponding to the set of available users at round t, $\mathcal{U}^{(t)}$. Specifically, the i-th row of $\mathbf{B}$ denoted by $\boldsymbol{b}_i^\top$ is included in $\mathbf{B}^{(t)}$ provided that $\text{supp}(\boldsymbol{b}_i) \subseteq \mathcal{U}^{(t)}$. If $\mathbf{B}^{(t)}$ is an empty matrix, then the server skips this round. Otherwise, the server selects a row from $\mathbf{B}^{(t)}$ uniformly at random if $p_i = p, \forall i \in [N]$. If the users have different dropout probabilities, the server selects a row from $\mathbf{B}^{(t)}$ that includes the user with the minimum frequency of participation $\ell_{\min}^{(t-1)} \stackrel{\text{def}}{=} \arg\min_{i \in \mathcal{U}^{(t)}} f_i^{(t-1)}$. If there are many such rows, then the server selects one of them at random.

Overall, the algorithm first designs a privacy-preserving family of sets of K users such that any set can be selected at any given round while ensuring the multi-round privacy guarantee T. Then a specific set is selected from this family at each round to ensure fairness. We describe the two components of Multi-RoundSecAgg in Algorithm 1 and Algorithm 2, which are provided in Appendix D.

5 Theoretical Results

We provide the theoretical and convergence guarantees of Multi-RoundSecAgg in Sec. 5.1 and Sec.5.2, respectively.

5.1 Theoretical Guarantees of Multi-RoundSecAgg

We first establish the theoretical guarantees of Multi-RoundSecAgg in terms of the key metrics.

Theorem 1. *Multi-RoundSecAgg with parameters* N, K, T *ensures a multi-round privacy guarantee of* T*, an aggregation fairness gap* $F = 0$*, and an average aggregation cardinality given by*

$$C = K\left(1 - \sum_{i=N/T-K/T+1}^{N/T} \binom{N/T}{i} q^i (1-q)^{N/T-i}\right),$$

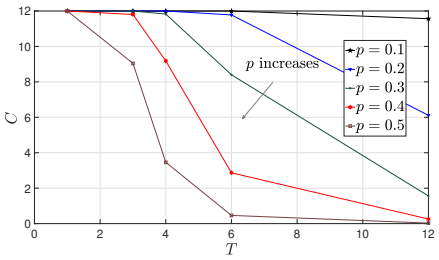


Figure 2: An illustration of the trade-off between the multi-round privacy guarantee T and the average aggregation cardinality C. In this example, $N = 120$ and $K = 12$.

where $q = 1 - (1 - p)^T$*, when all users have the same dropout probability* p.

We provide the proof of Theorem 1 in Appendix A.

Remark 4. (Trade-off between "Multi-round Privacy Guarantee" and "Average Aggregation Cardinality"). Theorem 1 indicates a trade-off between the multi-round privacy and the average aggregation cardinality since as T increases, C decreases which slows down the convergence as we show in Sec. 5.2. We illustrate this trade-off in Fig. 2.

Remark 5. (Necessity of Batch Partitioning (BP)). We show that any strategy that satisfies the privacy guarantee in Equation (5) must have a batch partitioning structure, and for given $N, K, T, K \leq N/2$, the largest number of distinct user sets in any strategy is at most $\binom{N/T}{K/T}$, which is achieved in our design in Section 4. We provide the proof in Appendix C.

Remark 6. (Non-linear Reconstructions of Aggregated Models). The privacy criterion in Equation (5) considers linear reconstructions of the aggregated models. One may also consider more general non-linear reconstructions. The long-term privacy guarantees of batch partitioning hold even under such reconstructions as the users in the same batch always participate together or do not participate at all. Hence, the server cannot separate individual models within the same batch even through non-linear operations.

5.2 Convergence Analysis of Multi-RoundSecAgg

We now provide the convergence guarantees of Multi-RoundSecAgg by first introducing a few common assumptions (Li et al. 2019; Yu, Yang, and Zhu 2019).

Assumption 1. $L_1, \ldots, L_N$ *in* (1) *are all* ρ*-smooth: for all* $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$ *and* $i \in [N]$, $L_i(\boldsymbol{a}) \leq L_i(\boldsymbol{b}) + (\boldsymbol{a} - \boldsymbol{b})^\top \nabla L_i(\boldsymbol{b}) + \frac{\rho}{2}\|\boldsymbol{a} - \boldsymbol{b}\|^2$.

Assumption 2. $L_1, \ldots, L_N$ *in* (1) *are all* μ*-strongly convex: for all* $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$ *and* $i \in [N]$, $L_i(\boldsymbol{a}) \geq L_i(\boldsymbol{b}) + (\boldsymbol{a} - \boldsymbol{b})^\top \nabla L_i(\boldsymbol{b}) + \frac{\mu}{2}\|\boldsymbol{a} - \boldsymbol{b}\|^2$.

Assumption 3. *Let* $\xi_i^{(t)}$ *be a sample uniformly selected from the local dataset* $\mathcal{D}_i$*. The variance of the stochastic gradients at each user is bounded, i.e.,* $\mathbb{E}\|\nabla L_i(\boldsymbol{x}_i^{(t)}, \xi_i^{(t)}) - \nabla L_i(\boldsymbol{x}_i^{(t)})\|^2 \leq \sigma_i^2$ *for* $i \in [N]$.

Assumption 4. *The expected squared norm of the stochastic gradients is uniformly bounded, i.e.,* $\mathbb{E}\|\nabla L_i(\boldsymbol{x}_i^{(t)}, \xi_i^{(t)})\|^2 \leq G^2$ *for all* $i \in [N]$.

We now state the convergence guarantees.

Theorem 2. *Consider a FL setup with N users to train a machine learning model from* (1)*. Assume K users are selected by Multi-RoundSecAgg with average aggregation cardinality C defined in* (7) *to update the global model from* (2)*, and all users have the same dropout rate, hence Multi-RoundSecAgg selects a random set of K users uniformly from the set of available user sets at each round. Then, the following is satisfied*

$$\mathbb{E}[L(\boldsymbol{x}^{(J)})] - L^* \leq \frac{\rho}{\gamma + \frac{C}{K}EJ - 1}\left(\frac{2(\alpha+\beta)}{\mu^2} + \frac{\gamma}{2}\mathbb{E}\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\|^2\right), \quad (10)$$

where $\alpha = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2 + 6\rho\Gamma + 8(E-1)^2G^2$, $\beta = \frac{4(N-K)E^2G^2}{K(N-1)}$, $\Gamma = L^* - \sum_{i=1}^{N} L_i^*$, *and* $\gamma = \max\left\{\frac{8\rho}{\mu}, E\right\}$.

We provide the proof of Theorem 2 in Appendix B.

Remark 7. Theorem 2 shows how the average aggregation cardinality affects the convergence. When the average aggregation cardinality is maximized, i.e., $C = K$, the convergence rate in Theorem 2 equals that of the random selection algorithm provided in Theorem 3 of (Li et al. 2019). In (10), we have the additional term E (number of local epochs) in front of J compared to Theorem 3 of (Li et al. 2019) as we use global round index t instead of using step index of local SGD. As the average aggregation cardinality decreases, a greater number of training rounds is required to achieve the same level of accuracy.

Remark 8. (Different Dropout Rates). When the dropout probabilities of the users are not the same, characterizing the theoretical and convergence guarantees of Multi-RoundSecAgg is challenging. This is due to the fact that batch selection based on the frequency of participation breaks the conditional unbiasedness of the user selection, which is required for the convergence guarantee.

6 Experiments

Our experiments consist of two parts. We first numerically demonstrate the performance of Multi-RoundSecAgg compared to the baselines of Sec. 3.3 in terms of the key metrics of Sec. 3.2. Next, we implement convolutional neural networks (CNNs) for image classification with MNIST (LeCun, Cortes, and Burges 2010) and CIFAR-10 (Krizhevsky and Hinton 2009) datasets to investigate how the key metrics affect the test accuracy.

Setup. We consider a FL setting with $N = 120$ users, where the server aims to choose $K = 12$ users at every round. We study two settings for partitioning the CIFAR-10 dataset across the users.

- **IID Setting.** In this setting, the 50000 training samples are shuffled and partitioned uniformly across the $N = 120$ users, where each user receives 417 or 416 samples.
- **Non-IID dataset.** In this setting, we utilize the *data-sharing strategy* of (Zhao et al. 2018), where the 50000 training samples are divided into a globally shared dataset $\mathcal{G}$ and private dataset $\mathcal{D}$. We set $|\mathcal{G}| = 200$ and $|\mathcal{D}| = 49800$. Then, we sort $\mathcal{D}$ by the labels, partition it into 120 shards of size 415, and assign each of the 120 users one shard. Each user has 200 samples of globally shared data and 415 samples of private dataset with one label.

For both settings, we implement LeNet (LeCun et al. 1999). While the state-of-the-art models (Kolesnikov et al. 2019; Tan and Le 2019) achieve 99% accuracy, LeNet is sufficient for our needs, as our goal is to evaluate various schemes, not to achieve the best accuracy. The hyperparameters of our experiment are provided in Appendix F.

Modeling dropouts. At each round, user $i \in [N]$ drops from the protocol with probability p_i. In the IID setting, p_i is selected from $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ uniformly at random. In the non-IID setting, to investigate how the aggregation fairness gap affects the test accuracy, we assume that p_i depends on the label of the private data. Specifically, the dropout probability of the users with the label 0 is 0.1 while each user with label 9 has a dropout probability of 0.5.

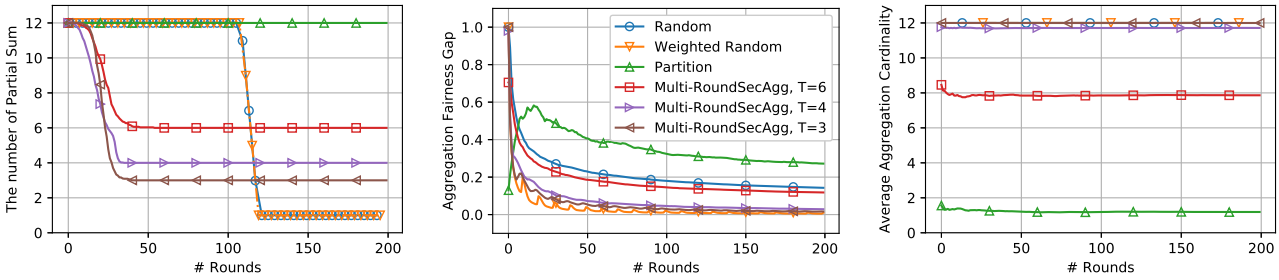
Implemented Schemes. For the benchmarks, we implement the three baselines introduced in Sec. 3.3, referred to as *Random*, *Weighted Random*, and *Partition*. For the proposed Multi-RoundSecAgg, we construct three privacy-preserving families of sets with different target multi-round privacy guarantees, $T = 6$, $T = 4$, and $T = 3$ which we refer to as Multi-RoundSecAgg ($T = 6$), Multi-RoundSecAgg ($T = 4$), and Multi-RoundSecAgg ($T = 3$), respectively. One can view the Random and Partition as extreme cases of Multi-RoundSecAgg with $T = 1$ and $T = K$, respectively. Table 1 summarizes the family size R defined in (8).

Scheme	Family size ($= R$)
Random selection	$\sim 10^{16}$
Weighted random selection	$\sim 10^{16}$
User partition	10
Multi-RoundSecAgg, T=6	190
Multi-RoundSecAgg, T=4	4060
Multi-RoundSecAgg, T=3	91389

Table 1: Family size with $N = 120$, $K = 12$.

Key Metrics. To numerically demonstrate the performance of the six schemes in terms of the key metrics defined in Sec. 3.2, at each round, we measure the following metrics.

- For the multi-round privacy guarantee, we measure the number of models in the partial sum that the server can reconstruct, which is given by $T^{(t)} := \min_{\boldsymbol{z}\in\mathbb{R}^J\}} \|\boldsymbol{z}^\top \mathbf{P}^{(t)}\|_0$, s.t. $\mathbf{P}^{(t)^\top}\boldsymbol{z} \neq \mathbf{0}$. This corresponds to the weaker privacy definition of Remark 1. We use this weaker privacy definition as the random selection and the random weighted selection strategies provide the worst privacy guarantee even with this weaker definition, as demonstrated later. On the other hand, Multi-RoundSecAgg provides better privacy guarantees with both the strong and the weaker definitions.
- For the aggregation fairness gap, we measure the instantaneous fairness gap, $F^{(t)} := \max_{i\in[N]} F_i^{(t)} - \min_{i\in[N]} F_i^{(t)}$ where $F_i^{(t)} = \frac{1}{t+1}\sum_{l=0}^{t} \mathbb{1}\{\{\boldsymbol{p}^{(l)}\}_i = 1\}$.



(a) Multi-round privacy guarantee. (b) Aggregation fairness gap. (c) Average aggregation cardinality.

Figure 3: The key metrics with $N = 120$ (number of users), $K = 12$ (number of selected users at each round).

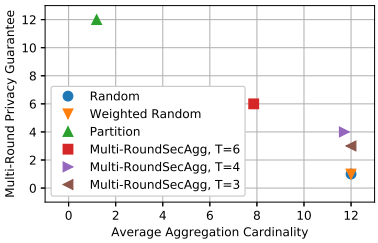
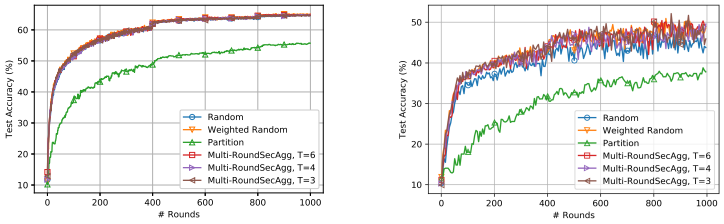


Figure 4: Trade-off between multi-round privacy guarantee versus average aggregation cardinality with $N = 120$, $K = 12$.



(a) IID data distribution. (b) Non-IID data distribution.

Figure 5: Training rounds versus test accuracy of LeNet in (LeCun et al. 1999) on the CIFAR-10 with $N = 120$ and $K = 12$.

- We measure the instantaneous aggregation cardinality as $C^{(t)} := \frac{1}{t+1}\sum_{l=0}^{t} \|\boldsymbol{p}^{(l)}\|_0$.

We demonstrate these key metrics in Figure 3. We make the following key observations.

- Multi-RoundSecAgg achieves better multi-round privacy guarantee than both the random selection and random weighted selection strategies, while user partitioning achieves the best multi-round privacy guarantee, $T = K = 12$. However, the partitioning strategy has the worst aggregation cardinality, which results in the lowest convergence rate as demonstrated later.
- Figure 4 demonstrates the trade-off between the multi-round privacy guarantee T and the average aggregation cardinality C. Interestingly, Multi-RoundSecAgg when $T = 3$ or $T = 4$ achieves better multi-round privacy guarantee than both the random selection and the weighted random selection strategies while achieving almost the same average aggregation cardinality.

Key Metrics versus Test Accuracy. To investigate how the key metrics affect the test accuracy, we measure the test accuracy of the six schemes in the two settings, the IID and the non-IID settings. Our results are demonstrated in Figure 5. We make the following key observations.

- In the IID setting, the Multi-RoundSecAgg schemes show test accuracies that are comparable to the random selection and random weighted selection schemes while the Multi-RoundSecAgg schemes provide higher levels of privacy. Specifically, the Multi-RoundSecAgg schemes achieve $T = 3, 4, 6$ based on the privacy-preserving family design while the random selection and random weighted selection schemes have $T = 1$, i.e., the server can learn an individual local model.
- In the non-IID setting, Multi-RoundSecAgg not only outperforms the random selection scheme but also achieves a smaller aggregation fairness gap as demonstrated in Fig. 3(b).
- In both IID and non-IID settings, the user partitioning scheme has the worst test accuracy as its average aggregation cardinality is much smaller than the other schemes as demonstrated in Figure 3(c).

We also implement additional experiments on MNIST and make similar observations in Appendix E.

7 Conclusion

Partial user participation may breach user privacy in federated learning, even if secure aggregation is employed at every training round. To address this challenge, we introduced the notion of long-term privacy, which ensures that the privacy of individual models are protected over all training rounds. We developed Multi-RoundSecAgg, a structured user selection strategy that guarantees long-term privacy while taking into account the fairness in user selection and average number of participating users, and showed that Multi-RoundSecAgg provides a trade-off between long-term privacy and average number of participating users (hence the convergence rate). Our experiments on the MNIST and CIFAR-10 datasets on both the IID and non-IID settings show that Multi-RoundSecAgg achieves comparable accuracy to the random selection strategy (which does not ensure long-term privacy), while ensuring long-term privacy guarantees.

References

Bell, J. H.; Bonawitz, K. A.; Gascón, A.; Lepoint, T.; and Raykova, M. 2020. Secure single-server aggregation with (poly) logarithmic overhead. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 1253–1269.

Bonawitz, K.; Ivanov, V.; Kreuter, B.; Marcedone, A.; McMahan, H. B.; Patel, S.; Ramage, D.; Segal, A.; and Seth, K. 2017. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 1175–1191.

Chen, W.; Horvath, S.; and Richtarik, P. 2020. Optimal Client Sampling for Federated Learning. *arXiv preprint arXiv:2010.13723*.

Cho, Y. J.; Gupta, S.; Joshi, G.; and Yağan, O. 2020. Bandit-based Communication-Efficient Client Selection Strategies for Federated Learning. *arXiv preprint arXiv:2012.08009*.

Cho, Y. J.; Wang, J.; and Joshi, G. 2020. Client Selection in Federated Learning: Convergence Analysis and Power-of-Choice Selection Strategies. *arXiv preprint arXiv:2010.01243*.

Choi, B.; Sohn, J.-y.; Han, D.-J.; and Moon, J. 2020. Communication-Computation Efficient Secure Aggregation for Federated Learning. *arXiv preprint arXiv:2012.05433*.

Elkordy, A. R.; and Avestimehr, A. S. 2020. Secure aggregation with heterogeneous quantization in federated learning. *arXiv preprint arXiv:2009.14388*.

Fredrikson, M.; Jha, S.; and Ristenpart, T. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 1322–1333.

Geiping, J.; Bauermeister, H.; Dröge, H.; and Moeller, M. 2020. Inverting Gradients–How easy is it to break privacy in federated learning? *arXiv preprint arXiv:2003.14053*.

Jain, V.; Sah, A.; and Sawhney, M. 2020. Singularity of discrete random matrices II. *arXiv preprint arXiv:2010.06554*.

Kadhe, S.; Rajaraman, N.; Koyluoglu, O. O.; and Ramchandran, K. 2020. FastSecAgg: Scalable Secure Aggregation for Privacy-Preserving Federated Learning. *arXiv preprint arXiv:2009.11248*.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2019. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*.

Kolesnikov, A.; Beyer, L.; Zhai, X.; Puigcerver, J.; Yung, J.; Gelly, S.; and Houlsby, N. 2019. Big transfer (bit): General visual representation learning. *arXiv preprint arXiv:1912.11370*, 6(2): 8.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical report, Citeseer.

LeCun, Y.; Cortes, C.; and Burges, C. 2010. MNIST handwritten digit database. *http://yann. lecun. com/exdb/mnist*.

LeCun, Y.; Haffner, P.; Bottou, L.; and Bengio, Y. 1999. Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*, 319–345. Springer.

Li, X.; Huang, K.; Yang, W.; Wang, S.; and Zhang, Z. 2019. On the Convergence of FedAvg on Non-IID Data. In *International Conference on Learning Representations*.

McMahan, H. B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Int. Conf. on Artificial Int. and Stat. (AISTATS)*, 1273–1282.

Nasr, M.; Shokri, R.; and Houmansadr, A. 2019. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, 739–753. IEEE.

Pejó, B.; and Biczók, G. 2020. Quality Inference in Federated Learning with Secure Aggregation. *arXiv preprint arXiv:2007.06236*.

Ribero, M.; and Vikalo, H. 2020. Communication-efficient federated learning via optimal client sampling. *arXiv preprint arXiv:2007.15197*.

So, J.; Güler, B.; and Avestimehr, A. S. 2021. Turbo-aggregate: Breaking the quadratic aggregation barrier in secure federated learning. *IEEE Journal on Selected Areas in Information Theory*, 2(1): 479–489.

Stich, S. U. 2018. Local SGD converges fast and communicates little. *arXiv preprint arXiv:1805.09767*.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 6105–6114. PMLR.

Tang, M.; Ning, X.; Wang, Y.; Wang, Y.; and Chen, Y. 2021. FedGP: Correlation-Based Active Client Selection for Heterogeneous Federated Learning. *arXiv preprint arXiv:2103.13822*.

Tran, T. 2020. The smallest singular value of random combinatorial matrices. *arXiv preprint arXiv:2007.06318*.

Yu, H.; Yang, S.; and Zhu, S. 2019. Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5693–5700.

Zhao, Y.; Li, M.; Lai, L.; Suda, N.; Civin, D.; and Chandra, V. 2018. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*.

Zhao, Y.; and Sun, H. 2021. Information Theoretic Secure Aggregation with User Dropouts. *arXiv preprint arXiv:2101.07750*.

Zhu, L.; and Han, S. 2020. Deep leakage from gradients. In *Federated Learning*, 17–31. Springer.